Corpus Linguistics 2013

Abstract Book

Edited by

Andrew Hardie and Robbie Love

Lancaster: UCREL

Table of Contents

Plenaries

Papers

Posters

Plenaries

What can translations tell us about ongoing semantic changes? The case of *must*

Karin Aijmer
University of Gothenburg
karin.aijmer@eng.gu.se

The grammaticalization of the modal auxiliaries is still under way. Even in a narrow time perspective we find changes indicating that the restructuring of the modal area is not complete. The changes affect both the epistemic and deontic meaning but have been particularly drastic for deontic *must.* Leech et al. (2009) compared modal auxiliaries in corpora constructed according to the same design but from different periods. We can also compare the modal auxiliaries across languages to establish similarities and differences.

The changes which have taken place in English can be highlighted by a comparison with Swedish and English where the same changes have not take place. My comparison is based on the occurrence of *must* and *måste* in the English-Swedish Parallel Corpus (ESPC).The translations provide a panorama of the substitutes of *must* and raw material for describing how they can be distinguished from their neighbours in the area of deontic imodality.

Taking a Language to Pieces: art, science, technology

Guy Cook
King's College
guy.cook@kcl.ac.uk

My talk explores the relationship between language as a lived experience, representations of language for (corpus) linguistic analysis, and the use of (corpus) linguistics in technology. My claim is that important distinctions are being overlooked. In the course of this exploration, I consider past conflicts between literary criticism and linguistics, the current fashionable view that only holistic depictions of language are worthwhile, and the ongoing subordination of academic research to politically partisan technological agendas.

In literature and everyday communication, the lived experience of language is both infinitely complex and inextricable from value judgments. Linguistics, in contrast, seeks to approach (though it never quite reaches) this experience by simplification and selection, and by putting evaluation to one side. Like medical diagrams which show bones or muscles or nerves in images which are very unlike a real person, linguistics seeks an understanding of one part of language at the time, not in the erroneous belief that this is reality, but as a necessary prelude to a better informed re-engagement with reality.

Corpus linguistics shares this commitment to idealisation, and it is this allegiance which underpins its unrivalled extension of both the description and the theory of language. Its data are selective and partial, and for this very reason, powerful. In its applications, however, like the other sciences, corpus linguistics inevitably re-engages with the values and complexity of language as a lived experience. It becomes political and evaluative, and open to question.

Corpus linguists need to maintain the distinctions between art, science and technology, and to see the strengths and weaknesses of each. A failure to do so is easily exploited by political and commercial opportunists, and poses a threat not only to the independence and the achievements of corpus linguistics, but to academic enquiry as a whole.

The textual dimensions of Lexical Priming

Michael Hoey
University of Liverpool
hoeymp@liverpool.ac.uk

Matthew Brook O'Donnell
University of Michigan
mbod@umich.edu

It has always been a claim of lexical priming theory that it accounts for textual phenomena as well as lexical and grammatical phenomena. Three symmetries have been proposed between lexical and textual features as identified in corpus linguistics. The first of these is between collocation and textual collocation (which approximates cohesion); the term 'collocation' is indeed ambiguous in the literature, being used both a corpus-linguistic phenomenon and a cohesive one, but the ambiguity is explained if we recognise the symmetry just mentioned.

The second is between semantic association (or semantic preference) and textual semantic association, and is the least explored of the three symmetries. Evidence, though, will be presented that suggests that textual semantic association links the lexicon to the text-semantic relations of the kind that were explored in the 1970s and since then have been largely neglected. The admittedly inadequate evidence that will be offered seems nevertheless to suggest that a more thorough exploration is required.

The third symmetry is that between colligation and textual colligation, and is the most thoroughly explored of the three symmetries. First evidenced in the early work of Halliday, textual colligation is an attempt to account lexically for choices affecting Theme-Rheme, paragraph boundaries and text initiation.

In a thorough exploration of text-initiation and paragraph boundaries in hard news stories, funded by the AHRC, and in conjunction with Michaela Mahlberg and Mike Scott, the authors have found that there is a strong association between text-positioning and lexical choice, both at the level of the single word and at the level of the cluster. Drawing also on experiments with informants, the authors seek to show that paragraphing is a very different phenomenon from that usually posited in the applied linguistic literature and one that can be evidenced using corpus linguistic techniques. In so far as textual colligation and the other textual symmetries discussed are shown to be supported, they also fail to disconfirm the claims of lexical priming theory.

No corpus linguist is an island: Collaborative and cross-disciplinary work in researching phraseology

Ute Römer
Georgia State University
uroemer@gsu.edu

Over the past few decades, corpus linguists have done a lot to move phraseology research from the periphery to the heart of linguistics (using Ellis' 2008 terminology). The work of John Sinclair and other researchers inspired by the British contextualist tradition has been particularly influential in this context. Phraseology has also become of core interest to researchers in related fields such as psycholinguistics, natural language processing, cognitive linguistics, and language acquisition and instruction (as testified by the range of contributions on the topic in the 2012 *Annual Review of Applied Linguistics*).

This talk argues that progress and development in phraseology research will depend to a large extent on successful collaborations between corpus linguists and scholars from other fields, and on a skillful combination of analytic techniques in a "methodological pluralism" sense (McEnery & Hardie 2012). The talk starts with a brief overview of a few important strands in current corpus-based phraseology research. It then presents findings from four phraseological studies that all benefited from the presenter's collaboration with researchers from neighboring disciplines, including a computational linguist, a genre expert, a psycholinguist, and a cognitive linguist:

- A study that develops an analytical model to determine the phraseological profile of a text type (Römer 2010);
- A study that attempts to measure formulaic language (FL) in corpora of academic writing by native and non-native speakers at different proficiency levels, using a variety of operationalizations of FL (O'Donnell, Römer & Ellis 2013);
- A study that combines quantitative and qualitative approaches to the distribution of attended and unattended *this* in advanced student writing across disciplines (Wulff, Römer & Swales 2012); and
- A study that examines verb-argument constructions in language use and in speakers' minds, drawing on corpus data and psycholinguistic evidence (Ellis, O'Donnell & Römer 2013; Römer, O'Donnell & Ellis submitted).

The talk closes with thoughts on future avenues for cross-disciplinary research on phraseology.

References

Ellis, N. C. 2008. Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in Language Learning and Teaching* (pp. 1-13). Amsterdam: John Benjamins.

Ellis, N. C., M. B. O'Donnell & U. Römer. 2013. Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning* 63(Supp. 1): 25-51.

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press.

O'Donnell, M. B., U. Römer & N. C. Ellis. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18(1): 83-108.

Römer, U. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1): 95-119. [Reprinted in: Biber, Douglas & Randi Reppen (eds.). 2012. *Corpus Linguistics. Volume I: Lexical Studies*. London: SAGE Publications. 307-329.]

Römer, U., M. B. O'Donnell & N. C. Ellis. Submitted. Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge.

Wulff, S., U. Römer & J. M. Swales. 2012. Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives. *Corpus Linguistics and Linguistic Theory* 8(1): 129-157.

Papers

A corpus-based study for assessing the collocational competence in learner production across proficiency levels

Maha N. Alharthi
Princess Nora University
dr.maha2006@gmail.com

It is widely acknowledged that EFL/ESL language learners face a considerable challenge in mastering L2 collocations in their written and spoken language, regardless of their L1 and/or the length of L2 instruction (Gitsaki 1996; Granger 1998; Groom 2009; Howarth 1998; Laufer and Waldman 2011; Nesselhauf 2003, 2005). This paper investigates Arab EFL learners' collocational use of the highly frequent light verbs (LVs): MAKE, DO, and HAVE. These three verbs were selected for two main reasons: First, they appear at the top of any corpus-based list of high-frequency verbs (apart from BE and the modal auxiliaries) (Altenberg and Granger 2001). Second, Arab EFL learners tend to produce collocational errors by confusing them with each other; e.g., they may produce *make my homework and *do a mistake. More attention should be given to these constructions in L2 instruction since a high percentage of errors, that have a disruptive impact on the processing by native speakers (Millar 2011), have been observed to occur in them (see Howarth 1998; Nesselhauf 2003).

The objective of this study is to describe the developmental patterns of the collocational knowledge of L2 learners at various proficiency levels through their production of light verb constructions (LVCs). This study investigated three different proficiency groups and compared them to each other in order to trace any possible developmental pattern in group performance.

Proficiency is a controversial topic in which contradictory results have emerged. Some studies indicate that the use of collocations is related to proficiency (Gitsaki 1996; Al-Zahrani 1998), whereas others indicate no correlation (Howarth 1998; Laufer and Waldman 2011). Hopefully, this study may provide further empirical information that may help describe how collocational competence develops. More importantly, the SLA literature reveals very few studies of collocations which used error analysis approaches and/or elicitation tasks to investigate the collocational problems encountered by Arab L2 learners (e.g., Al-Zahrani 1998; Farghal and Obiedat 1995).

This study adopts Sinclair's (1991b:170) frequency-based approach to collocation which is defined as co-occurrence of words at a certain distance with significant frequencies. It attempts to answer the following questions:

1. Is there a relationship between the learners' collocational use of LVCs and their language proficiency?
2. Do learners tend to use these LVs in more collocational combinations than non-collocations?

To answer these questions, authentic learner data has been investigated using a subset of the BUiD Arab Learner Corpus (BALC) [1], which consists of examples of Arabic L1 learner English at various proficiency levels, ranging from post-beginners to upper intermediate. A frequency list of each LVC was generated using AntConc. These lists were presented to a native speaker to assess the appropriateness of these constructions in the context of two sentences derived from the corpus. The following table represents the descriptive data for each sub-corpus:

	Do			
Level	No.	% of Appro. LVCs	% of Devi LVCs	LVCs
1	162	44 (8)	56 (10)	18
2	240	85 (92)	15 (16)	108
3	338	90 (110)	10 (12)	122
	Make			
Level	No.	% of Appro. LVCs	% of Devi LVCs	LVCs
1	22	22 (2)	78 (7)	9
2	141	48 (36)	52 (39)	75
3	267	78 (71)	22 (20)	91
	Have			
Level	No.	% of Appro. LVCs	% of Devi LVCs	LVCs
1	481	95 (91)	5 (5)	96
2	747	99 (268)	2 (4)	272
3	1538	98 (542)	2 (12)	554

Table 1: Percentages & Frequency of Appropriate vs. Deviant LVCs

To find out whether the above differences are large enough to be significant, the chi-square test for independence was applied.

Table 2 shows a statistically significant relationship between the learners' collocational

[1] BALC was compiled by Randall and Groom (2009) for the purpose of investigating L2 learners' acquisition of English spelling.

competence and their proficiency in two of the LVs: DO and MAKE. The more proficient learners produce significantly more appropriate LVCs than the lower group. For Have, the chi-square test returns an insignificant result. In order to quantify the strength of the observed correlation independently of the sample size, the effect size was computed. The measure of the effect size for Do and Make gives the following results: 0.32 and 0.36, respectively, which reveals an intermediately strong correlation.

Light Verb	df	X^{2}(X-squared)	P-value	Cramer's V
DO	2	25.31	3.20E-06	0.32
MAKE	2	22.26	1.47E-05	0.36
HAVE	2	4.53	0.1038	0.07

Table 2: Chi-Square Tests:
Appropriate vs. Deviant LVCs

Finding the most appropriate collocations for MAKE ranks first as the most difficult for learners at different proficiency levels, revealing a decrease in the number of errors with a growth in proficiency. Second, the collocations of DO rank next in the hierarchy of difficulty. HAVE ranks third, displaying very low percentages of errors for all levels. This result is consistent with the results of Nesselhaulf (2004) and Erman (2009).

To find out which proficiency group is most responsible for the significant effect obtained for Do and Make, Pearson residuals were computed.

DO	Appropriate LVCs	Deviant LVCs
Level 1	-1.85	4.36
2	0.16	-0.14
3	0.66	-1.55

Table 3: The Pearson residuals: DO

As shown above, the strongest effect is the large number of the deviant LVCs produced by post-beginners, followed by their production of small number of the appropriate constructions and the small quantity of the deviant constructions produced by upper-intermediate. By contrast, the residuals for the intermediate proficiency level are closer to zero, since their observed frequencies are close to the expected ones.

MAKE	Appropriate LVCs	Deviant LVCs
Level 1	-1.52	1.96
2	-1.57	2.02
3	1.90	-2.44

Table 4: The Pearson residuals: MAKE

For MAKE, the strongest effect seems to come from the dispreference of deviant constructions by upper-intermediate students, followed by the deviant collocations produced by intermediate learners and post-beginners where the observed frequency is greater than expected. By contrast, their production of appropriate LVCs are less than expected.

This result can be illustrated in the following association plots, which indicate that the significant effect is mostly due to the fact that the deviant LVCs are more likely produced by post-beginners.

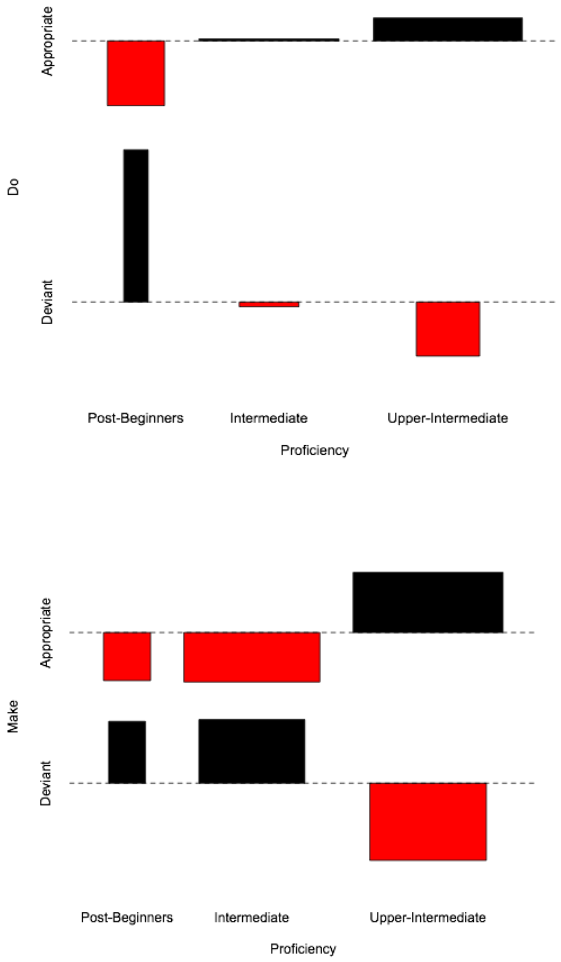


Figure 1: The relation between the collocational competence of LVCs and Proficiency

Concerning the second research question about

the difference between the number of the collocational occurrences of each LVC and those of non-collocation? The descriptive data is shown in the following table:

	Do		
Level	No.	Appro. LVCs	Non-Col.
1	162	8	144
2	240	92	127
3	338	110	216
	Make		
Level	No.	Appro. LVCs	Non-Col.
1	22	2	13
2	141	36	66
3	267	71	176
	Have		
Level	No.	Appro. LVCs	No
1	481	91	316
2	747	268	502
3	1538	542	984

Table 5: Frequency of Appropriate LVCs vs. Non-collocations

To compare the differences in collocation frequencies among the three sub-corpora, I related the number of the collocational use of each LV to the number of non-collocation occurrences in each sub-corpus.

Light Verb	df	X^{2}(X-squared)	P-value	Cramer's V (The effect size)
Do	2	61.35	4.77E-14	0.30
Make	2	3.53	0.17	0.10
Have	2	26.19	2.17E-06	0.10

Table 6: Chi-Square Test: Collocations vs. Non-Collocations

A significant correlation was revealed among the three proficiency groups and their collocational preferences of Do and Have. No significant correlation was found in their preferences of collocation patterns for Make.

The nature of this association then becomes clear from the following residuals: appropriate collocation of Do and Have are not preferred by post-beginners (negative residuals of ≈-5.59 and -3.83, respectively). Instead, they tend to use these LVs in non-collocations more than expected (positive residuals of ≈3.67 and 2.71, respectively). The more proficient groups prefer using LVs in appropriate collocations.

DO	Appropriate Collocations	Non-Collocations
Level 1	-5.59	3.67
2	3.20	-2.10
3	1.19	-0.78

Table 7: The Pearson residuals: DO

HAVE	Appropriate Collocations	Non-Collocations
Level 1	-3.83	2.71
2	0.71	-0.50
3	1.48	-1.05

Table 8: The Pearson residuals: HAVE

The association plots in Figure 2 indicate that the more proficient learners display more preference of using LVs in collocational patterning than in non-collocations. However, the post-beginners learners follow the opposite pattern.

To sum up, the main finding of the study is that a clear correlation can be observed between learners' collocational competence of LVCs and their proficiency. Collocation competence was observed to increase as the level of proficiency increases. The results of the lower group production data may suggest that they process and store these constructions analytically rather than holistically. They start out with single words and break these expressions down into their constituent parts. Later in the acquisition process, they start using some forms of prefabricated patterns. Developing the learners' awareness of the significance of these constructions may help improve their collocation competence.

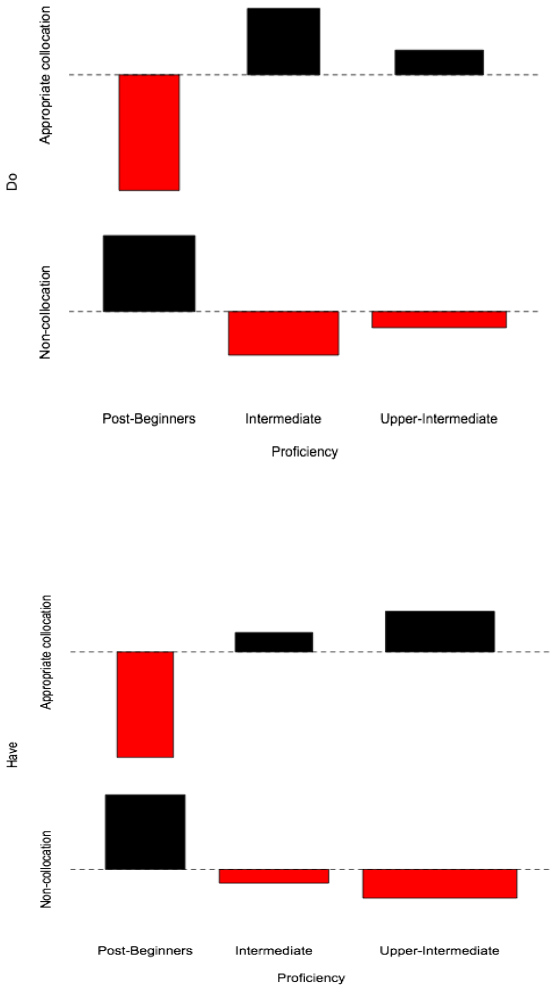


Figure 2: The relation between collocation occurrences vs. non-collocation

References

Al-Zahrani, M.S. 1998. *Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi university*. Unpublished PhD thesis, Indiana University of Pennsylvania.

Altenberg, B and Granger, S. 2001."The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing". *Applied Linguistics* 22(2): 173-195.

Erman, B. 2009. "Formulaic Language from a learner Perspective: What the learner needs to know". In Corrigan, R. et al. *Formulaic Language: Acquisition, Loss, Psychological reality, and functional explanations*, pp.323-46. John Benjamins B.V.

Farghal, M. and Obiedat, H. 1995. "Collocations: A neglected variable in EFL". *International Review of Applied Linguistics in Language Teaching* 33 (4): 315–332.

Gitsaki, C. 1996. *The development of ESL collocational knowledge*. PhD thesis. The University of Queensland.

Granger, S. (ed.). 1998. Learner English on Computer. London: Longman.

Granger, S. 1998a. "Prefabricated patterns in advanced EFL writing: collocations and formulae". In A Cowie. *Phraseology: Theory, Analysis, and Applications*. Oxford: OUP.

Gries, S. 2009. *Quantitative corpus linguistics with R*: a practical introduction. London: Routledge.

Gries, S. (to appear). "Frequency tables, effect sizes, and explorations". In Dylan Glynn & Justyna Robinson (eds.). *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam & Philadelphia: John Benjamins.

Gries, St. Th., & Wulff, S. 2005. "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora". *Annual Review of Cognitive Linguistics* 3: 182–200.

Groom, N. 2009. "Effects of second language immersion on second language collocational development". In A. Barfield & H. Gyllstad (eds.) *Researching collocations in another language: multiple interpretations*. London: Palgrave Macmillan.

Handl, S. 2008. "Essential collocations for learners of English". In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 43–66). Amsterdam: John Benjamins.

Hasselgren, A. 1994. "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary". *International Journal of Applied Linguistics* 4(2): 237-258.

Howarth P. 1998. "Phraseology and Second Language Proficiency". *Applied Linguistics*, t. XIX, s. 24–44.

Kjellmer, G. 1991. "A Mint of Phrases". In A. Cowie *Phraseology: Theory, Analysis, and Applications*. Oxford: OUP.

Laufer, B. and Waldman, T. 2011. "Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English". *Language Learning* 61(2): 647-72.

Leech, G. 2006. *A Glossary of English Grammar*. Edinburgh Univ. Press.

Millar, N. 2011. "Processing malformed formulaic language," *Applied Linguistics* 32(2):129-148.

Nesselhauf, N. 2003. "The Use of Collocations by Advanced Learners of English and Some Implications for Teaching". *Applied Linguistics* 24 (2): 223–242.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Palmer, F. R. 1981. *Semantics*. A New Outline. Cambridge: CUP.

Quirk, R., Greenbaum, S., Leech G. and Svartvik J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

'*Sure* he has been talking about coming for the last year or two": the *Corpus of Irish English Correspondence* and the use of discourse markers

Carolina P. Amador-Moreno
University of Extremadura
camador@unex.es

Kevin McCafferty
University of Bergen
Kevin.McCafferty@if.uib.no

Few features of Irish English have been studied diachronically at all, and the area of discourse markers is likewise largely neglected even as regards present-day Irish English (Barron & Schneider 2005; Amador-Moreno 2010; Corrigan 2010). This study analyses the use of four pragmatic markers: the variants *anyway* and *anyhow*, *like* and *sure* in the *Corpus of Irish English Correspondence* (CORIECOR), which contains private correspondence from the late seventeenth century to the early twentieth, in order to survey their diachronic development. CORIECOR is a corpus of personal letters which covers the timespan from 1750-1940. The corpus contains some 4700 texts (approx. 3 million words), of which 4100 (2.5m words) are correspondence maintained between Irish emigrants and their relatives, friends and contacts. The letters were sent mainly between Ireland and other countries such as the United States and Canada, Great Britain, New Zealand, and Australia, and therefore provide an empirical base for studies of historical change in IrE and its contribution to other major overseas varieties.

We look at how the use of *anyhow* is significant in the letters, showing that it was a widespread discoursal feature in Ireland by the 19th century. We will also discuss the presence of *like* in the corpus, which bears out Tagliamonte's claim that 'discourse *like* had already made a grammatical shift towards discourse particle (rather than discourse marker) well before its surge in frequency in North America' (Tagliamonte 2012: 172). Finally, we analyse the use of *sure*, a distinctive trait of Irish English (IrE), which may have been an IrE pragmatic marker for up to 400 years, surviving in spite of stereotyping and normative stigmatisation. IrE *sure* is different from AmE *sure* in that it tends to be uttered as part of a larger intonation group and is produced with a reduced vowel and no stress or intonational prominence. Our study suggests that the emphatic AmE uses of *sure* might have grammaticalised from the IrE (and possibly also BrE) uses taken to North America by emigrants, as indicated by Aijmer (2009:339).

Our paper argues that the evidence of this corpus of private correspondence seems to indicate that the variant *anyways*, which prescriptivists condemned as Irish, is absent from the corpus. This suggests that such claims may not have been based on observation of real usage. The presence of *anyhow* and *anyway* in the letters also supports the hypothesis that colloquialisation may have played an important role in the rise of speech-like features, triggering a change from below, as literacy enabled more of the population to express themselves in writing, so that the linguistic traits of lower social strata were recorded in writing too. We will also shown that *like* was already a DM long before the appearance of striking new uses in North America in the late twentieth century. Comparison of our data with similar NAmE data might help explain the development of DM uses in certain structural positions, particularly in order to account for the prevalence of clause-final *like* in IrE, as opposed to its disappearance in AmE, despite evidence of transportation to the New World by Irish emigrants. Our study also documents the use of *sure* in various structural positions over the last few centuries and suggests that unstressed uses of *sure* could have followed an interesting developmental path. One possibility is that it was first imported into IrE from EModE during the period of British settlement, then re-exported through emigration.

The paper also brings attention to the value of emigrant letters for the study of language variation and change. Letters are among the more 'oral' text types available for linguistic study (Schneider 2002), and corpus-based studies covering nearly 1000 years of English language history show personal correspondence to be more vernacular, and more sensitive to linguistic variation and change, than other text types (e.g. Nevalainen & Raumolin-Brunberg 2003).

This initial investigation into the use of DMs in CORIECOR draws attention to the need for further diachronic analysis of IrE itself, as well as comparisons with other varieties. Analyses of this kind would address issues related to the diffusion of DMs between varieties, thus providing testing grounds for the grammaticalisation hypothesis that assumes a move from strictly textual to more interpersonal and pragmatic meanings. Future analysis of these DMs in CORIECOR, accounting for social aspects such as the nature of relationships between letter-writer and reader, social status, level of education, gender, and regional distribution will help shed further light on the use and development of these features in Irish English.

References

Aijmer, Karin 2009. The pragmatics of adverbs. In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English.* Cambridge: Cambridge University Press. 324-340.

Amador-Moreno, Carolina P. 2010. *An Introduction to Irish English*. London: Equinox.

Barron, A. And Schneider, K. 2005. *The Pragmatics of Irish English.* Berlin: Mouton de Gruyter.

Corrigan, K. P. 2010. *Irish English, Volume 1 – Northern Ireland*. Edinburgh: Edinburgh University Press.

Nevalainen, T. & H. Raumolin-Brunberg 2003. *Historical sociolinguistics. Language change in Tudor and Stuart England.* London: Longman.

Schneider, E.W. 2002. Investigating variation and change in written documents. In P. Trudgill, J.K. Chambers & N. Schilling-Estes (eds.), *The handbook of language variation and change*. Oxford: Blackwell. 67-96.

Tagliamonte, Sali A. 2012. *Roots of English. Exploring the history of dialects.* Cambridge: Cambridge University Press.

Developing *AntConc* for a new generation of corpus linguists

Laurence Anthony
Waseda University
anthony@waseda.jp

1 Introduction

Concordance software is one of the most important but often forgotten components of any corpus study. Over the years, various concordance software tools have been released in what McEnery and Hardie (2012) have described as four generations of tool development. Today, both 3rd-generation software tools, such as *AntConc* (Anthony, 2012), *WordSmith Tools* (Scott, 2012), and *MonoConc Pro* (Barlow, 2000), and 4th-generation web-based tools, such as those hosted at *byu.corpus.edu* (Davies, 2012) are popular choices for most corpus research work.

In recent years, *AntConc* (Anthony, 2012) has seen a rapid growth in popularity among researchers, teachers, and language learners due to its rich set of features, freeware license, multiplatform support, and easy-to-use interface. For researchers, *AntConc* performs speedily and accurately on a wide-range of small and mid-sized corpora. It also offers flexible handling of tags, metadata, and language encodings, and provides a wealth of functions and features. In 2012 alone, the software was downloaded over 120,000 times by users in over 80 countries, and it has become one of the software tools of choice in many corpus linguistics departments looking to introduce students to corpus linguistics through a free and easy tool (Anthony, 2012). For teachers and learners, *AntConc* can perform basic operations, such as producing KWIC concordance lines and keyword lists, in a quick and easy way. Also, it can be used both inside the classroom and as part of student homework projects on Windows, Macintosh OS X, and Linux computers. Finally, to motivate learners to use corpora in their learning, it offers a modern and attractive-looking interface.

Although *AntConc* has many strong features, it also has a number of weaknesses when compared to the most popular web-based and commercial tools. To address these issues, various design and performance improvements have been introduced in the latest version of the software. In this paper, I first review the current status of corpus analysis tools discussing their respective strengths and weaknesses, and explaining the motivation to introduce changes to *AntConc*. Next, I describe the changes introduced in the latest version of *AntConc*. As part of the discussion, I explain the choice of programming

language, the importance of including a flexible and explicit token definition, and the approach used by *AntConc* to handle whitespace issues that provides far greater transparency and flexibility over other tools. Many of these changes are directly relevant not only to corpus linguists but also teachers and learners who use corpus tools as part of a Data-Driven Learning (DDL) approach.

2 Current status of corpus analysis tools

Recently, Tribble (2012) conducted a major survey of the most popular tools used by corpus linguists around the world. Based on responses from 891 linguists, he showed that three tools are predominantly used today: *corpus.byu.edu*, *WordSmith* Tools and *AntConc* (see Figure 1). Viewing these results, it is apparent that the most popular tools are fast, easy-to-use, and feature-rich.

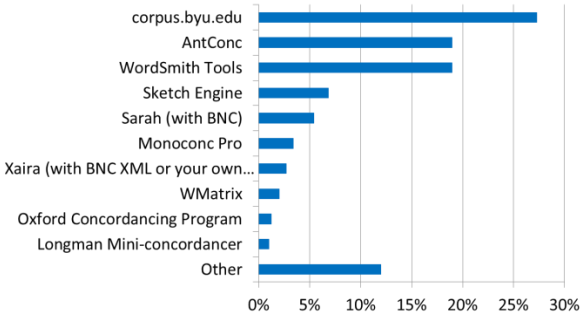


Figure 1. Survey results in response to the question: "Which computer programs do you use for analysing corpora?" Reponses: 891. (Tribble, 2012)

On the other hand, Tribble also reports that more advanced tools are being increasingly desired by corpus researchers. As a result, there is a growing interest in software development and coding, for example, using the R statistical package.

From the results of Tribble's survey, it is clear that for many corpus linguists a number of areas need addressing. Firstly, most web-based tools (e.g. corpus.byu.edu) provide only a window to a general English language corpus and offer no access to the raw corpus data. This makes it difficult to use them in many situations, for example, when the researcher is interested in the specialized language use in English for Specific Purposes (ESP) research. In this context, general corpus tools, such as *WordSmith Tools* (Scott, 2012) and *MonoConc Pro* (Barlow, 2000) can be used successfully as they can process almost any corpus for which the raw data is available. On the other hand, these tools come with a restrictive commercial license that often prevents their use in countries and regions where budgets are limited. In particular, they are difficult to use outside of the classroom, for example, as part of student homework projects. Also, statistical information is becoming increasingly important in corpus work, but the current tools do not easily generate the latest statistical results needed by researchers.

AntConc is a freeware tool that is able to process raw corpus data of various kinds. As a result, it can be useful in many contexts. However, in recent years, *AntConc* has begun to fall behind other tools in terms of speed, mainly due to its database architecture. *AntConc* processes all data in active memory and uses only a primitive indexing system. In addition, some common computer operations, such as drag-and-drop, are not available due to aging nature of the programming languages used to develop the software, i.e., *Perl* and *TclTk*. Various other limitations have recently become apparent, such as its limited support for handling annotated files and its limited statistical functions. In view of these limitations, an effort to update *AntConc* has been undertaken over the past three years. The results of this work are described in the following section.

3 New Features in *AntConc*

Programming language: This biggest change to *AntConc* has been to recode the software in the *Python* programming language together with the *Qt* graphical user interface package. The use of Python allows more advanced statistics modules to be included directly in the software thus addressing a major weakness in previous versions. Also, the use of *Qt* allows *AntConc* to adopt a more modern appearance and enables standard computer operations, such as drag-and-drop, to be incorporated. The use of *Qt* also allows rich-text tables to be utilized, leading to fast rendering of color-highlighted results.

Database architecture: The new version of *AntConc* incorporates a SQLite backend database that can operate in an indexed mode or on plain text files. This gives the program the ability to search for results on much larger, multi-level annotated files, while also being able to search in plain text files using regular expressions.

Multi-language support: *AntConc* has always offered multi-language support. However, the new version extends this to include flexible definitions of words that can extend or replace various Unicode character classes. It can also handle cross-platform line breaks and various forms of whitespace in and between text lines.

Performance improvements: Various other coding changes have led to the latest version of *AntConc* performing at greatly increased speeds over previous versions. This has led to some operations that took minutes to perform in the past completing in a matter of seconds or fractions of a second. The

software no longer has a sluggish feel and can handle very large corpora of 100s of millions of words without problem. A screenshot of the new software is shown in Figure 2.

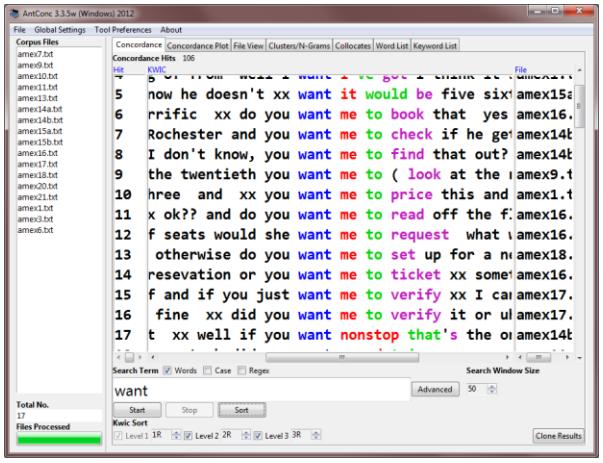


Figure 2. Screenshot of latest version of *AntConc*. (Anthony, 2012)

4 Conclusions

Users of *AntConc* should find the latest version to be much improved over previous versions. It is hoped that these changes will enable to software to continue meeting the needs of the corpus linguistics community.

References

Anthony, L. (2012). *The Past, Present, and Future of Software Tools in Corpus Linguistics*. Presentation given at KACL 2012, Busan, Korea.

Anthony, L. (2012). *AntConc* (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Barlow, M. (2000). *MonoConc Pro* [Computer Software]. Available from http://www.athel.com/mono.html.

Davies, M. (2012). corpus.byu.edu.

McEnery, T and Hardie, A (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Scott, M. (1996). *WordSmith Tools* [Computer Software]. Available from http://www.lexically.net/software/index.htm.

Tribble, C. (2012). T*eaching and Language Corpora: Quo Vadis?* 10th Teaching and Language Corpora Conference (TALC). Warsaw, 11th-14th July 2012.

Bridging lexical and constructional synonymy, and linguistic variants – the Passive and its auxiliary verbs in British and American English

Antti Arppe
University of Alberta
arppe@ualberta.ca

Dagmara Dowbor
University of Alberta
dowbor@ualberta.ca

1 Introduction and Background

In the case of *constructional alternations* – or *synonymous syntactic variants* – studies utilizing multivariate/multicausal predictive models have been recently undertaken, though primarily for binary/dichotomous alternations of only two outcomes, and typically for English, e.g. the possessive alternation: X of Y vs. Y's X (Gries 2002; Rosenbach 2003), verb particle placement (Gries 2003a), and the dative alternation: GIVE NP NP vs. GIVE NP PP (Gries 2003b; Bresnan 2007), Exeeptions scrutinizing polytomous settings with more than two outcomes are various studies on German word order variation and the German active vs. *werden*-passive vs. *bekommen*-passive (Bader & Häussler). Inspired by the latter study, Arppe (2011) has scrutinized the English four-way constructional alternation of active vs. *be*-passive vs. *get*-passive vs. *become*-passive, which can be seen to merge lexical with constructional alternation, using British English data (British National Corpus: BNC) (see examples 1-4):

(1) BECOME: ... how the siege$_{\text{PATIENT}}$ became interpreted by today's protestant loyalists$_{\text{AGENT}}$ ... [original sentence in BNC]
(2) GET: ... how the siege got interpreted by today's protestant loyalists ...
(3) BE: ... how the siege was interpreted by today's protestant loyalists ...
(4) ACTIVE: ... how today's protestant loyalists interpreted the siege ...

Based on a statistical multivariate analysis of the British English corpus data using *polytomous logistic regression* (Arppe 2008, 2012) on a range of contextual predictors, (a) active constructions were observed in the British English corpus data to be significantly associated with having an explicitly expressed AGENT or other argument which could be turned into subject in a corresponding active construction, or a co-ordinated verb, (b) *be*-passive constructions with ABSTRACTIONs, ACTIVITIES/EVENTs, ARTEFACTs, [forms of] COMMUNICATION, HUMAN GROUPs or [forms of] POSSESSION as PATIENTs (i.e. grammatical subjects

in passive constructions) and expressions of NECESSITY in the verb chain; (c) *become*-passives with HUMAN GROUPs as PATIENTs or with adverbials of QUANTITY/DEGREE; and (d) *get*-passives with HUMAN INDIVIDUALs as PATIENTs and clausal meta-arguments as well as phrasal elements (e.g. *get bogged down*).

Later on, Arppe, Bolger (2012) & Arppe, Bolger & Dowbor (2012) have contrasted these corpus-based results against other types of linguistic evidence, suggesting (a) a curvilinear/complex relationship between corpus-based (relative) frequency, as expected *Probability* given a multiple of contextual factors, and graded *Acceptability* ratings; (b) a (roughly) linear relationship between *Proportion* (in selection) and *Probability* (while the relationship between *Proportion* and *Acceptability* appears again curvilinear), and (c) a significant inverse effect between perceptive/cognitive *processing ease* as reflected in reading in an eye-tracking experiment and corpus-based *Probability*.

This study extends the scrutiny to specifically the three alternative passive constructions and to new, substantially more extensive data from American English, with both a more detailed account of the phenomenon as well as possible comparisons between American and British English in mind.

2 Data, linguistic annotation and statistical analysis techniques

Our data sample was obtained from Corpus of Contemporary American English (COCA), consisting of 3,000 concordance lines of 500 instances per auxiliary verb by modality (spoken vs. written), which were then coded for ~80 morpho-syntactic and conceptual variables. Some initial univariate analysis results concerning these variables have already been presented by Dowbor et al. (2012), and have informed us in this follow-up study.

The multivariate analysis method we selected, polytomous logistic regression applying the *one-vs-rest* heuristic (Arppe 2012), allows us to (1) estimate the relative weights of the linguistic explanatory variables, with each passive auxiliary contrasted against the two others, as well as also (2) model the impact of their joint occurrences as expected probability distributions for the alternative passive constructions, illuminating unique prototypical uses of each construction and their most exemplary sentences – as well as those contexts in which the passive auxiliaries appear interchangeable.

3 Results

The considerably larger American English corpus sample allowed us to undertake a more detailed analysis of the contextual factors influencing and associated with the selection of any of the three passive auxiliary constructions. Table 1 presents the impacts of perhaps the two most important contextual properties with respect to this phenomenon, namely the conceptual subtypes of the (preceding) grammatical subject and (following) prepositional object arguments. For this relatively simple model the *Accuracy* is already relatively high at 61.4%, and the confidence of the model in terms of its overall probability estimates aggregated in MacFadden's pseudo-R-squared statistic is moderately good at R_L^2=0.236.

In terms of grammatical subjects, i.e. functional PATIENTs of the passive constructions, we can note that an ACTIVITY, AMOUNT, COGNITION, COMMUNICATION, INSTRUMENT, and TEXT are significantly in favor of the *be* auxiliary. In turn, a PERSON or PEOPLE as subject are significantly in favor of the *get* auxiliary, but none of the conceptual subtypes is significantly associated with the *become* auxiliary. Moving onto the prepositional objects, often indicating an argument that could be turned into the subject in a corresponding active construction, an ACTIVITY or TEXT are significantly in favor of the *be* auxiliary, while a NAME or SUBSTANCE are in favor of the *become* auxiliary, and an INSTITUTION, THING, or no prepositional object at all (*none*) are significantly in favor of the *get* auxiliary. Our data sample already contains further coding on e.g. clause types and adverbials, which could be used to fine-tune the analysis substantially further.

Contextual property	be	become	get
(Intercept)	0	0	-
SUBJ.ACTIVITY	+	-	-
SUBJ.AMOUNT	+	-	0
SUBJ.ANIMAL	0	0	0
SUBJ.BODY_PART	0	0	0
SUBJ.COGNITION	+	-	0
SUBJ.COMMUNICATION	+	0	0
SUBJ.EVENT	0	0	-
SUBJ.INSTITUTION	0	0	0
SUBJ.INSTRUMENT	+	-	0
SUBJ.PEOPLE	-	0	+
SUBJ.PERSON	-	0	+
SUBJ.PLACE	0	0	0
SUBJ.PSYCH_STATE	0	0	0
SUBJ.QUALITY	0	0	0
SUBJ.STATE	+	0	0
SUBJ.SUBSTANCE	0	0	0
SUBJ.TEXT	+	0	-
SUBJ.THING	0	0	0
PO.ACTIVITY	+	-	0
PO.COGNITION	0	0	0
PO.COMMUNICATION	0	0	0
PO.EVENT	0	-	0

PO.INSTITUTION	-	0	+
PO.INSTRUMENT	0	-	0
PO.NAME	-	+	0
PO.NONE	0	-	+
PO.PEOPLE	0	0	0
PO.PERSON	0	0	0
PO.PLACE	0	0	0
PO.PSYCH_STATE	0	0	0
PO.QUALITY	0	0	0
PO.STATE	0	0	0
PO.SUBSTANCE	-	+	0
PO.TEXT	+	0	0
PO.THING	0	0	+
PO.TIME	0	0	0
PO.TYPE	0	+	-

Table 1. Impact of conceptual subtypes of subject (SUBJ) and following prepositional phrase objects (PO) on the selection of one of the three passive auxiliary construction in the COCA sample: +: significant odds in favor of auxiliary; –: significant odds against auxiliary; 0: non-significant odds.

Comparing this American English model against the British English one (Arppe 2011) is not the most straight-forward task as the semantic/conceptual subgroup coding schemes are not exactly the same. Nevertheless, we can for instance note that in both American and British English an ACTIVITY, AMOUNT (~POSSESSION), or INSTRUMENT (~ARTEFACT) or form of COGNITION or COMMUNICATION (~TEXT) as a PATIENT, i.e. grammatical subject of the passive constructions, are significantly in favor of the *be* auxiliary, as is the case with PERSON or PEOPLE in relation to the *get* auxiliary. However, we are not able to replicate the association of (HUMAN) GROUPs as significant subjects of the *become* passive auxiliary.

4 Conclusions

As an interim conclusion we may claim we have a good start in systematically modelling the contextual factors that drive the selection of any one of the three alternative passive auxiliary constructions in British English.

References

Arppe, A. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy*. Publications of the Department of General Linguistics, University of Helsinki, No. 44.

Arppe, A. 2012. polytomous: *Polytomous logistic regression for fixed and mixed effects*. R package version 0.1.4.

Arppe, Antti (2011). "From modeling lexical synonyms to constructional alternations." *Corpus Linguistics 2011 Conference* (CL2011), 20-22.7.2011, University of Birmingham, UK.

Arppe, A. & P. Bolger (2012). "Grammar triangulated – contrasting frequency, acceptability and processing." *Workshop on Grammar between gradience and frequency*, Annual Conference of DGfS (Deutsche Gesellschaft für Sprachwissenschaft), Frankfurt am Main, Germany, 6-9 March 2012.

Arppe, A., P. Bolger & D. Dowbor (2012). "The more evidential diversity, the merrier – contrasting linguistic data on frequency, selection, acceptability and processing". New Ways of Analyzing Syntactic Variation (NWASV), 15-17 November 2012, Nijmegen, the Netherlands.

Bader, M. & J. Häussler (unpublished manuscript). The primacy of grammaticality.

Bresnan, J., A. Cueni, T. Nikitina, & R. Harald Baayen. (2007). "Predicting the Dative Alternation." In: *Cognitive Foundations of Interpretation*. Boume, G., Kraemer, I. & J. Zwarts. Royal Netherlands Academy of Science, Amsterdam, the Netherlands, pp. 69-94.

Dowbor, D., A. Arppe & S. Rice (2012). "The English Passive Under Corpus Scrutiny: Some Cognitive Models of Transitivity Revisited." *Conceptual Structure, Discourse, and Language* (CSDL-11), 17-20 May 2012, Vancouver,

Canada.Gries, S. Th. (2002). "Evidence in linguistics: Three approaches to genitives in English." In: Brend, R. M., W. J. Sullivan and A. R. Lommel (eds.), LACUS Forum XXVIII: *What Constitutes Evidence in Linguistics?* Fullerton: LACUS, pp. 17–31.

Gries, S. Th. (2003a). *Multifactorial analysis in corpus linguistics: a study of particle placement*. Continuum, London, England.

Gries, S. Th. (2003b). "Towards a corpus-based identification of prototypical instances of constructions." *Annual Review of Cognitive Linguistics*, Vol. 1, pp. 1-27

Rosenbach, A. (2003). "Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English." In: Rohdenburg, G. and B. Mondorf (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, pp. 379–411.

An open-access gold-standard multi-annotated corpus with huge user-base and impact: The Quran

Eric Atwell, Nora Abbas, Bayan AbuShawar, Claire Brierley, Kais Dukes, Majdi Sawalha, and AbdulBaquee Muhammad Sharaf
University of Leeds
e.s.atwell@leeds.ac.uk

1 The Quran for corpus linguistics

The Quran has many advantages as a dataset for Corpus Linguistics research. Linguistics touches many domains and applications; but in general CL research involves analysis of a domain-specific corpus of text documents enriched with linguistic and semantic tags. Ideally, we want a domain where: a source text corpus is freely available as plain text and also with linguistic annotations, with no IPR or privacy restrictions; a large expert community exists, which has already developed standard "tagging schemes" or ontologies for the domain; and a large user group exists, to assist with linguistic and semantic tagging, and to evaluate our systems and results, and also to use the text analytics tools we deliver, so that our research has impact. The corpus which best meets these research criteria is the Quran. The source Classical Arabic text is freely available, both raw text and a range of tagged and annotated versions; Quran scholars over the past thousand years have developed a rich tradition of Arabic linguistics to formally describe and tag the language and meaning of the Quran; and billions of Muslims worldwide constitute the largest user-group ever for a single source text corpus – other globally-read texts such as the Bible or Harry Potter are more widely accessed in a range of translations than their original source texts.

In this paper, we present a range of research results and resources which illustrate the attractions of the Quran as a dataset for corpus linguistics.

2 The Quranic Arabic Corpus

The Quranic Arabic Corpus (Dukes, Atwell and Habash, 2011) is a multimodal language resource that integrates deep morphological and syntactic tagging based on traditional Quranic Arabic grammar textbooks, interlinear word-by-word English translation as well as multiple aligned verse-by-verse English translations, multiple speech recordings, and resources for visualization and collaborative analysis for the Classical Arabic language of the Quran. Freely available online at **http://corpus.quran.com**, the website is a widely-used research resource and also a popular study guide for Quranic Arabic, used by over 1.2 million visitors over the past year. (Dukes and Atwell 2012) provide a description of the underlying software system that has been used to develop the corpus annotations. The multimodal data is made available online through an accessible cross-referenced web interface.

3 Text Mining the Quran

(Sharaf and Atwell 2012a,b) describe tools and resources for text mining the Quran including QurSim verse similarity dataset, and QurAna anaphoric co-reference markup and named entity markup.

QurSim is a dataset derived from the Quran, in which semantically similar or related verses are linked together. This will be a valuable evaluation resource for computational linguists investigating similarity and relatedness in short texts. Furthermore, this dataset can be used for evaluation of paraphrase analysis and machine translation tasks. QurSim is characterised by: (1) superior quality of relatedness assignment; as we have incorporated relations marked by highly-reputed Quran experts, this markup provides a gold standard for various evaluation tasks, (2) the size of our dataset: over 7,600 pairs of related verses are collected from scholarly sources with several levels of degree of relatedness. This dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs. QurSim is incorporated into online query pages where users can visualize for a given verse a network of all directly and indirectly related verses. Empirical experiments showed that only 33% of related pairs shared root words, demonstrating the need to go beyond common lexical matching methods, and incorporate semantic, domain knowledge, and other corpus-based approaches.

QurAna is a large dataset created from the original Quran text, where personal pronouns are tagged with their antecedents. These antecedents are maintained as an ontological list of concepts, useful for information extraction tasks. QurAna is characterized by: (a) the large number of pronouns tagged with antecedent information (over 24,500 pronouns), and (b) maintenance of an ontological concept list out of these antecedents. This annotated dataset could benefit researchers in obtaining empirical rules in building new anaphora resolution approaches.

4 Boundary-Annotated Quran

(Brierley et la 2012), (Sawalha et al 2012) report on the Quran Boundary-Annotated Qur'an dataset of 77430 words and 8230 sentences, where each word is tagged with prosodic and syntactic information at two coarse-grained levels. A boundary-annotated and part-of-speech tagged corpus is a prerequisite for developing phrase break classifiers. Boundary annotations in most other speech corpora are descriptive, delimiting intonation units perceived by the listener. We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from Tajwīd (recitation) mark-up included in the traditional source text of the Quran, which we then interpret as additional text-based data for computational analysis. This mark-up is prescriptive, and signifies a widely-used recitation style.

5 Qurany 'Search for a Concept' Tool

(Abbas 2009) presents Qurany, a bilingual English/Arabic information extraction tool for the Quran that significantly enhances recall and precision when searching for concrete and abstract concepts. Keyword search is extended to synonyms used in any of eight parallel English translations of the Quran as well as the original Arabic, and search for lemmas and morphemes. Qurany also offers users a comprehensive hierarchical classification of Quran abstract topics or themes using expert knowledge imported from 'Mushaf Al Tajweed', a highly-reputed classical commentary which provides in effect an ontology of the Quran, via an index of topics which covers nearly 1100 concepts in the Quran. The Qurany ontology-enriched dataset is also available as a Google-searchable website, for example search via Google for prayers site[1] .

6 Quran Gold Standard for Evaluating Arabic Morphological Analyzers

Sawalha and Atwell (2008, 2013) compare several morphological analysers and morphological tagging schemes for Arabic corpora. They found that the range of analysers and schemes had been demonstrated on different corpora, making direct comparisons difficult; so, they used a chapter from the Quran as a gold standard text for evaluating Arabic morphological analysers. This inspired their SALMA Standard Arabic Linguistics Morphological Analysis tag set expounding traditional fine-grained morphological features, applied to a chapter of the Quran in annotation of a Gold Standard SALMA-tagged sample corpus.

[1] http://www.comp.leeds.ac.uk/nora/html/

7 Using the Quran to train an Machine-Learning Chatbot

A chatbot is a machine conversation system which interacts with human users via natural conversational language. (AbuShawar and Atwell 2004, 2005a,b) present software to Machine-Learn conversational patterns from a transcribed spoken corpus, and use these patterns to generate a range of chatbots speaking various languages and sublanguages learnt from spoken transcripts, for example from the British National Corpus of English, the Minnesota French Corpus, and the Corpus of Spoken Afrikaans amongst others. These versions of the chatbot demonstrated the generality and adaptability of the chatbot-training algorithms, but had little use beyond that. However, when the algorithms were trained using the Quran, this produced our first chatbot to find a practical use: in Muslim school Quran lessons, the Quran chatbot provided an interesting and unusual teaching aid for pupils learning to recite the Quran.

8 Conclusions

This paper presents a range of research and resources to illustrate the use of the Quran for corpus linguistics research. We are able to capitalise on a rich Islamic tradition of Quranic Arabic analysis and grammar, developed over centuries to enable non-Arabic-speaking Muslims to access the Classical Arabic source text. And we are able to find practical uses and users of our results, outside the Corpus Linguistics community. This helps motivate us, and also helps to persuade research funding agencies that Corpus Linguistics can have significant "Impact". We have advocated (Atwell et al 2010) that understanding the Quran should be a new Grand Challenge for Computer Science and Artificial Intelligence; the Quran offers a grand challenge but also an outstanding resource for Corpus Linguistics.

References

Abbas, N. 2009. *Quran 'Search for a Concept' Tool and Website*. Unpublished thesis, University of Leeds.

Abu Shawar, B; Atwell, E. 2005. A chatbot system as a tool to animate a corpus. *ICAME Journal*, vol. 29, pp. 5-24

Abu Shawar, B.; Atwell, E.2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, vol. 10, pp. 489-516.

Abu Shawar, B; Atwell, E. 2004. An Arabic chatbot giving answers from the Qur'an. Bel, B and Marlien, I (editors) *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*, Volume 2, pp. 197-202 ATALA. Fez, Morocco.

Atwell, E; Dukes, K; Sharaf, A; Habash, N; Louw, B;

Abu Shawar, B; McEnery, T; Zaghouani, W; El-Haj, M. 2010. Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence. *GCCR'2010 Grand Challenges in Computing Research*. Edinburgh, Scotland

Brierley, C; Sawalha, M; Atwell, E. 2012. Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing. *LREC'2012 Language Resources and Evaluation Conference.* Istanbul, Turkey

Dukes, K; Atwell, E. 2012. LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. *LREC'2012 Language Resources and Evaluation Conference.* Istanbul, Turkey.

Dukes, K; Atwell, E, Habash, N. 2011. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. *Language Resources and Evaluation Journal*, pp.1-30.

Sawalha M, Atwell E. 2013. A standard tag set expounding traditional morphological features for Arabic language Part-of-Speech tagging. *Word Structure Journal*, to appear.

Sawalha, M; Brierley, C; Atwell, E. 2012. Predicting Phrase Breaks in Classical and Modern Standard Arabic Text. *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Sawalha, M; Atwell, E. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers. *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics.* Manchester, England.

Sharaf, A; Atwell, E. 2012a. QurSim: A corpus for evaluation of relatedness in short texts. *LREC'2012 Language Resources and Evaluation Conference.* Istanbul, Turkey.

Sharaf, A; Atwell, E. 2012b. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Triangulating levels of focus and the analysis of personal adverts on Craigslist

Paul Baker
Lancaster University
p.baker@lancaster.ac.uk

This paper evaluates the effectiveness of triangulating methods when comparing three small corpora of personal adverts found online. Cohen and Manion (2000: 254) say that triangulation is an 'attempt to map out, or explain more fully, the richness and complexity of human behavior by studying it from more than one standpoint' while Altrichter et al. (2008: 147) argue that this approach 'gives a more detailed and balanced picture of the situation'.

Three small corpora (42,000-76,000 words) were collected from the India, Singapore and Australia subsections of the website Craigslist. All adverts were placed in the category 'men seeking women'. Personal adverts have the capacity to tell us something about constructions of gender and desire, and the extent to which such constructions relate to societal restrictions or norms. It was decided to use geographic location as a main variable in order to determine the extent to which the sorts of linguistic patterns (and related gendered discourses, Sunderland 2004) were dependent on culture or society, or whether there were indications that certain discourses appeared to be more global in nature. To quote from a well-known gendered discourse (which is also the title of a song by the group The Used released in 2009), is it true that 'men are all the same'?

The research questions set were 1) how do males relate to gender when advertising to meet women on Craigslist, and 2) to what extent are there differences or similarities between advertisers from the different countries? Three separate approaches were taken. First, using Wmatrix (Rayson 2008), the words in the three corpora were assigned semantic tags. Key semantic tags were identified by comparing pairs of corpora together. Tags that were key across two comparisons were viewed as particularly salient. For example, when the Australian corpus was compared against the Singaporean corpus, one tag that was key for Singapore was F1 (Food). This tag was also key for Singapore when it was compared against India, which gives an indication that the F1 tag seems to be strongly associated with the Singaporean adverts (compared to the other two contexts). On the other hand, another tag which was key for Singapore when compared against Australia was L1- (Dead). However, this tag was not key when Singapore was

compared with India, so L1- was not analysed as the difference did not occur across the two comparisons that involved Singapore. Consistently, key tags were subjected to concordance analyses in order to gain a better understanding of why advertisers used them and how this related to gendered discourses of male identity and desire.

Secondly, it was decided to examine combinations of adjectives that the advertisers used to describe themselves. It was hypothesised that adjective choice would reveal the types of traits that male advertisers felt would be valuable and likely to elicit replies from women, which would subsequently be indicative of gendered discourses. The twenty most frequent self-describing adjectives were elicited for each corpus and then grouped into related sets e.g. personality, physical appearance, ethnicity, age etc. Chi-squared tests were used in order to compare the sets of traits across the three corpora. Additionally, collocational networks of the top 20 collocates were calculated for each corpus in order to gain a better idea about how different sets of adjectives were used together, sometimes to reinforce a particular identity construction.

Finally, in order to focus more closely on gendered discourses across the corpora, it was decided to carry out analyses of expanded concordance lines of a small set of words relating to women (*woman*, *lady*, *girl*, *female* and their plurals). Cases where advertisers made generalising comments about women rather than specific ones were given particular analytical focus, especially when they referred to the notion of a generic woman. Additionally, adverts where the advertiser expressly referred to societal norms or adverts which related women to men in some way also proved to be a good 'site' for the identification of gendered discourses. Similar cases were grouped together in order in an attempt to identify gendered discourses.

The paper concludes with an evaluation of the efficacy of the three methods, particularly focusing on whether this form of triangulation is adept at confirming previously found results (e.g. the extent to which all three methods produced similar findings), or whether each method resulted in completely different answers to the research questions.

References

Altrichter, H., Feldman, A., Posch, P. & Somekh, B. 2008. *Teachers investigate their work; An introduction to action research across the professions.* London: Routledge.

Cohen, L. and Manion, L. 2000. *Research methods in education*. London: Routledge.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics.* 13 (4): 519-549.

Sunderland, J. (2004) *Gendered Discourses*. London: Palgrave.

Robust corpus architecture: a new look at virtual collections and data access

Piotr Bański
IDS Mannheim / University of Warsaw
banski@ids-mannheim.de

Elena Frick
IDS Mannheim
frick@ids-mannheim.de

Michael Hanl
IDS Mannheim
hanl@ids-mannheim.de

Marc Kupietz
IDS Mannheim
kupietz@ids-mannheim.de

Carsten Schnober
IDS Mannheim
schnober@ids-mannheim.de

Andreas Witt
IDS Mannheim
witt@ids-mannheim.de

1 Introduction

The present contribution has two logical components: the first presents the basic aims and design principles of KorAP – a new corpus analysis platform that is being developed at the Institut für Deutsche Sprache in Mannheim. In the second part, we concentrate on two closely related issues that have arisen in the process of the development of the internal data architecture for KorAP but have consequences for innovative corpus design in general. These issues prompt a reformulation of the definition of the concept of virtual collections and a new assessment of the consequences of this reformulated concept for the practical considerations of access permissions and security in general.

2 Aims

KorAP (Korpusanalyseplattform der nächsten Generation, cf. Bański et al. 2012), currently in the prototype phase, is an innovative corpus analysis platform designed to address the demands of modern linguistic research. The platform is intended to facilitate new linguistic findings by making it possible to manage and analyse primary data and annotations in, eventually, the petabyte range, while at the same time fully satisfying the demands of a scientific tool, by both allowing an undistorted view of the primary linguistic data, and giving equal status to the various possible analyses of those data: in KorAP documents, the primary ("raw") text is physically separated from all its possible interpretations, from which the user may choose and which she may compare (this is an example of radical stand-off architecture, similar to that used in the American National Corpus, cf. Ide and Suderman, 2006).

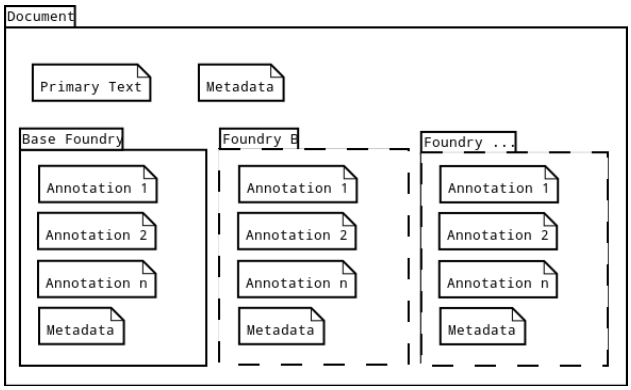


Figure 1. KorAP data model, where the primary ("raw") text is separated from annotations, organized into "foundries".

An additional important aim of the project is to make corpus data as openly accessible as possible in light of unavoidable legal restrictions, for instance by providing a sandbox that enables users to apply their own tools, working on data that cannot be released, or by supporting distributed virtual collections (Kupietz et al. 2010). The KorAP software itself will be released under an open license.

3 Virtual collections and security

The term "virtual collections" was first introduced, i.e. imported from the context of digital libraries, by the D-SPIN project (the former name of the German part of CLARIN, cf. Bankhardt 2009) as a generalization of "virtual corpora" and as a fundamental concept for the development of a standardized way to persistently identify research data consisting of language resources, in order to facilitate the implementation of maxims such as replicability, in linguistics and adjacent disciplines.

It has been further elaborated within the CLARIN project, also in the form of a basic implementation of a registry for virtual collections ("CLARIN-VCR")[1], and has been characterized as "distributed collections of corpora or data" (Broeder et al. 2007) that can "include a large number of resources created by different teams at different institutions" (ISO 24619:2011). An illustration of one practical implementation of the concept is provided by DeReKo (Deutsches Referenzkorpus, Kupietz et al. 2010), cf. Figure 2. Whenever applied to corpora, this term has usually denoted some amount of text accompanied by a single instance of grammatical description, most often inline – embedded in the text by means of XML elements or attributes.

[1] See http://clarin.ids-mannheim.de/vcr/app/public

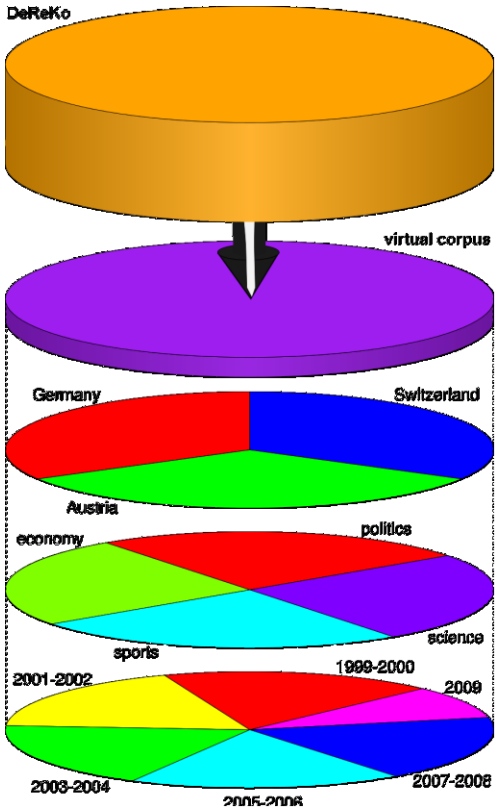


Figure 2. Illustration of DeReKo (Deutsches Referenzkorpus) seen as a "primordial sample", from which individual virtual collections can be created according to metadata-based choices, in this case: country of origin, topic, and time period. Copied from Kupietz et al. (2010).

This by now widely adopted characterisation of virtual collections and the current state of the CLARIN-VCR have proven insufficient for a tool such as KorAP, which explicitly allows for many concurrent annotation layers to provide equally valid points of view on the underlying data. KorAP explicitly distinguishes the primary ("raw") text of the individual corpus documents from its linguistic descriptions, even down to the level of supporting multiple tokenizations, with hierarchies of annotation layers built upon each tokenization stream.

KorAP's robust support for metadata describing not only the primary text but also each annotation layer, together with the traditional concept of virtual collections, result in multi-faceted virtual collections that, apart from texts, can combine e.g. annotations produced by the same tool or within the same school of linguistic thought, including collections of resources bearing the same distribution licenses. This is illustrated in Fig. 3.

As the complexity of linguistic resources increases, assigning permissions directly to users becomes an administrative challenge and as a consequence poses high security risks. To deal with such security issues and improve administration and performance, existing corpus analysis systems and modern research platforms such as TextGrid (TextGrid-Konsortium 2009: 18f) opt most often for role- and/or group-based approaches (MPI 2006: 83f.). Although these concepts offer graded access control by allowing grouping to corpora and virtual collections, they lack flexibility to handle KorAP's demands for access control. KorAP's ability to separate the raw text from its various grammatical representation results in a large variety of collections that can be created.

Furthermore, KorAP has been designed to support

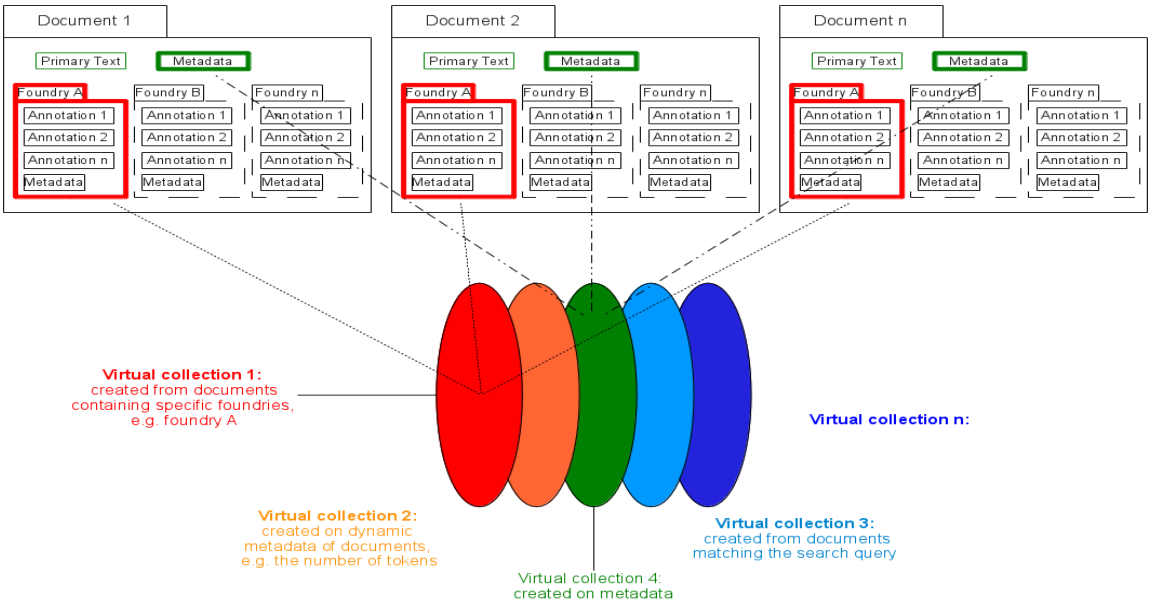


Figure 3. A sample of ways in which KorAP virtual collections can be created

user-supplied corpora, which imposes additional requirements on an access control system: reliance on mappings to roles does not allow for the envisioned definition of fine-grained access control policies, and a group-based approach would result in massive overkill.

References

Bankhardt (2009): D-Spin – Eine Infrastruktur für deutsche Sprachressourcen. In: Sprachreport 1/2009. S. 30-31 – Mannheim: Institut für Deutsche Sprache, 2009. (Sprachreport 1/2009)

Bański, P. Fischer, P.M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, C., and Witt, A. 2012. The new IDS corpus analysis platform: Challenges and prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul.

Broeder, D., Declerck, Th., Kemps-Snijders, M., Keibel, H., Kupietz, M., Lemnitzer, L., Witt, A., Wittenburg, P. (2007): Citation of Electronic Resources: proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Dokument N366.

Ide, N., Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. In Proceedings of the Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy.

ISO 24619:2011. Language Resource Management – Persistent Identification and sustainable Access in Language Technology Applications. (PISA). Technical report. International Organization for Standardization.

Kupietz, M/Belica, C/Keibel, H/Witt, | (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010), S. 1848-1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf

Kupietz, M/Bankhard, C (eds.) (2009): D-SPIN Report R7.1 – Legal Aspects in the Provision of Language Resources: The German Context. (http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.1.pdf)

Kupietz, M/Bankhard, C (eds.) (2010): D-SPIN Report R7.3 – Initial Localisation of CLARIN Best Practices and Business Models. (http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.3.pdf)

Max-Planck-Institute for Psycholinguistics (MPI). (2006). DAM-LR Distributed Access Management for Language Resources – Deliverable 8.1 Definition Report (pp. 1–92). Nijmegen. (http://www.mpi.nl/dam-lr/documents.html)

TextGrid-Konsortium (2009): Abschlussbericht – Öffentliche Fassung. http://www.textgrid.de/fileadmin/berichte-1/abschlussbericht-1.pdf

Exemplar theory and patterns of production

Michael Barlow
University of Auckland
mi.barlow@auckland.ac.nz

1 Introduction

Corpus analyses are used to provide insights into grammar from a cognitive or social perspective. One strand of cognitive approaches is associated with usage-based approaches (Langacker 1987) in which cognitive representations of language are closely linked to usage: language production and comprehension.

Since corpora are typically an amalgamation of the speech or writing of many different individuals, the patterns of production of individuals are hard to discern (Mollin 2009). In this paper, we review some data on individual differences in spoken production and explore how these differences might be explained within an Exemplar theory framework.

2 Corpus data

For this study, the speech of four White House press secretaries is examined. These are Mike McCurry (1994-98), Ari Fleischer (2001-03), Scott McClellan (2003-06), and Robert Gibbs (2008-2011). The speech samples are taken from transcripts, which, are sufficiently accurate representations of the spoken interactions for the purposes of the present study. The amount of speech transcribed is considerable and here we will work with individual speech corpora consisting of three samples of 200,000 words of running text per speaker. An important advantage of working with this particular dataset is that the general context of the discourse is held constant across the different samples. The content changes day by day, but the overall format of press conferences varies very little and almost all the interactions consist of the press secretary answering questions posed by members of the media.

To highlight the distinction between single-speaker and the mixed-speaker samples usually associated with corpus studies, we compile a mixed-speaker corpus from the White House press conferences. Six new files each containing 200,000 words from the White House press conference data are created. In this case, however, we use a complementary data set that excludes the contribution of the press secretary and instead consists of the speech of members of the press corps and White House officials.

To recap: we are working with three 200,000 word samples of the speech of each press secretary. In the case of Mike McCurry, we actually have enough data for six samples and we will label the triples as Mike1 and Mike2. For the mixed-speech samples, we also take six samples and in this case we randomly combine the samples to create two batches: XMix1 and XMix2. It is necessary to work with more than one sample in order to get information on variation in usage over time.

3 Results

It is clear that the role of the media representatives differs from that of the White House press secretaries and this will have a considerable impact on so it may well not be appropriate to assign any significance to the greater or lesser use of linguistic expressions by the press corps (as a group of mixed speakers) compared with the press secretaries. What we are focusing on is the nature of the variation among samples of files consisting of the language of multiple speakers in comparison with the inter-speaker variation. Thus these mixed speaker samples provide a control that indicates the level of variation you would expect to see from one typical corpus sample to the next.

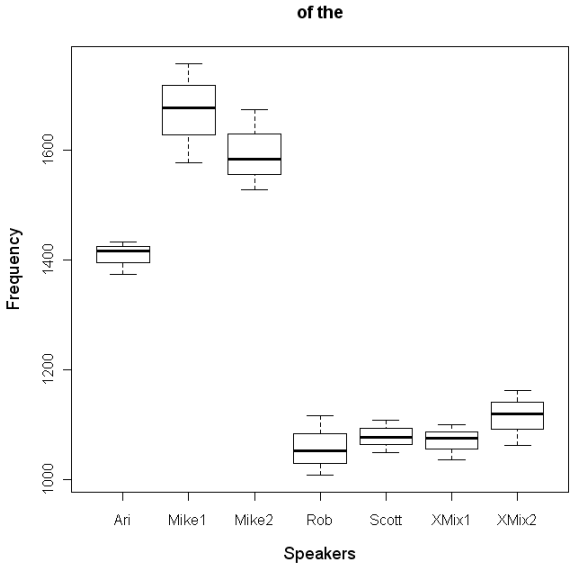


Figure 1. Use of *of the* by Different Speakers

Here we give the results of the frequency of use of two bigrams *of the* and *that we*, as shown in Figures 1 and 2. The data are representative and similar patterns occur with other bigrams, constructions, tenses, negation, etc.

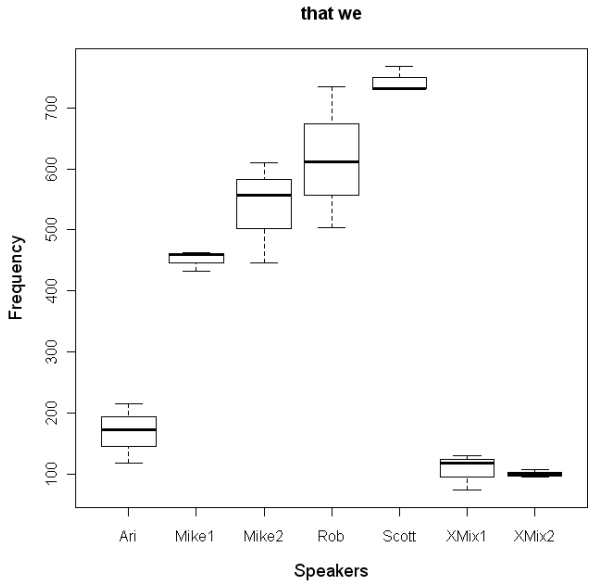


Figure 2. Use of *that we* by Different Speakers

We can see from the box plots in these figures that there is little variation in the mixed speaker samples. This is representative of the results obtained from most corpus analyses in which individual variation is washed out. The box plots also show that the individual differences are quite marked and are stable over a period of about a year. For example, Rob and Scott are low users of *of the* --- around 1100 instances per 200,000 words; whereas Mike is a high user: 1600 instances per 200,000 words. In the case of that we, Scott is a high user with 700 instances and Ari is a low user at around 200 instances per 200,000 words.

4 Exemplar theory

Taking these results into account, we see that for any particular speaker the patterns of production differ from the patterns of comprehension and hence the relation between usage and grammar is more complex than we might have thought.

One perspective on usage-based theory is an exemplar-based approach (Pierrehumbert 2001, Bybee 2006, Hay and Bresnan 2006), which might offer a way of explaining the data, as discussed below.

Walsh et al 2010: state:

> Central to Exemplar Theory are the notions of frequency, recency, and similarity. Extensive storage of language input exemplars takes place, categorization of input is made by comparison with extant exemplar memory traces, production is facilitated by accessing these stored exemplars, and the exemplar memory is in a constant state of flux with new inputs updating it and old unused exemplars gradually fading from memory.

Various issues concerning the degree of abstraction and the decay of memory of instances remain to be resolved and different models of exemplar theory have been proposed (Bybee 2006, Snider 2008, Bod 2006).

Hay and Bresnan (2006) discuss phonetic and syntactic exemplar theories and link exemplars to contextual information:

> individual exemplars are not only phonetically rich, but are also indexed with a variety of social information (the identity of the individual, their gender, regional origin, approximate age, what they are wearing, their hairstyle …, anything that could be perceived as sociolinguistically or sociologically relevant)

Hay and Bresnan suggest that syntactic memories could also be socially indexed, which means that units larger than the word are linked with appropriate social variables.

Exemplar theory is to a large extent a data-driven theory of grammar with a focus on the establishment of grammatical "categories" using exemplars. Thus the main focus is on comprehension and "grammar building" rather than production. What we have seen above is that the patterns of instances of a category resulting from comprehension may be different from those associated with production and an elaborated theory is needed.

Exemplar saliency index

$$\begin{bmatrix} \textit{interlocutor(s)} & \begin{bmatrix} \textit{name} & \textit{John} \\ \textit{gender} & \textit{male} \\ \textit{role} & \textit{reporter} \\ \textit{demeanour} & \textit{hostile} \\ \dots & \dots \end{bmatrix} \\ \textit{speaker} & \begin{bmatrix} \textit{SELF (I/me)} \end{bmatrix} \\ \textit{identity} & \begin{bmatrix} \textit{spokesperson} \\ \textit{insider} \\ \textit{expert} \\ \dots \end{bmatrix} \\ \textit{purpose} & \begin{bmatrix} \textit{deflect criticism} \\ \textit{maintain status} \end{bmatrix} \\ \textit{setting} & \begin{bmatrix} \textit{White House briefing room} \end{bmatrix} \\ \dots & \end{bmatrix}$$

Figure 3. The Saliency Index

5 Saliency index

We posit a saliency frame that represents the kinds of information linked to particular language expressions. The structure in Figure 3 is not meant to be a static structure; it simply represents the kind of information relevant to a press secretary taking part in a press conference and facing questions from

the media. This frame includes not only social information but also information on intentions or other relevant functional and cognitive aspects of the discourse.

One important aspect of the frame relevant here is that the speaker is essentially tracking their own production along with contextual and other variables. This is represented by the speaker variable, which in this example is linked to SELF.

References

Bod, Rens 2006. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review* 23. 291-320

Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language* 82, 4.

Hay, J. and Bresnan, J. 2006. 'Spoken syntax: The phonetics of giving a hand in New Zealand English. '*The Linguistic Review 23: Special Issue on Exemplar-Based Models in Linguistics*, 321-349.

Langacker, R. 1988. A usage-based model. In Rudzka-Ostyn, B. (ed), *Topics in Cognitive Linguistics*, 127-61. Amsterdam: Benjamins.

Mollin, Sandra. 2009. "'I entirely understand' is a Blairism: The methodology of identifying idiolectal collocations." *International Journal of Corpus Linguistics*, 14:3, 330-355.

Pierrehumbert, J. B. 2001. Exemplar dynamics: Word frequency, lenition and contrast, in Bybee, J. and Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure*, John Benjamins, Amsterdam, 137-157

Snider, Neal. 2008. *An Exemplar Model of Syntactic Priming.* Unpublished Doctoral Dissertation. Stanford University

Walsh, Michael, Wade, Travis, Möbius, Bernd, & Schütze, Hinrich. 2010. Multi-level exemplar theory. *Cognitive Science 34:4, 537-582*

The construction of otherness in the public domain: a CDA approach to the study of minorities in Ireland

Leanne Bartley
Universidad de Granada
lbartley@ugr.es

E. Hidalgo Tenorio
Universidad de Granada
ehidalgo@ugr.es

1 Introduction

In the last two decades, there have been significant changes across Europe politically and economically speaking. Ireland is no exception. The Celtic Tiger started to roar, literally and metaphorically, in 1994, when economist Kevin Gardiner made use of this expression with the intention of comparing the growth of the Republic with examples of a similar situation present years before in Hong Kong, Singapore, South Korea and Taiwan.

Curiously enough, at that stage of economic expansion, the seed was planted for a country which had suffered the devastating effects of the British Empire's colonization only to fall into the trap of reproducing the same patterns that led to the affliction of other colonized people in the past (Said 1995): those allegedly considered different, inferior, and socially, economically and/or morally dangerous were excluded, however, in a somewhat diverse society consisting of many other marginalised minorities. Consequently, the representation of these vulnerable groups, who share the label of "the Other", ends up being more than relevant given that it can help us detect the bigoted viewpoints held by many about those who are underrepresented or misrepresented in discourse. Furthermore, this enables the analysis of the construction of one's own identity, and how one often assumes we are entitled to judge others. On the one hand, those categorized under "otherness" are associated with illness, underdevelopment, ignorance and crime; for that reason, they are demonized discursively, and, as such, become society's scapegoats. On the other hand, "otherness" can also generate discourses of pity and compassion, although arguably proves a less common reaction from exposure to discrimination. Either way, the people belonging to a minority usually have to endure the stigma attached to certain stereotypes. The ideas the vast majority commonly accept are based on mental models that allow prejudice to be considered as the grounds for normalcy. The identification of discursive strategies that naturalise prejudiced thought is one of the objectives of this panel.

During a time when those more conservative

tended to win elections, often by a landslide, the spreading of racist and anti-multicultural discourse only backed the feelings of fear and rejection encouraged by the elite with regards to difference. Then, some representatives of parties such as Fianna Fáil and Fine Gael had no problem in expressing their convictions against immigration policies, and achieved a sufficient number of votes to return to Parliament. Although the extreme right did not receive a massive amount of support, racism is still a problem in Ireland. Sectarianism and resentment have become more prominent and, accordingly, a cause-effect relationship, although inaccurate, has been established, for example, between immigration and recession. Attacks on ethnic minorities, as much physical as symbolic, have dramatically increased in the last ten years.

In 2008, 183 people were arrested and charged, and it is believed there were more but victims kept quiet about anyone else involved.[1] Other evidence of this is intimidation often suffered by the Asian community and those from Eastern Europe and the Balkans; the introduction of medical centres for traveller;[2] or the harassment young gays and lesbians are subject to in schools in Dublin. The graffiti encouraging foreigners to be thrown out of the country is an anecdote by comparison with the killing of 15 year-old Nigerian teenager, Toyosi Shittabey, in Tyrrelstown.[3] This open act of violence proves the degree of extreme tension reached in Ireland as a result of this new ideological panorama in which hate towards marginalized groups is now clearer than ever.

2 Theoretical framework

This is the context of this research, which will focus on the main tenets developed in the rather heterogeneous theoretical framework of CDA (Wodak and Chilton 2005). We will examine how people's values and beliefs help create their perception of the world, the events taking place in that world, and the participants involved in various social practices. With the aim of analyzing the image different ideologies leave behind in discourse, and, alternatively, to reveal hidden identities through an individual's textual construction, we will study the modality patterns in the corpus we have selected, after first reconsidering the notion itself.

Prejudice is unavoidable in language; through its use, it is possible to perpetuate the same ideas shared among individuals belonging to various social groups and networks. For that reason, there is a need to analyse the strategies used by the subjects in our investigation in order to adequately configure their discursive universe. It is through others' observation of what we say and how we say it that we develop as individuals. The verbal choices we make explicitly (or we are barely conscious of) serve to identify us. This was something that was determined by stylisticians in East-Anglia in the 1970s (Fowler 1986), and cropped up later in critical linguistics and CDA (Fairclough 2003).

The act of representation of subjects is a privilege for those holding symbolic power and exercising dominance in a given context. Those who, due to their status, have access to this are able to voice themselves in the process of constructing the dominant group's discourse; they decide who can and cannot be represented, and under which terms and conditions (Fairclough 1995). Highlighting certain attributes instead of others, hiding several participants in certain social situations, or mystifying actions or states in which one becomes involved can serve to privilege some and silence others. The social reality does not change as a result of the mere fact that individuals reorder their syntactic structures in a particular way; however, perceptions are most certainly modified, and that is reflected in how they are presented and how events are construed in texts. Lexical selection identifies who "the Other" is from our perspective. It also acts as evidence for who we appear to be, and what both our stance as well as our view of the world are.

With the above in mind, it is important to briefly mention the concept of modality, especially given its nature to act as a means of revealing one's personal perspective regarding a particular topic. Halliday (2004: 116) states that it embodies, on the whole, the likelihood of a proposition, or alternatively the degree of desirability of a proposal. Modality is a measurable concept occurring on a scale of two extremes, with one end of the scale pertaining to *yes* and the other to *no*. Along this scale, varying degrees of probability, obligation and inclination are present (ibid: 147), which leads us to a controversial issue tackled in this panel: the different types of modality that have been proposed and, furthermore, which need redefining: namely, *epistemic*, *deontic*, *evidential* and *dynamic* modalities (see Palmer 1986). The problematic notions here concern the two latter; *evidentiality*, defined as the speaker's assessment of an event, will depend upon the source of the information: direct evidence, reporting, inference (Marín Arrese 2004). The issue is that inference and direct evidence are essential to the

1. See http://www.irishexaminer.com/opinion/editorial/racism-in-ireland--abuse-is-an-affront-to-our-history-101659.html, "Racism in Ireland – Abuse is an affront to our history", 24 September 2009, *Irish Examiner*.
2. See http://flag.blackened.net/revolt/rbr/travrbr2.html "Racism in Ireland. Travellers fighting back".
3. See http://www.irishcentral.com/story/roots/ireland_calling/ireland-of-the-welcomes----and-the-racism-90834164.html "Ireland of the welcomes... and the racism", 14 April 2010, John Spain.

description of epistemic modality as well (Cornillie 2009). As for *dynamic* modality, two different aspects altogether are considered: volition and ability (Wärnsby 2006: 21). However, it must first be highlighted that, if modality reflects attitude and stance, it becomes difficult to account for capacity (Huddleston and Pullum 2002; Gisborne 2007). Moreover, we will attempt to establish whether dynamic modality is in fact a subcategory in its own right or whether volition is simply a sub-class of deontic modality, given that forces are involved in both cases.

3 Materials and method

For our purposes here we have resorted to LexisNexis, collecting 2,336 news articles from 3 Irish broadsheet and tabloid newspapers with over 1,500,000 words from 1997 until 2012. Taking into account the number of texts, the research presented in this panel will be carried out in the line of corpus linguistics (McEnery and Wilson 1996, Baker 2006). Wordsmith Tools 3.0 (Scott, 1999), AntConc 3.2.1.0 (Anthony 2010) or the Wmatrix platform (Rayson 2007) enable concordances to be carried out; and allow the examination of key words in context, as well as the study of semantic prosodies. Moreover, this is possible in such a way that the researcher can refer to empirical evidence, and not to mere intuition at the time of interpreting cultural elements they are faced with.

In the analysis phase, we will proceed with a qualitative approach which will benefit from a quantitative perspective. Thus, with the frequency lists generated, and subsequently through the use of concordance lists, we will classify explicit markers of modality that will show which relationship exists between minorities and Irish society (e.g. *We think immigrants / travellers / homosexuals should / have to…; immigrants / travellers / homosexuals are X but…*); later, statistical tests will be performed so that, if findings prove statistically significant, it will be possible to identify the strategic ideological bent in Ireland.

References

Anthony, L. 2010. AntConc 3.2.1.0. http://www.antlab.sci.waseda.ac.jp/software.html

Baker, P. 2006. *Using corpora in discourse analysis*. London: Continuum.

Cornillie, B. 2009. "Evidentiality and epistemic modality". *Functions of Language* 16 (1): 44-62.

Fairclough, N. 1995. *Critical discourse analysis*. London: Longman.

Fairclough, N. 2003. *Analysing discourse. Textual analysis for social research*. London & NY: Routledge.

Fowler, R. 1986. *Linguistic criticism.* Oxford: OUP.

Gisborne, N. 2007. "Dynamic modality". *SKASE Journal of Theoretical Linguistics* 4 (2): 44-61.

Halliday, M.A.K. 2004. *An introduction to functional grammar* (revised by Christian M.I.M. Matthiessen). London: Edward Arnold.

Huddleston, R.D. and Pullum, G.K. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Marín Arrese, J.I. (ed.) 2004. *Perspectives on evidentiality and modality.* Madrid: Editorial Complutense.

McEnery, T. and Wilson, A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Palmer, F.R. 1986. *Mood and modality*. Cambridge: CUP.

Rayson, P. 2007. "Wmatrix: a web–based corpus processing environment". Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/

Said, E. 1995. *Orientalism: Western conceptions of the Orient*. London/N.Y.: Penguin Books.

Scott, M. 1999. *Wordsmith tools 3.0* .Oxford: OUP.

Wärnsby, A. 2006. *(De)coding modality: the case of must, may, måste, and kan*. Lund Studies in English: Lund.

Wodak, R. and Chilton, P. (eds.) 2005. *A new Agenda in (critical) discourse analysis*. Amsterdam & Philadelphia: John Benjamins.

Exploring the Firthian notion of collocation

Sabine Bartsch
Technische Universität Darmstadt
bartsch@linglit.tu-darmstadt.de

Stefan Evert
FAU Erlangen-Nürnberg
stefan.evert@fau.de

1 Introduction: exploring the Firthian notion of collocation

Collocations are pervasive in natural language. According to Altenberg (1991: 128) "roughly 70% of the running words in the corpus form part of recurrent word combinations of some kind". The study of such word combinations in corpora of authentic language dates back to the earliest collocation studies by J. R. Firth (1957), who is commonly credited with introducing the concept within British Contextualism.

In contrast to most prior comparative evaluation studies, which focused on the extraction of lexicalised multiword expressions relevant for traditional paper dictionaries, our research takes as its point of departure a strictly Firthian (1951; 1957) definition of collocation as the habitual and recurrent juxtaposition of words with particular other words. Such a definition lends itself to the automatic, statistically based identification of collocations in large corpora.

Corpus-based exploration of collocations on a larger scale has become feasible with the availability of large electronic corpora, computational tools and procedures for their linguistic exploration as well as the necessary computing power in the late 20th c. Since then, a substantial number of corpora of different sizes and composition have become available, opening up new possibilities for the systematic study of collocation and many other linguistic phenomena. Progress has been made in particular by harnessing ever larger corpora, a wide range of increasingly sophisticated statistical measures of association (cf. Pecina 2005: 15), and state-of-the-art software tools for automatic linguistic annotation and analysis.

2 Research issues

The aim of the research presented in this paper is a critical evaluation of (i) the statistical techniques used for the automatic identification of collocations and multiword expressions and (ii) some tacit assumptions regarding suitable types of corpus data.

A first set of issues concern the statistical association measures that form the core of automatic collocation identification, such as log-likelihood (G^2), t-score (t), Mutual Information (MI), and the Dice coefficient to name but some of the most widely used ones (see Evert 2008 for details). Previous evaluation studies have focused on a comparison of different association measures for the automatic identification of lexicalised multiword expressions. To our knowledge, no evaluation study has specifically targeted the Firthian notion of collocation yet.

Moreover, the results of these studies are inconclusive: which association measure is most useful seems to depend on factors such as language, type of multiword expression as well as corpus size and composition. One common observation is that the MI score is biased towards low-frequency data and typically performs much worse than e.g. the log-likelihood ratio. The continuing popularity of MI among computational lexicographers suggests that the situation may be different if the aim is to identify collocations in a Firthian sense.

A second set of issues concern the question of suitable corpus size, where the tacit assumption – put bluntly – has always been that "bigger is better" for statistical approaches. However, no systematic exploration of the influence of corpus size and composition has been undertaken yet. It may well be that a smaller, but clean and balanced corpus is better suited for the identification of habitual general-language collocations than e.g. a large collection of Web pages or newspaper text.

Finally, a third set of issues concern the question of whether collocation studies can and will benefit from different kinds of corpus enrichment by means of annotation, i.e. whether collocations are best researched in unannotated plain-text corpora with little pre-processing (e.g. Sinclair 1991) or whether the task might not benefit from more abstract layers of annotation such as part of speech tagging or even syntactic parsing. At least some definitions of collocation strongly suggest this (e.g. Bartsch 2004) while a strictly Firthian notion of collocation does not seem to entail syntactic constraints.

3 Motivation for this research

The motivation for the research presented in this paper is to improve our understanding of the role that (i) the composition of the underlying corpus (ranging from clean, balanced reference corpora to huge, messy Web collections), (ii) automatic linguistic annotation (part-of-speech tagging, syntactic parsing, etc.), and (iii) the mathematical properties of statistical association measures play for the automatic identification of collocations from corpora.

In addition to improving techniques for the reliable automatic identification of collocations for

lexicographical purposes, our research was driven by three questions: Are bigger corpora always better, or is a balanced composition more important? To what extent does automatic linguistic annotation improve collocation identification? Can we find evidence for the postulated presence of syntactic relations between collocates (Bartsch 2004: 79), in contrast to the traditional window-based operationalisation (Sinclair 1966) of the Firthian notion of collocation?

4 Approach: corpora, annotation, scores

The research discussed in this paper is an attempt to jointly address these issues on the basis of a gold standard set of collocations which are lexicographically attested and which are used to evaluate candidates extracted from corpora of different sizes on the basis of different levels of annotation.

As a gold standard for our evaluation study, we used two-word collocations from 224 entries of the BBI Combinatory Dictionary (Benson et al. 1986), a pre-corpus collocation dictionary whose underlying concept of collocation comes very close to the Firthian definition. The 224 node words were chosen based on their frequency in various corpora and to include a range of words with interesting collocational behaviour. All lexical collocations (with nouns, verbs, adjectives and adverbs as collocates) found in these entries were extracted to form a gold standard of 2,949 known collocations.

Based on this gold standard, we have compared (i) a standard range of widely-used association measures (including log-likelihood, t-score, the Dice coefficient, co-occurrence frequency, and several variations of Mutual Information, (ii) corpora of widely different sizes and composition (ranging from the 100-million-word British National Corpus over a subset of the English Wikipedia to the Google Web 1T 5-Gram database compiled from 1 trillion words of English Web text), and (iii) different levels of automatic linguistic annotation (raw surface forms, POS tagging and lemmatisation, dependency parsing) and extraction methods (collocational spans of various sizes vs. syntactic relations). In order to ensure a fair comparison, candidates extracted from the different corpora were restricted to combinations of the 224 node words with a set of ca. 7,700 common lexical words. Random samples of the candidates were manually checked in order to ensure that the evaluation is not biased against e.g. novel or domain-specific collocations found in Web corpora but not listed in the BBI dictionary.

Following the methodology of Evert and Krenn (2001), evaluation is carried out by computing precision and recall of the *n* highest-ranked collocation candidates for each combination of association measure, base corpus, level of annotation and extraction method. Results for such *n*-best lists of various sizes can be visualised in the form of a precision-recall graph as shown in Fig. 1 for candidates extracted from the British National Corpus based on a collocational span of 3 words to the left and right of the node.

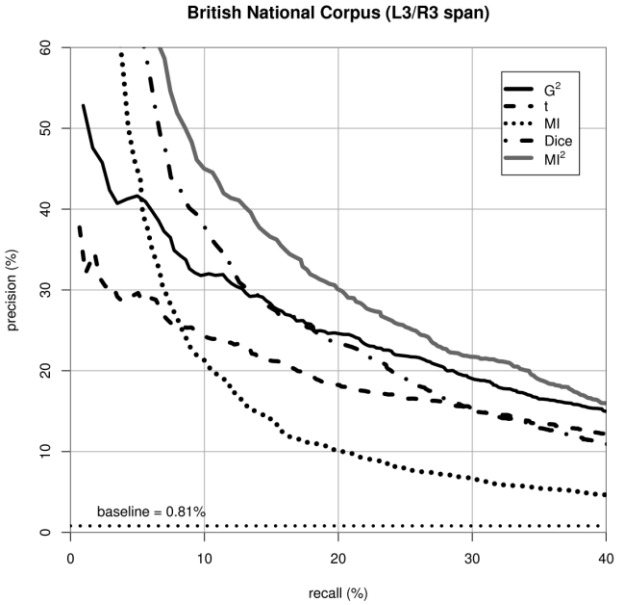


Figure 1. Sample of evaluation results for the British National Corpus with L3/R3 collocations span.

From this graph we can see that (i) Mutual Information (dotted line) indeed performs much worse than log-likelihood (solid line) and the Dice coefficient (dashed line), but that (ii) the uniformly best association measure is a variant of MI known as MI^2 (thick grey line). Looking at the centre of the plot, one can see that MI^2 identifies 20% of the collocations listed in the gold standard at an average precision of 30%. In other words, taking a sufficient number of highest-ranking candidates to find ca. 600 gold standard collocations (i.e. 20% of 2,949), three out of ten candidates from this part of the ranking are true collocations.

5 Preliminary results

The experiments carried out so far suggest the following general conclusions:

- "bigger is worse" – habitual collocations can be identified in the British National Corpus with significantly higher precision than in any of the larger corpora;
- collocations tend to form a direct syntactic relation as proposed by Bartsch (2004) – collocational spans achieve lower precision than candidates based on syntactic dependencies (and longer spans are worse than shorter spans);
- the Mutual Information measure correlates

best with the Firthian notion of collocation according to the expert judgment of lexicographers (and confirmed by one of the authors), but only if suitable statistical adjustments are made to counter the low-frequency bias of the original MI measure.

6 Conclusion and future work

The aim of the research presented in this paper is to explore the Firthian notion of collocation on the basis of statistical measures applied to corpora of differing sizes, composition and level of annotation. As the preliminary results briefly discussed in section 5 above indicate, assumptions concerning the ideal corpus size for statistical explorations of collocations – namely that bigger corpora are better – must be seriously challenged. Furthermore, these findings suggest that collocation studies can benefit from more abstract layers of annotation in order to take into account grammatical relations, extending and refining models of collocational relations. The results most strikingly suggest that there is much to be gained from further research towards improving statistical measures of association for modelling and analysing natural language phenomena.

References

Altenberg, B. 1991. "Amplifier Collocations in Spoken English." In S. Johansson and A.-B. Stenström. (eds.) 1991. *English Computer Corpora. Selected Papers and Research Guide*. Berlin, New York: Mouton de Gruyter. 127-147.

Bartsch, S. 2004. *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Dissertation. Tübingen: Verlag Gunter Narr.

Benson, M., Benson, E., Ilson, R. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Amsterdam, New York: John Benjamins.

Evert, S. 2008. "Corpora and collocations". In: A. Lüdeling and M. Kytö. (eds.), *Corpus Linguistics. An International Handbook*, chapter 58. Berlin: Mouton de Gruyter.

Evert, S. and Krenn, B. 2001. "Methods for the qualitative evaluation of lexical association measures". In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Firth, J. R. [1951] 1957. "Modes of meaning." In: *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press. 190-215.

Pecina, P. 2005. "An extensive empirical study of collocation extraction methods". In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, MI.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

A corpus-based study of the Non-Obligatory Suppression Hypothesis (of Concepts in the Scope of Negation)

Israela Becker
Tel-Aviv University
israela2@post.tau.ac.il

1 Introduction

According to *negation as suppression* view, prevalent in psycholinguistics, negation is an instruction from a speaker to an addressee to eliminate the concept within the scope of negation from the mental representation and replace it with its antonym, should this be available. If no antonym is available, the negated concept simply decays (Hasson and Glucksberg 2006; Kaup et al. 2006; MacDonald and Just 1989; Mayo et al. 2004). An alternative approach – the *non-obligatory suppression* of a concept within the scope of negation – was introduced by Giora (2006) via the *negation as mitigation* hypothesis. Giora assumed that the negation operator does not necessarily suppress the concept within its scope, but often *retains* it for pragmatic considerations (namely, mitigation). Giora and her colleagues (Giora et al. 2005a, 2005b; Giora et al. 2007) provided extensive psycholinguistic evidence supporting the *retention* of the concept in the scope of negation.

The current research aims to provide a corpus-based test for the *non-obligatory suppression* hypothesis, by using the notion of 'polarity strength'.

2 'Polarity strength'

'Polarity strength' is a numerical value disclosing the degree of positivity or negativity associated with a concept, i.e. the strength of sentiment expressed when using this concept. 'Polarity strength' dates back to the late 1950s when Osgood, Suci, and Tannenbaum (1957) developed a technique for measuring the connotative meaning of concepts, known as the 'semantic differential', where participants rated a concept along several scales, one of which is the 'potency scale' (which lies between 'strong' and 'weak') determining the concept's 'polarity strength'. Attempts to calculate (rather than rate) the 'polarity strength' of concepts were also made in the field of computational linguistics (some prominent works are Esuli and Sebastiani 2006; Kamps et al. 2004; Williams and Anand 2009).

Here I propose a different potency measurement that would serve the purpose of my research, which is to test the *non-obligatory suppression* hypothesis:

I (deliberately) **naïvely** assume that an adjective and its negated antonym (e.g., 'bad' and 'not good') are interchangeable. When a speaker intends to use a certain concept, she can either use the direct concept (e.g., 'bad') or its indirect logical equivalent (e.g., 'not good'). This is a reasonable assumption when canonical adjectival pairs are in focus. Statistically speaking – an adjective and its negated antonym are the two possible outcomes that constitute a sample space, from which the speaker could select either one or the other. Each of these two outcomes could be assigned a probability. The sum of the two probabilities amounts to 1. In other words, if a direct outcome (e.g., 'bad') is avoided, it is replaced by no other than its negated indirect version (e.g., 'not good').

Why would then speakers refrain from using a direct adjective? Its strength is likely to be a reason. That is, the stronger an adjective, the more infrequent it is with respect to its negated-mitigated version (its negated antonym). Thus, a *Strength Index* (*SI)* can be formulated:

$$SI_{Adjective} = \frac{(Negated\_Antonym)}{(Negated\_Antonym + Adjective)}$$

where the term "Adjective" refers to the number of times an adjective appears in a corpus ('bad', for instance), and the term "Negated_Antonym" refers to the number of times the negated antonym (e.g., 'not good') appears in the same corpus. The ratio between the numerator and the denominator expresses the extent to which a negated adjective is preferred over its direct antonym. The higher the *SI*, the less preferred (and consequently, more replaced) the direct expression is with respect to its negated antonym. In other words, the higher the *SI*, the stronger the adjective.

So how could, then, the numerical values of the *SI* account for the retention of the concept in the scope of negation?

3 Hypothesis and predictions

Since the numerical value of the *SI* is meant to disclose the potency, namely the 'strength' of an adjective, participants' evaluation as regards the strength of adjectives can be checked for correlation with *SI* values. We can then sketch potential profiles of the aforementioned correlation (between the behavioral data and the numerical values of the *SI*).

Whether a concept in the scope of negation is retained or suppressed will be determined in light of **prior discourse expectations** (whether positive or negative). The decision whether the correlation (either strong or weak) indicates suppression or

	Strong Correlation between participants' evaluation and *SI* values of:	Weak Correlation between participants' evaluation and *SI* values of:	Retention or Suppression? (when prior positive expectations are assumed)
Pattern 1	Negative adjectives		Retention
		Positive adjectives	Retention
Pattern 2	Negative adjectives		Retention
	Positive adjectives		Suppression
Pattern 3		Negative adjectives	Suppression
		Positive adjectives	Retention
Pattern 4		Negative adjectives	Suppression
	Positive adjectives		Suppression

Table 1. The potential correlation patterns between participants' evaluation of the strength of positive and negative adjectives and *SI* values (given prior positive expectations).

retention of the concept in the scope of negation has to conform with prior expectations. In the following, we will explain how prior positive expectations determine the retention or suppression of the concept in the scope of negation. Analysis in light of prior negative expectations is straightforward.

When prior positive expectations are assumed, 4 correlation patterns are possible, as summarized in Table 1. Due to the limited scope of this abstract, I will discuss only one of the 4 potential patterns, termed "pattern 1". If findings of this study can be accounted for by this particular pattern rather than the others, they will provide support for the *non-obligatory suppression* hypothesis (Giora et al., 2005a, 2005b; Giora et al. 2007).

Following Colston (1999), positive expectations make 'good' the natural candidate speakers may decide to use. 'Bad' is not part of positive expectations nor is 'not bad'. If the concept in the scope of negation *is* retained, then 'not bad' will *not* replace 'good' because it would be in conflict with prior positive expectations. Therefore, 'good' and 'not bad' would *not* constitute a single sample space. Consequently, the *SI* of positive adjectives (given positive expectations) will *fail* to express the strength of positive adjectives. As a result, speakers' evaluation of the strength of positive adjectives will fail to correlate with *SI* values. Along the same lines, if the concept in the scope of negation *is* indeed retained, then 'bad' (which is *not* part of positive expectations) is expected to be replaced by 'not good', as 'good' is the natural choice for the speaker when positive expectations are assumed. Therefore, 'bad' and 'not good' *do* constitute a single sample space, and as such are interchangeable. Consequently, the *SI* of negative adjectives (given positive expectations) will correctly capture the strength of negative adjectives. As a result, speakers' evaluation of the strength of negative adjectives will strongly correlate with the *SI* values. The other possible correlation patterns (which are listed in table 1) are analyzed along the same lines.

As already suggested, the only correlation pattern (out of the four suggested patterns) that fully supports the retention hypothesis is "pattern 1" which shows strong correlation between behavioral data and *SI* values of negative adjectives, and weak correlation between behavioral data and *SI* values of positive adjectives.

4 Procedure, findings, and conclusions

Eight canonical bi-polar (morphologically-unrelated) adjectival pairs (16 adjectives) of an emotive nature – which were established as emotive by Stone et al. (1966)[1], and as canonical by several research groups (Deese, 1964; Gross et al. 1989; Jones et al. 2007; Paradis, 2010; Van de Weijer et al. 2012) – were extracted out of the Blitzer, Drezde and Pereira (2007) Sentiment Dataset[2]. This dataset is devoted to customer reviews extracted from the www.amazon.com website[3], and consists of 1.1M tokens. I focused on the salient meaning of each adjective as appearing in WordNet (Miller et al. 1990), while manually filtering non-salient meanings as well as idioms. The number of counts for each adjective ranged from fifty to several hundreds. *SI* values were calculated for each adjective.

Next, *SI* values were compared with behavioral data, namely participants' ratings of the 'polarity

[1] http://www.wjh.harvard.edu/~inquirer/
[2] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index.html
[3] http://www.amazon.com

strength' of the 16 adjectives (presented in isolation) on a –5 to 5 evaluative scale. Figure 1 presents the participants' ratings vs. the *SI* results calculated for the 16 adjectives. The correlation between participants' rating of negative adjectives and *SI* values is statistically significant, $r(758)=-0.43$, $p<0.001$, while the correlation between participants' rating of positive adjectives and *SI* values is statistically insignificant, $r(758)=0.004$, $p<0.9$.

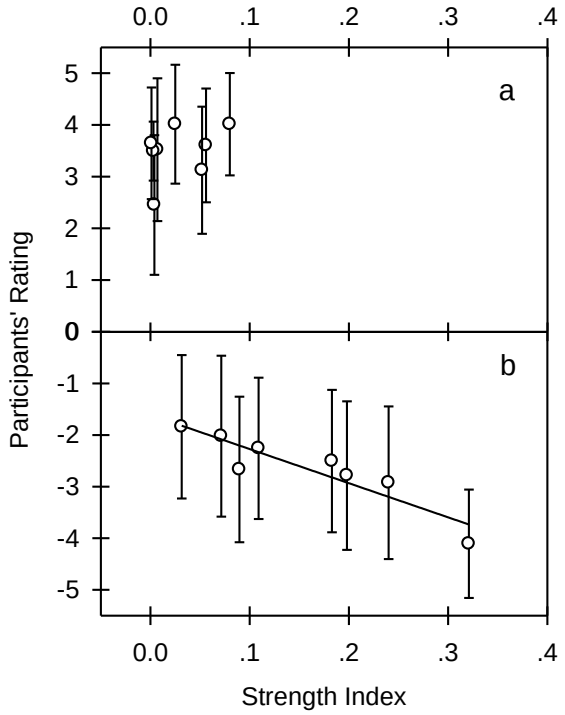


Figure 1. Participants' ratings of the strength of the 16 adjectives vs. the adjectives' calculated values of the *Strength index* of (a) 8 positive adjectives and of (b) 8 negative adjectives.

The strong correlation of behavioral data with *SI* values of negative adjectives and the weak correlation of behavioral data with *SI* values of positive adjectives is predicted by the correlation pattern termed here "pattern 1". As suggested earlier, "pattern 1" reflects *retention* of the concept in the scope of negation (given positive prior expectations). Results, thus, lend support to the *non-obligatory suppression* approach.

References

Blitzer, J., Dredze, M., & Pereira, F. (2007). *Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification.* Paper presented at the the 45th Annual Meeting of the Association of Computational Linguistics.

Colston, H. L. (1999). "Not Good" is "Bad," but "Not Bad" is not "Good": An analysis of three accounts of negation asymmetry. *Discourse Processes, 28*(3), 237-256.

Deese, J. (1964). The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior, 3*(5), 347-357

Esuli, A., & Sebastiani, F. (2006). *SENTIWORDNET: A publicly available lexical resource for opinion mining.* Paper presented at the the 5th Conference on Language Resources and Evaluation LREC'06.

Giora, R. (2006). Anything negatives can do affirmatives can do Just as well, except for some metaphors. *Journal of Pragmatics, 38*, 981-1014.

Giora, R., Balaban, N., Fein, O., & Alkabets-Zlozover, I. (2005a). Negation as positivity in disguise. In H. L. Colston & A. N. Katz (Eds.), *Figurative language comprehension: Social and cultural influences* (pp. 233-258). Hillsdale, NJ: Erlbaum.

Giora, R., Fein, O., Aschkenazi, K., & Alkabets-Zlozover, I. (2007). Negation in context: A functional approach to suppression. *Discourse Processes, 43*(2), 153-172.

Giora, R., Fein, O., Ganzi, J., Alkeslassy-Levi, N., & Sabah, H. (2005b). On negation as mitigation: The case of negative irony. *Discourse Processes, 39*(1), 81-100.

Gross, D., Fischer, U., & Miller, G. A. (1989). The organization of adjectival meanings. *Journal of Memory and Language, 28*, 92-106.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? an examination of negated metaphors. *Journal of Pragmatics, 38*, 1015-1032.

Jones, S., Paradis, C., Murphy, L. M., & Willners, C. (2007). Googling for 'opposites': A web-based study of antonym canonicity. *Corpora, 2*(2), 129-154.

Kamps, J., Marx, M., Mokken, R. J., & de Rijke, M. (2004). *Using WordNet to measure semantic orientations of adjectives*. Paper presented at the the Fourth International Conference on Language Resources and Evaluation.

Kaup, B., Luedtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics, 38*, 1033-1050.

MacDonald, M. C., & Just, M. A. (1989). Changes in activation level with negation. *Journal of Experimental Psychology, 15*(4), 633-642.

Mayo, R., Schul, Y., & Burnstein, E. (2004). ''I am not guilty'' vs ''I am innocent'': Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40*, 433-449.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to Word-Net: An on-line lexical database. *International Journal of Lexicography, 3*, 235-244.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Paradis, C. (2010). *Good, better and superb antonyms: a conceptual construal approach*. Prague, Czech Republic: Charles University in Prague, Faculty of Philosophy and Arts.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis* (Vol. 8). Cambridge, MA: The MIT Press.

Van de Weijer, J., Paradis, C., Willners, C., & Lindgren, M. (2012). As lexical as it gets: The role of co-occurrence of antonyms in a visual lexical decision experiment. In D. Divjak & S. T. Gries (Eds.), *Frequency effects in cognitive linguistics* (Vol. 2, pp. 255 – 279). Berlin, Germany: Mouton de Gruyter.

Williams, G. K., & Anand, S. S. (2009). *Predicting the polarity strength of adjectives using WordNet*. Paper presented at the Proceedings of the Third International ICWSM Conference.

Integrating visual analysis into corpus linguistic research

Monika Bednarek
University of Sydney
Monika.Bednarek@sydney.edu.au

Integrating the analysis of meaning-making systems other than language is becoming more important for the discipline of Linguistics, including for corpus linguistic research: Nowadays, many texts that corpus linguists are interested in exploring are instances of 'multimodal' (Kress and van Leeuwen 2006) discourses – that is, discourses where meanings made in modes (systems) other than language (for example, images) are of crucial importance. Recent research has seen the development of multimodal corpora and resources, such as the Nottingham Multi-Modal Corpus (e.g. Adolphs and Carter in press). New software programmes such as O'Donnell's UAM Image Tool[1] allow the annotation and quantitative analysis of image corpora. No doubt we will see similar developments of such useful resources in the future, which will help us to integrate visual and other multimodal analysis into corpus linguistic research.

However, until such resources become more wide-spread, and when researchers do not have access to multi-modal corpora per se, there are other ways of integrating visual analysis into corpus-linguistic research. Below I focus on illustrating some possible ways of doing so using the example of media discourse. More specifically, I discuss studies of two types of media discourse: fictional television discourse (which includes moving images) and newspaper discourse (which includes still images). The common context of these data is that they are both examples of mass media discourse, and that images, whether still or moving, make important contributions to the meanings made in such discourse. The common approach to these data is that of corpus-assisted discourse studies (Partington 2009) or corpus-based discourse analysis (Baker 2006).

The first study concerns an examination of emotionality in televisual dialogue (Bednarek 2010), where computerised analysis of a one-million word corpus of television dialogue (analysing emotive interjections such as *oh my god, for heaven's sake*) with the help of Wordsmith software) was combined with in-depth multimodal analysis of one scene. This analysis of one scene explored emotionality expressed through language, gesture, facial expressions, and head movements. The in-depth

[1] http://www.wagsoft.com/ImageTool/

multimodal analysis showed how these resources work together to construe two different identities for the characters interacting in the scene. At the same time, the computerised analysis of the whole corpus demonstrated that the temporary (multimodal) construal of emotionality in one scene instantiates one of the character's more general (stable) identities throughout the TV series. Combining quantitative and qualitative perspectives in this way enables us to know whether a speaker's expressive behaviour is tied to a particular situation and context or whether it is part of the speaker's more stable identity. It also allows us to explore contradictory meanings (where the emotion established through visuals differs from the emotion established through language).

The second study concerns an exploration of 1000 so-called 'stand-alone' news stories. A stand-alone is a type of news story where the image dominates – both the verbal text and often the page and has therefore also been called an 'image-nuclear' news story (Caple 2009). This is based on joint research (Bednarek and Caple 2012) where quantitative database analysis of images and quantitative corpus linguistic analysis of words (of all 1000 stories) – with a focus on composition and evaluation – was combined with qualitative, in-depth analysis of selected news stories to provide insights into verbal and visual meaning-making in this new news story genre. The quantitative analyses allowed insights into how meanings were constructed through both language and images *throughout* the corpus, whereas the in-depth analysis of two soft news and two hard news stories allowed an exploration of how meanings in image and text relate to each other (for example, through verbal-visual play; compare also Caple and Bednarek 2010) and to what extent these four stories exhibit the features that the quantitative analyses uncovered.

Both studies demonstrate different ways of integrating visual analysis into corpus-linguistic research and pose specific challenges as well as offering unique rewards to researchers.

References

Adolphs, S. and Carter, R. in press. *Spoken corpus linguistics: from monomodal to multimodal* (*Routledge advances in corpus linguistics*). London/New York: Routledge.

Baker, P. 2006. *Using corpora in discourse analysis*. London/New York: Continuum [Bloomsbury].

Bednarek, M. 2010. *The language of fictional television. drama and identity*. London/New York: Continuum.

Bednarek, M. and Caple, H. 2012. *News discourse*. London/New York: Continuum [Bloomsbury].

Caple, H. 2009. "Multisemiotic communication in an Australian broadsheet: a new news story genre". In C. Bazerman, A. Bonini and D. Figueiredo (Eds.). *Genre in a changing world, perspectives on writing*. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press. Available for download from: http://wac.colostate.edu/books/genre/

Caple, H. and Bednarek, M. 2010. "Double-take: unpacking the play in the multi-modal news story". *Visual Communication* 9 (2): 211-229

Kress, G. and van Leeuwen, T. 2006. *Reading images. The grammar of visual design*. Second Edition. London/New York: Routledge.

Partington, A. 2009. "Evaluating evaluation and some concluding thoughts on CADS". In J. Morley and P. Bayley (Eds.). *Corpus-assisted discourse studies on the Iraq conflict*. London/New York: Routledge.

Individual and gender variation in spoken English: Exploring *BNC 64*

Vaclav Brezina
VU University Amsterdam
v.brezina@vu.nl

1 Introduction

The study of language variation pioneered by Labov in the 1960s has focused mainly on phonetic/phonological variables. As Labov (2006 [1966]) pointed out, phonetic/phonological variables lend themselves fairly easily to variationist research due to their frequent occurrence and semantic stability of different variants (cf. Gordon, 2007; Tagliamonte, 2006, p. 70ff). Since the emergence of large language corpora searchable by a computer, a number of linguists have been intrigued by the possibility of exploring variation also at the syntactic and discourse levels (cf. Bauer, 2008; Pichler, 2010; Smith, 2007; Watt, 2007). However, this type of corpus-based sociolinguistic research has to overcome a number of methodological problems (cf. Baker, 2010).

So far, corpus-based sociolinguistic studies have typically offered general comparison of frequencies of a target linguistic variable in socially defined sub-corpora (e.g. speech of all men vs. speech of all women in the corpus). This procedure, however, emphasises inter-group differences and ignores within-group variation as most of these studies do not use any measure of dispersion (cf. Gries, 2006). This study therefore seeks to fill the gap in corpus-based sociolinguistic research by exploring both individual and social (gender) variation in spoken language data. It is based on *BNC 64*, a socially balanced corpus extracted for the purposes of this study, in which the speech of individual speakers can be easily traced. The study tests two hypotheses about spoken language:

1) *Gender-based sociolect hypothesis:* Male and female speakers differ in the use of a number of lexico-grammatical features in their speech.

2) *Individual style hypothesis:* Different individual speakers are consistent in employing different lexico-grammatical features in their speech.

2 Method

The study is based on *BNC 64*, a 1.5-million-word corpus of casual speech extracted from the *BNC – demographic*. *BNC 64* is a corpus which represents the speech of 64 selected speakers (32 men and 32 women) who provide between 6.4 and 64 thousand tokens each. In addition to gender, the corpus is also balanced for age, socio-economic status and region (see Table 1). In *BNC 64*, the transcribed speech from each individual speaker is stored in a separate file which enables us to easily explore both individual and social variation in the corpus.

Gender	Age	Socio-econ. status		Region
32 M 32 F	A (14-34): 24 B (35-54): 27 C (55+): 13	AB: 14 C1: 16	MC: 30	different regions across the UK
		C2: 17 DE: 13	WC: 30	
		UU: 4		

Table 1: *BNC 64* – Basic characteristics

Based on the review of sociolinguistic literature (cf. Holmes, 1995; Talbot, 2003) and Biber's (Biber, 1991 [1988]; Biber & Conrad, 2009) analysis of register variation, ten lexico-grammatical features were selected and searched for in the corpus: *the*, *lovely*, nominalizations, other nouns, word length, contractions, predicative adjectives, private verbs, personal pronouns and hedges. These variables represent a wide range of linguistic means that are indicative of differences in the style of communication. In this study, these variables were employed to test the two hypotheses: Gender-based sociolect hypothesis and Individual style hypothesis.

In order to test the Gender-based sociolect hypothesis, the ten linguistic variables were searched for in the individual files of *BNC 64*. Statistical significance of the gender-based distribution of the individual linguistic features was tested by Mann-Whitney U test and confidence intervals were calculated. Finally, the distribution of all ten linguistic variables in the speech of the 64 speakers was analysed using correspondence analysis. This procedure (similar to factor and principle component analysis) allows us to simultaneously analyse multiple dependent variables and reduce them to two factors which explain the largest percentage of variation in the data. The product of the analysis is a two-dimensional correspondence plot in which both the dependent variables (lexico-grammatical features) and the independent variable (speaker's gender) are displayed. The correspondence plot thus allows for a good visual inspection of complex data.

In order to test the Individual style hypothesis, each of the 64 files (representing one speaker) was divided into half and the sub-files were analysed in a similar way as when testing the Gender-based sociolect hypothesis. The investigation, however, focused on finding out whether a) individual speakers show consistent linguistic use of the selected variables and b) whether individual

speakers can be distinguished from each other on the basis of the selected linguistic features.

3 Results

The results show clear gender-based and individual-based patterns in the data. The findings related to the Gender-based sociolect hypothesis are presented in Table 2 and Figure 1. As can be seen from Table 2, five out of the ten linguistic variables showed statistically significant differences between the male and the female speakers (with nouns showing a borderline significance). Men preferred the definite article, nominalizations, and predicative adjectives while women preferred the adjective *lovely* and personal pronouns.

almost 70 per cent of the variation. The vast majority of male speakers are placed left of the Y-axis with a clear preference for nominalizations, predicative adjectives and the definite article. On the other hand, female speakers cluster around personal pronouns and the adjective *lovely* in the upper quarter of the plot.

Selected results (use of the definite article and personal pronouns) related to the Individual style hypothesis are displayed in Figure 2 below; here each speaker is represented by two samples. As can be seen, the samples from the same speaker appear close to each other in the plot which indicates consistent linguistic behaviour of individual speakers and supports the Individual-style

	the	*lovely*	NOUNS	NOMINALIZ	W_LENGTH
Mann-Whitney U	289	312	368	317	380.5
p	.003**	.007**	.053	.009**	.077
	CONTR	PRED_AJ	PRIV_VB	PRON	HEDGES
Mann-Whitney U	372	296	401	290	432
p	.060	.004**	.136	.003**	.283

Table 2: Differences between the male and female use of the ten linguistic variables

The complexity of the linguistic behaviour of men and women in *BNC 64* is displayed in the correspondence plot in Figure 1 below. Here the two factors extracted from the ten variables explain

hypothesis. This tendency was also confirmed by strong correlations between samples from the same speaker. In addition, Figure 2 shows a clear gender-based pattern similar to that displayed in Figure 1.

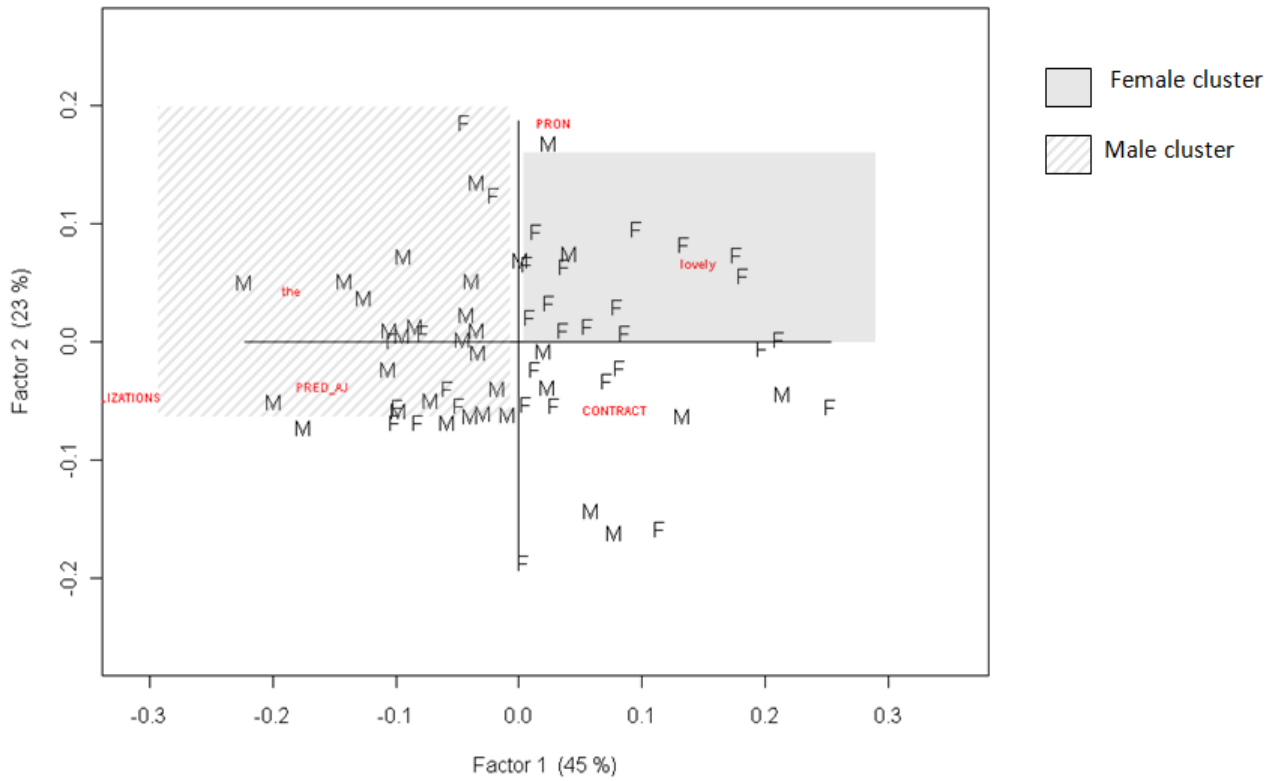


Figure 1: Gender-based pattern in *BNC 64:* Correspondence analysis

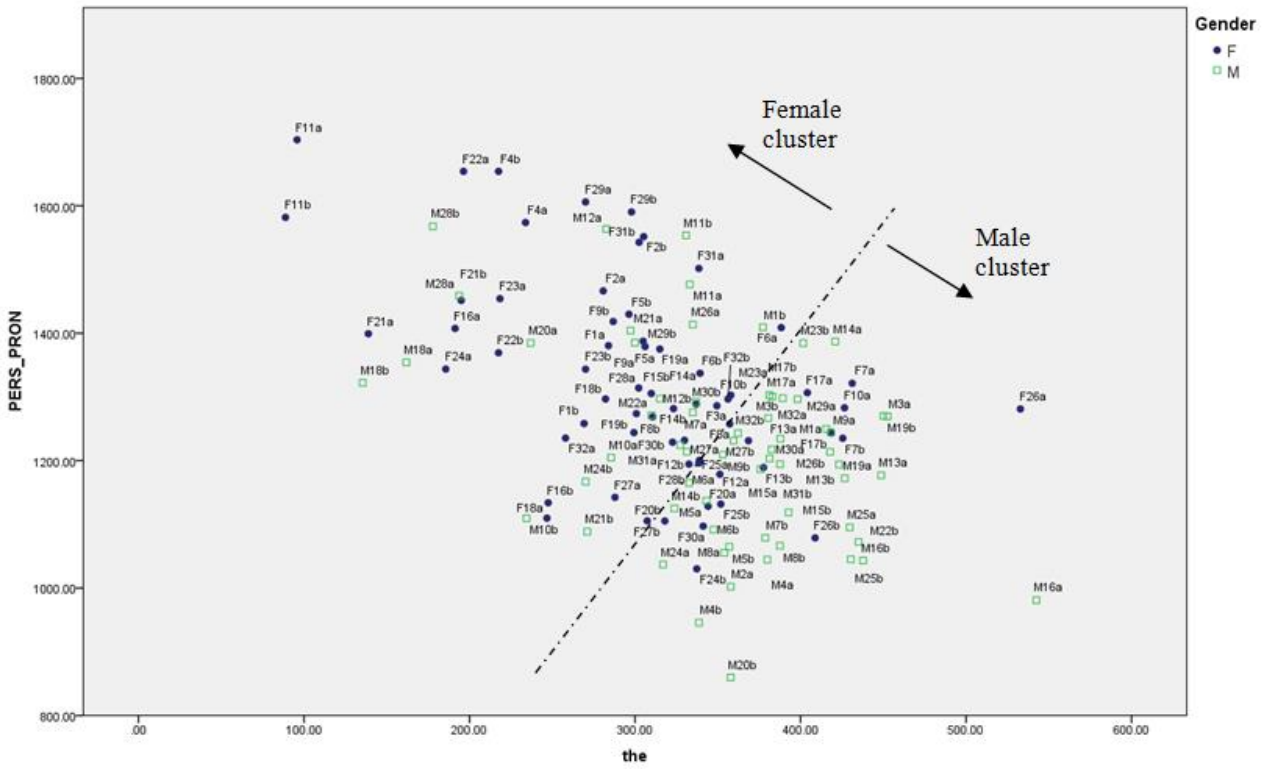


Figure 2: Individual-style pattern in *BNC 64*: The use of the definite article and personal pronouns

4 Conclusions

The data show support for both of the tested hypotheses. It seems that the difference between male and female speech in *BNC 64* can be conceptualised along a general dimension similar to Biber's involved, narrative & personalised vs. informational, descriptive & detached. The analyses showed that male speakers consistently prefer a less involved and more informational style than female speakers. In addition, the analysis revealed that individual speakers themselves are fairly consistent in the use of the key lexico-grammatical features as shown by different samples from the same speaker clustering close to each other. The paper further discusses the findings in relation to Eckert's (2009) framework of three waves of variation study and points out some of the methodological principles required for the validity of corpus-based sociolinguistic studies.

References

Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Bauer, L. (2008). Inferring variation and change from public corpora. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The Handbook of language variation and change* (pp. 97-114). Oxford: Blackwell.

Biber, D. (1991 [1988]). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Eckert, P. (2009). *Three waves of variation study: The emergence of meaning in the study of variation*. Retrieved 1/12/2012, from http://www.stanford.edu/~eckert/PDF/ThreeWavesofVariation.pdf

Gordon, M. (2007). Techniques of analysis: phonological variation. In C. Llamas, L. Mullany & P. Stockwell (Eds.), *The Routledge companion to sociolinguistics* (pp. 19-27). London: Routledge.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora, 1*(2), 109-151.

Holmes, J. (1995). *Men, women and politeness*. London: Longman.

Labov, W. (2006 [1966]). *The social stratification of English in New York City*. Cambridge: Cambridge University Press.

Pichler, H. (2010). Methods in discourse variation analysis: Reflections on the way forward. *Journal of Sociolinguistics, 14*(5), 581-608.

Smith, J. (2007). Techniques of analysis: morphosyntactic variation. In C. Llamas, L. Mullany & P. Stockwell (Eds.), *The Routledge companion to sociolinguistics* (pp. 28-40). London: Routledge.

Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Talbot, M. (2003). Gender stereotypes: Reproduction and challenge. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of language and gender* (pp. 414-424). Oxford: Blackwell.

Watt, D. (2007). Variation and the variable. In C. Llamas, L. Mullany & P. Stockwell (Eds.), *The Routledge companion to sociolinguistics* (pp. 3-11). London: Routledge.

Automatically identifying instances of change in diachronic corpus data

Andreas Buerki
Humboldt-Universität zu Berlin
andreas.buerki@hu-berlin.de

1 Introduction

With the increasing availability of diachronic corpora, machine-aided identification of linguistic items that have undergone significant change is set to become an important task. This importance is heightened further if, as Hilpert and Gries (2009:386) have argued, approaching linguistic change in a data-driven manner can reveal otherwise unnoticed phenomena. Key to this endeavour is being able to tell apart relevant change from noise and random or other synchronic variation. This non-trivial task differs in important ways from the much more widely investigated comparison of linguistic features between two (usually contemporary) corpora and has to date not received the attention it should perhaps be afforded.

In this paper, a number of methods for identifying relevant change are reviewed and a procedure suggested which has not so far been documented. This new procedure is based on a simple chi-square test for goodness of fit, combined with additional parameters. Its operation is illustrated using the example of a study conducted to investigate motivation of recent and ongoing change in Multi-word Expressions (MWEs) using data taken from the 20-million word Swiss Text Corpus (STC). The STC is a corpus of 20th century written German as used in Switzerland (Bickel et al 2009). Results of the application of the proposed method indicate that the procedure yields high-quality instances of significant change in the data and is applicable to MWEs as well as a range of other linguistic items. It is able to identify instances of change with fewer arbitrary decisions and able to identify a wider range of different types of change than other suggested methods. Additionally, it shows that both the structure of the data as well as particular research interests will guide the choice of method used to identify relevant change.

2 Goals, data and possible procedures

The method presented is designed to deal with data that consist of a large number of linguistic items and their frequencies in different time periods. An example for this type of data is seen in table 1 where one of nearly 18,000 MWEs extracted from the STC is shown with associated frequencies over five time periods.

Item	Frequencies in 5 periods
Im Laufe der Jahre *'in the course of time'*	23 26 76 35 31

Table1: Data structure

The method presented can be applied to any linguistic item as long as types and tokens can be extracted from corpus materials and counted (e.g., words, constructions, morphological features, etc.).

Before discussing ways to identify instances of change, it is pertinent to consider two preliminary questions. The first concerns what should be considered relevant change. For present purposes, a relevant instance of change is defined as an observable instance of change in the data which is not due to noise or accidental variation, but rather reflects what could reasonably be thought to indicate diachronic change in the language of which the corpus is a sample.

The second question concerns the types of change that should be identified. These are taken to be

1. the appearance of (new) types
2. the disappearance of (old) types
3. semantic shifts (stable form)
4. change in form (stable semantics)
5. notable in- and/or decreases in frequency.

Following other diachronic corpus studies, these five types of change are investigated from the vantage point of changes in frequency which, as will be demonstrated, can be used not only to identify change of types 1, 2 and 5, but also of types 3 and 4.

Only a small number of methods for identifying relevant shifts in frequencies have been suggested to date. These include the use of a coefficient of variance (CV) as applied in Baker (2011) and a rank-order correlation measure where frequencies are correlated with the sequence of time periods in the corpus (i.e. periods 1 to n) suggested in Hilpert and Gries (2009). A further possibility is to use Belica's (1996) coefficient of difference (D), the values of which can be squared for each period and then summed to arrive at an overall measure of change. A fourth option which has not so far been applied is to use a chi-square test for goodness of fit to test if the differences in frequency across the periods is significantly different from what could be expected due to chance.

To find the most useful method for purposes of identifying relevant change in the data outlined, all four methods were applied to MWEs extracted from the STC. In all cases, the following additional parameters were set: only MWEs were considered which showed a frequency of at least four occurrences per a million words in at least two of the five time periods (i.e. they could occur less often or

not at all in maximally three of five time periods). This ensured that items identified occured with notable frequency at one point, but also allowed for patterns where an item might have appeared or disappeared (or both) during the period of investigation. Further, frequencies were capped at three times the number of documents in which they occurred. This was used to prevent burstiness (caused, for example, by a topical concentration of certain items in individual documents) from unduly influencing the tracing of diachronic developments.

For the rank-order correlation method (we used Spearman's rho) and the method using the chi-square test, levels of significance were defined which provided a non-arbitrary divide between significant and non-significant change. For Spearman's rho, which was more restrictive, a significance level of alpha = 0.05 was used, for the chi-square test, a more stringent significance level of alpha = 0.001 was specified. For the other two methods, the highest scoring third of changes was considered to have undergone relevant change (cf. Baker 2011).

3 Results of the evaluation

Instances of relevant change identified by each of the four methods were compared by looking at individual test cases as well as the overall number of items identified. Table 2 shows the number of MWE-types identified as having undergone relevant change in each of the four methods. The total number of MWE-types occurring with minimum frequency in at least two time periods was 17,645.

Method	Types with relevant change
Top third (CV and D)	5,881
Spearman's rho	1,268
Chi-square	7,563

Table 2: Number of types with relevant change

The figures of table 2 show that the correlation method identifies the fewest types as having undergone relevant change. In fact, significance is only reached for perfect correlations (i.e. either a progressive in- or decrease in frequency over the five time periods). This is because the five data points provided by the temporal structuring of the source data provide insufficient statistical power; more data points would make this measure more meaningful, but such are not available in many cases.[1] The correlation based method, requiring a perfect rank-correlation is therefore too narrow to be useful for the data structures described. Comparing the detailed results of the remaining methods shows that in a number of instances the chi-squared based method appears to make more sensible decisions than the other two. Unlike the other two remaining methods, it also provides a non-arbitrary cut-off

point for identifying significant change. A cut-off point, furthermore, which is easily interpretable: items whose frequencies differ in ways that are unlikely to be due to chance are identified as having undergone relevant change. This links in well with the definition of relevant change given above.

4 Conclusions

In the study of MWE-change in which the methods were evaluated, the chi-square-based approach served as the most useful method for identifying relevant change among a vast number of potential changes. It showed important advantages over other possible procedures. The chi-square-based method was shown to be 1) broad enough to include a variety of patterns, rather than only a progressive in- or decrease in frequency, 2) to fit well with the definition of relevant change used, 3) to provide a statistically robust, non-arbitrary cut-off point and 4) it was well suited for application to data that cannot provide a large number of data points. The additional parameters used (an item needing to appear in at least two periods with a minimum frequency of 4/M and the cap on frequencies) added to the robustness of results and therefore their usefulness which was confirmed when a sample of MWEs identified as having undergone significant change was investigated in detail to establish motivations for change. While a comprehensive and broad identification of all significant change in the data was advantageous in the application reported on, for other purposes, a more selective method might be desired. In such cases, a measure such as the coefficient of variance used in Baker (2011) could additionally be used to rank change and limit investigation to top-ranking cases.

References

Baker, P. (2011). Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics*, *39*(1), 65-88.

Belica, C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, *1*(1), 61-73.

Bickel, H., Gasser, M., Häcki Buhofer, A., Hofer, L., & Schön, C. (2009). Schweizer Text Korpus. *Linguistik Online*, *39*(3), 5-31.

Hilpert, M., & Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora. *Literary and Linguistic Computing*, *24*(4), 385-401.

[1] Neither is it possible to apply the more sophisticated methods suggested by Hilpert and Gries (2009), for the same reason.

Reader engagement in Turkish EFL students' argumentative essays[1]

Duygu Çandarlı
Boğaziçi University
duygu.candarli@boun.edu.tr

Yasemin Bayyurt
Boğaziçi University
bayyurty@boun.edu. tr

Leyla Martı
Boğaziçi University
marti@boun.edu.tr

This study aims to explore reader representation in three different corpora of students' essays, namely BUELC (Boğaziçi University English Language Learner Corpora), LOCNESS (the Louvain Corpus of Native English Essays) and TEC (Turkish Essays Corpora). It is generally pointed out that the notion of audience and reader engagement in writing is significant, which means that writers interact with their readers and form a dialogue with them while writing. The concept of engagement, a component of Hyland's interaction model (2005a), is based on the Bakhtinian notion of dialogism which regards language as inherently dialogic (Bakhtin 1981). According to him, all language users tend to express their opinions by taking the real or imagined audience into account. In the case of student writing, the concept of audience is a bit vague as the students write their essays for their instructors to receive a grade or for their peers to get feedback. Also, they might have imaginary audience in their mind. Although it is high likely that students have a difficulty in engaging a dialogue with their instructors because of the power-relations and lack of audience awareness, it is possible for students to project their authority and construct a dialogue with their instructors and/or imaginary audience. In this way, they can become participating members of a wider discourse community and contribute to the on-going dialogue (Tang 2009). However, as argued in Hyland (2005b), it may be difficult for students "who are not used to seeing writing as interactive or to imagining the perceptions, interests and requirements of a potential audience" (p.364). Since students seem to have abstract notion of audience, their writing tends to be somewhat voiceless and interpersonal. Hyland (2005b) investigated reader engagement in Cantonese undergraduates' writing in English in a comparison with professional writing. The results revealed that student texts included markedly less engagement features than research articles, which had almost twice as many engagement features than student texts. Reader pronouns were the most common devices for published articles while directives were most frequent in the student texts. It was noted that reader pronouns highlight sharedness and solidarity between the writer and the reader whereas directives tended to manipulate readers into advocating writer's arguments. Hyland (2005b) concluded that fewer engagement features in the student essays could be attributed to institutional power, rhetorical confidence and probably cultural preferences. As student writers wrote their essays in an educational setting, they may not have been self-confident enough to engage with their tutors in their essays in Chinese culture.

Although there is a growing body of literature on contrastive rhetoric and metadiscourse strategies in Turkish students' L1 and L2 essays, to the knowledge of the researchers, very few studies have examined reader engagement in Turkish students' essays. It is believed that this study attempts to fill this gap by describing to what extent L1 and L2 essays of Turkish learners of English show the features of reader representation in comparison with the essays of monolingual American students. The argumentative essays from Boğaziçi University English Language Learner Corpora and Turkish essays of the same students are compared with a reference corpus of the LOCNESS. Reader engagement markers, including directives, questions, references to shared knowledge and personal asides are examined. AntConc 3.2.4 is used for data analysis, and pragmatic annotation is conducted with the UAM Corpus Tool. According to Hyland's model (2005a), attitude markers have stance functions; however, they can also be engagement resources in some contexts. In a recent study on stance and engagement in pure mathematics research articles, McGrath and Kuteeva (2012) argue that attitude markers are highly multifunctional, and they can be employed to engage with the readers as a rhetorical strategy apart from conveying writer's stance. Therefore, the functions of attitude markers are also discussed. Results indicate that Turkish essays written by Turkish students include considerably more directives and attitude markers than English essays written by the same group and American students' essays, which is in line with the findings of Hyland's study (2005b) that revealed extensive use of directives by Chinese students. An unexpected result of the study is that English essays of Turkish students have more reader pronouns than the other groups. The contrastive analysis of three different corpora sheds light on the interplay of bidirectional transfer, writing instruction, individual learner preferences and audience awareness.

[1] This study was supported by the Research Fund of Bogazici University, Project No: 5691)

References

Bakhtin, M. 1981. *The dialogic imagination: Four essays*. (trans. C. Emerson & M. Holquist). Austin: University of Texas Press.

Hyland, K. 2005a. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7 (2): 173-191.

Hyland, K. 2005b. Representing readers in writing: student and expert practices. *Linguistics and Education* 16 (4): 363-377

McGrath, L. and Kuteeva, M. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes* 31 (3): 161-173.

Tang, R. 2009. A dialogic account of authority in academic writing. In M. Charles, S. Hunston & D. Pecorari (eds.) *Academic writing: at the interface of corpus and discourse* (pp.170-188). London: Continuum.

It was X that type of cleft sentences and their Czech equivalents in InterCorp

Anna Čermáková
Charles University
anna.cermakova@ff.cuni.cz

František Čermák
Charles University
frantisek.cermak@ff.cuni.cz

1 Introduction

Despite a considerable and often conflicting number of views on how to view the phenomenon of clefting, how to call it or to which school it should be ascribed (clefting to generative grammar, topic-comment to Prague structuralism, etc.) the overall attention has so far been largely concerned with formal, mostly syntactic aspects, leaving other aside. To try to do more justice to the phenomenon branded here by one of the usual labels as clefting (or cleft sentences), which is typical for languages with a firm word order, such as English and Germanic languages in general, one has to switch over typologically from analytic to inflectional languages and view it functionally, i.e. as a means of expression of what is often called focus or comment, i.e. drawing the reader´s attention to a part of information that seems more important and new. In contrast to rather formal and endless discussions about syntactic properties, other issues, largely forgotten, should be stressed, too, such as emphasis laid on the information conveyed: indeed, it seems that although emphasis is very much part of this phenomenon, it is not clear how to distinguish it from a neutral cleft sentence of the type *It was this book that caught his imagination*.

On the basis of a large multilingual parallel corpus InterCorp, we will look in detail at one of the possible cleft-constructions *It was X that* and its Czech equivalents trying to bridge the divide between two typologically different languages. There is no equivalent grammatical construction in Czech and comparison of both languages shows that, although the Czech equivalents may not be so pronounced, there are more options to choose from, naturally the word order and intonation (in speech) playing an important role.

2 Background

It-clefts consist of the pronoun *it* and the verb *be*, both words belong to the most frequent words in English and so does their combination. The pattern *it is* occurs 128,471 per million words in the BNC. Over a third of the occurrences are in sentence initial position. The pattern *it was* occurs 122,318 p.m. and again more than third is in sentence initial position.

The most frequent complementation of *it is* are adjectives followed by *to*-infinitive (*possible to, important to, difficult to, necessary to*) and adjectives followed by *that*-clause (*clear that, likely that, possible that*). Patterning on the right of *it was* is less tidy (*it was the first, it was one of, it was not until, it was in the, it was held that*) and occurrences of frequent n-grams are much lower than in the case of the present tense *it is.* The most frequent is *it was the first* (followed by *time* in nearly half of the cases), which occurs 698 times, while the most frequent complementation with the present tense *it is possible* occurs nearly three times more often (1971).

The situation in Czech is rather different, however. While the Czech equivalents, the pronoun *to* and the verb *být*, belong to the most frequent words in the language as well and the individual frequencies of occurrence of the two are comparable to those in English (data from SYN2010 corpus, 100 mil. word balanced corpus of Czech), the frequency of occurrence of their combination is much less prominent. The present tense *to je* occurs only 25,463 p.m. and *je to* 53,855 p.m., the past tense *bylo to* 40,2 p.m. and *to bylo* 14,068 p.m. The patterning of the complementations of these combinations is much less structured and we find it both on the right and on the left sides. One of the explanations for these differences is the importance of the pattern *it* followed by the verb *be* in English, which can represent the above mentioned *it*-type cleft sentences and introductory-*it* sentences, which look much the same on surface.

Both of these constructions modify the word order in order to change or put an emphasis on a certain sentence element. *A Communicative Grammar of English* (Leech and Svartvik 2003) defines the introductory-*it* construction as "a means of postponing a subject clause to a later position in the sentence, either for end-weight or for end-focus" (p. 165) and further says "Occasionally introductory *it* displaces a clause in object position" (p. 166) while stressing that these are not to be confused with the *it*-type cleft sentences, which are defined as "useful for fronting an element as topic, and also for putting focus (usually for contrast) on the topic element. It does this by splitting the sentence into two halves, ‚high-lighting' the topic by making it the complement of *it + be* ..." (p. 163).

3 Comparison of English and Czech: Functional equivalents

Different languages express emphasising, high-lighting or attention shifting in various ways. Clefts are one of the means typical for English. The phenomenon of the English cleft sentences has received wide attention in grammars and literature (e.g. Patten 2012) even though some authors claim that they are in fact relatively rare (Roland et al. 2007). ‚Clefts' are being studied from contrastive language perspective as well, Gundel (2002) has compared English and Norwegian cleft constructions and claims they are more frequent in Norwegian, while Ahlemeyer and Kohlhof (1999) claim translated English-German texts reveal that only about a third of English *it*-clefts (or less, depending on the text type) are translated with the German equivalent. On a recent LREC conference (2010) Bouma et al. (2010) introduced a project of a specialized multilingual (Dutch, English, German and Swedish) parallel corpus of cleft sentences.

The above mentioned contrastive studies compare typologically the same languages (Germanic), which all have regular equivalent grammatical construction. In this study we would like to compare English *it*-type cleft sentences (with the verb form *was*) followed by *that*-clause and their translation equivalents in Czech. Czech is typologically different language with free word order and has no regular equivalent grammatical construction to *it*-type cleft sentences.

For the study we use data from the multilingual parallel corpus project *InterCorp.* We use an English-Czech subcorpus consisting of English original texts (35 novels by 26 British and American authors, the size of the corpus is 2 690 316 tokens in the English part). We focus on sentence initial constructions *It was* followed by 1 to 4 words followed by *that.* Cleft sentences are manually identified from all the results.

The analysis reveals how translators occasionally struggle with the cleft constructions. While we expected frequent usage of particles, such as *právě:*

> ***It was this night that*** *he told me the strange story of his youth with Dan Cody...*
> ***A právě tu noc*** *mi vyprávěl podivný příběh svého mládí s Danem Codym...*

we have found that translators often opt to ignore the cleft constructions altogether or translate rather mechanically, word by word which may sound exagerated in Czech:

> ***It was a refrain that*** *was often heard in moments of overwhelming emotion.*
> ***Byl to refrén, který*** *se často ozýval ve chvílích vzrušených emocí.*

Most frequently we find that the translation reflects the cleft construction simply by the word order prefering the initial sentence position.

> ***It was gin that*** *sank him into stupor every night...*
> ***Gin*** *ho každou noc uváděl do stavu strnulosti...*

We have additionally identified that translators often opt for a combination of word order and different

demonstrative pronouns (without any verbal element), which is interesting especially in cases where English uses the indefinite article (there are no articles in Czech and demonstratives are occasionally used in translations to render the definite article).

> ***It was the avoidance that*** *incriminated them.*
> ***To ta vyhýbavost*** *svědčila proti nim.*
>
> ***It was a memory that*** *he must have deliberately pushed out of his consciousness over many years.*
> ***Tu vzpomínku*** *musel po mnoho let úmyslně vypuzovat z vědomí.*

This study aims to identify and classify the most frequent translation options of the *It was X that* construction and compare the nature of 'emphasizing' between the two languages. Offering, eventually, a quantified table of the most common types of equivalents, we expect this case study to open up further questions and suggest avenues for further research.

References

Ahlemeyer, B. and Kohlhof, I. 1999. "Bridging the Cleft: An analysis of the translation of English it-clefts into German". *Languages in Contrast* 25: 1-25.

Bouma, G., Ovrelind, L. and Kunh, J. 2010. Towards a large parallel corpus of clefts. In *Proceedings of LREC 2010*. Malta.

Duffer, A. 2009. Clefting and discourse organization: Comparing german and romance. In A. Dufter and D. Jacob (eds.). *Focus and background in romance languages*. Amsterdam: John Benjamins.

Gundel, J. 2002. "Information structure and the use of cleft sentences in English and Norwegian". *Language and Computers* 16: 113-128.

InterCorp – Czech National Corpus. The Institute of the Czech National Corpus, Praha. Available at: www.korpus.cz

Johansson, M. 2001. "Clefts in contrast: a contrastive study of clefts and wh clefts in English texts and translations". *Linguistics* 39 (3), 547-582.

Leech, G. and Svartvik, J. 2003. A Communicative Grammar of English, 3rd Edition. Pearson Education.

Patten, A. 2012. *The English It-Cleft: A Constructional Account and a Diachronic Investigation*. Mouton de Gruyter.

Karlík, P., Nekula, M. and Rusínová, Z. 1995. *Příruční mluvnice češtiny.* Praha: NLN.

Roland, D., Dick, F. and Elman, J.L. 2007. "Frequency of basic English grammatical structures: A corpus analysis". *Journal of Memory and Language* 57 (3): 348-379.

SYN2010 – Czech National Corpus. The Institute of the Czech National Corpus, Praha, 2010. Available at: www.korpus.cz

The power of personal corpora: Students' discoveries using a do-it-yourself resource

Maggie Charles
University of Oxford Language Centre
maggie.charles@lang.ox.ac.uk

1 Introduction

Since the ground-breaking work of Johns (1991a, b), there have been many reports on the direct use of corpus data by students, particularly those studying academic writing in English at university level (e.g. Weber 2001; Bianchi and Pazzaglia 2007; Yoon 2008; Flowerdew 2012). Despite this ongoing research effort, there are only a few accounts of students constructing their own personal corpora. Two recent studies (Lee and Swales 2006; Gavioli 2009) report on students who built individual corpora to answer research questions they had formulated themselves. However, in both cases, most corpus consultation took place outside class and the aim was for students to produce extended pieces of individual work. The issue of how such personal corpora could be used and incorporated in regular class sessions has not been addressed. Nor is it clear how tools other than the concordancer could contribute to corpus pedagogy.

This paper reports on a study of 40 advanced-level EAP students who attended a course on academic writing in which each participant built a personal corpus from research articles (RAs) in their discipline. Class tasks examining specific discourse functions (e.g. making claims) formed the basis of the work, leading to a 'one task, multiple corpora' approach. This study presents data on the discoveries that students made using this combination of personal corpus consultation and class tasks and focuses particularly on the pedagogical applications of Word List, Collocates/Clusters and Plot.

2 Background and Data

The course lasted six weeks, with one two-hour session per week. In the first two sessions, students were introduced to the AntConc software (Anthony 2011) and each participant began to construct their own corpus of RAs. Each of the subsequent sessions focused on a specific corpus tool or technique, which the students used to investigate the occurrence and lexico-grammatical realisations of a given discourse function in their personal corpus (for further details see Charles 2007, 2011, 2012).

Over half of the participants (53%) were doctoral students, 15% Master's students, 15% post-doctoral researchers and 8% were taking other qualifications. A wide range of disciplines was represented (53% natural sciences; 23% social sciences; 25% arts/humanities).

There was considerable variation in the size of the personal corpora, the number of words ranging from 15,057 to 976,452, an average of about 150,000 words per corpus. The numbers of corpora in each band of the word count appear in Table 1. Although these personal corpora are quite small, they provide a useful and highly specialised resource for student investigations.

Word Count Band	Number of Personal Corpora
Under 50,000	6
50,000 – 100,000	12
100,000 – 150,000	8
150,000 – 200,000	6
Over 200,000	8
Total	40

Table 1. Number of personal corpora in each band of the word count

In order to examine each discourse function, the course employed a two-stage procedure: students were first introduced to the function within the more familiar environment of a text, and then moved on to hands-on corpus investigation. The tutor demonstrated the use of a specific corpus tool and students were given worksheets which provided a series of searches and focusing questions to help them to notice important aspects of their data and to interpret their results. Participants were asked to note down examples and data from their own corpus and to comment in writing on their findings. They then discussed their results with fellow-students and a whole-class feedback session concluded each class. This paper discusses data, examples and commentary taken from student worksheets.

3 Discoveries with Word List

The Word List tool was used in connection with work on the function 'Making and Countering Arguments'. Students were asked to record and comment on the frequency and sentence position of specific linking adverbials of contrast, result and addition. Biber et al. (1999) state that the most frequent position for linking adverbials in academic writing is sentence initial; however other researchers (e.g. Shaw 2009) have suggested that position may vary according to discipline. Word List in AntConc shows the frequency of capitalised and non-capitalised words separately, making it easy to determine whether a given adverbial occurs more

frequently in sentence initial position or not. The three contrast adverbials examined were *however*, *nevertheless* and *nonetheless*. Clear differences in preferred sentence position were found by students when they compared their findings to those in other disciplines. In natural sciences there was a preference for sentence initial use, but in arts/humanities non-sentence-initial position was privileged. For example, in a corpus of 71,000 words in literary studies, Renate[1] found 72%"88% of non-sentence-initial use, while Amina's corpus of 361,000 words in computer science showed the opposite tendency, with sentence-initial use of 72% – 81%.

In asking students to compare their own data to research findings, such tasks problematise established accounts and encourage students to adopt a critical and discerning attitude towards reference sources. Working with their own personal corpus both increases students' knowledge of the usage in their own field and gives them hard evidence to back up their language choices.

4 Discoveries with Collocates/Clusters

The Collocates and Clusters tools were used to investigate the function 'Making and Modifying Claims', focusing particularly on subject-verb combinations that occurred with the reporting verbs *suggest* and *show*. These verbs were chosen because they are associated with making claims and are likely to show disciplinary variation (Charles 2006). Consulting a corpus of 156,000 words in social work, Guo found three subjects that constructed claims with the verb *suggest*: *evidence* (10), *results* (9) and *findings* (4). By contrast, Karla's corpus of 143,000 words in politics revealed use of the pronouns *we* (40) and *I* (4) to construct claims with the verb *show*.

Collocates and Clusters allow students to discover the phraseology typical of the way in which their field performs the discourse function studied, while comparison with the findings of other students is important in revealing whether and how their discipline differs from others. Thus, although both students worked in social science disciplines, Guo's data clearly show how claims are made on an empirical basis in social work, while the phraseology identified by Karla reveals a more personal stance in politics.

Tasks such as these show the potential of the 'one task, multiple corpora' approach, which enables students to benefit from the corpus findings of all members of the class and draws attention to the diversity of academic writing. This approach not only enables generalisations to be made, but also allows for individualisation in the conclusions drawn.

5 Discoveries with Plot

The Plot tool provides a graphic representation of the frequency and distribution of the search term in each file of the corpus. In a corpus where each file belongs to the same genre and has a conventional generic structure, Plot allows the user to get a rough idea of the position of the search term in the generic structure. For example, in these personal corpora, each file is an RA, so a high concentration of instances towards the beginnings of files suggests that the search term is associated with the introduction section. Thus Plot can be used to help link search items to specific generic stages. Students employed this tool to investigate 'Making and Modifying Claims'. They were asked to make Plots for the modal verbs *may*, *could* and *might*, to comment on their distribution and to ascertain whether the verbs were being used to modify claims.

Consulting 141,000 words in history, Daniel found 108 instances of *may*, fairly evenly distributed across the files. However, they served to qualify the author's opinion rather than to make claims. By contrast, Yu found 354 *may* in 179,000 words of environmental science and noted two different tendencies. *May* could be evenly distributed throughout a file, in which case it often hedged claims. Alternatively, when it occurred towards the ends of files, in the conclusion, it referred to possibilities for future work.

By giving an indication of where a term occurs in the generic structure, Plot not only helps students to identify the parts of the text most likely to yield useful findings, but also enables them to interpret those findings in the light of their knowledge of the genre.

6 Conclusions

This paper has discussed the tools students can use and the types of discoveries they can make with personal corpora, following the 'one task, multiple corpora approach'. It highlights several factors likely to contribute to the success of corpus pedagogy: first, the worksheets provide guidance, which supports the students' investigations; second, the worksheets require students to write down examples, and third, to comment on them. Thus students are encouraged both to focus on the lexico-grammar of specific examples, and also to generalise from their data. Finally, discussion with others makes students more aware of the wide range of disciplinary practices and stimulates them to articulate and explain their own disciplinary knowledge.

[1] Student names are pseudonyms.

References

Anthony, L. (2011). AntConc (Version 3.2.4): http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Bianchi, F. and Pazzaglia, R. 2007. "Student writing of research articles in a foreign language: metacognition and corpora". In R. Facchinetti (ed.) *Corpus Linguistics 25 Years On.* Amsterdam: Rodopi, 261-287.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Charles, M. 2006. "The construction of stance in reporting clauses: a cross-disciplinary study of theses". *Applied Linguistics* 27 (3): 492-518.

Charles, M. 2007. "Reconciling top-down and bottom-up approaches to graduate writing: using a corpus to teach rhetorical functions". *Journal of English for Academic Purposes* 6 (4): 289-302.

Charles, M. 2011. "Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing classes". In A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds.) *New Trends in Corpora and Language Learning*. London: Continuum, 26-43.

Charles, M. 2012. "'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building". *English for Specific Purposes* 31: 93-102.

Flowerdew, L. 2012. "Exploiting a corpus of business letters from a phraseological, functional perspective". *ReCALL* 24 (2): 152-168.

Gavioli, L. 2009. "Corpus analysis and the achievement of learner autonomy in interaction". In L. Lombardo (ed.) *Using Corpora to Learn about Language and Discourse*. Bern: Peter Lang, 39-71.

Johns, T. 1991a. "From printout to handout: grammar and vocabulary teaching in the context of data-driven learning". In T. Johns and P. King (eds.) *Classroom Concordancing.* Birmingham: ELR University of Birmingham, 27-37.

Johns, T. 1991b. "Should you be persuaded: two samples of data-driven learning materials". In T. Johns and P. King (eds.) *Classroom Concordancing*. Birmingham: ELR University of Birmingham, 1-16.

Lee, D. and Swales, J. 2006. "A corpus-based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora". *English for Specific Purposes* 25 (1): 56-75.

Shaw, P. 2009. "Linking adverbials in student and professional writing in literary studies: what makes writing mature". In M. Charles, D. Pecorari and S. Hunston (eds.) *Academic writing: At the Interface of Corpus and Discourse.* London: Continuum, 215-235.

Weber, J.-J. 2001. "A concordance- and genre-informed approach to ESP essay writing". *ELT Journal* 55 (1): 14-20.

Yoon, H. 2008. "More than a linguistic reference: the influence of corpus technology on L2 academic writing". *Language Learning and Technology* 12 (2): 31-48.

Basic vocabulary and absolute homonyms: a corpus-based evaluation

Isabella Chiari
Sapienza Università di Roma
isabella.chiari@uniroma1.it

1 Introduction

The basic vocabulary of Italian, first produced in 1980 (De Mauro 1980), has now been given a significant revision with a release to be published in 2013 in the form of a frequency dictionary with genre dispersion, derived from a 18ml word corpus of Italian (balanced on the following subcorpora: press, literature, non fiction, entertainment, computer mediated communication and spoken language). The *New Basic Vocabulary of Italian Language*, NVDB (Chiari and De Mauro 2012) has given the possibility of addressing for the first time the issue of absolute homonym disambiguation, more specifically homograph, in the definition of basic vocabulary.

2 The issue of homonymy and its corpus incidence evaluation

The evaluation of the presence of *absolute homonymy*, similar in some respects to polysemy as noted by Jespersen (1928) and also by Buyssens (Buyssens 1960; Buyssens 1965) and Katz (1972), also called *lexical homonymy*, has been rarely estimated in the form of corpus incidence, but more often assessed on the general dictionary lemma list (Doniyor 2009; Parent 2012). The definition itself of absolute homonymy is, for certain aspects, ambiguous, since it includes mostly homographic (or homophonic) lemmata, belonging to *the same part of speech*, that bear different diachronic evolution and separate etymologies (Martinet 1974), but in minor cases also includes forms with the same etymology that develop different significant meaning evolutions.

The issue of homonymy and it's relationship with polysemy is of capital importance both in corpus linguistics and in lexicography (Messelaar 1985; Mojela 2007; Moldovan 1981), as well as in computational linguistics. The problem is only partially addressed in word-sense disambiguation systems (Edmonds and Kilgarriff 2002; Ide and Véronis 1998).

On the level of dictionary lemma list, some estimations have been presented and they seem to be strongly dependent on each specific language structure, especially from the morphological point of view. Estimates on the English language, for example, consider about 2,500 lemmata absolute homographs in the Oxford English Dictionary (about 89% of which are monosyllabic) (Doniyor 2009), that is about 0.5% of the overall lemma list. Of these absolute homographs only 7% do not exhibit different etymologies. While in Korean the estimation reveals that about 30% of dictionary entries are (absolute) homonyms, mostly nouns (Kang 2005).

The issue is extremely relevant since it is also used for explanation of specific choices made in diachronic linguistics using the well known 'conflict of homonyms' theory by Gilliéron (Cruz Cabanillas 1997; Cruz Cabanillas 1999; Menner 1936), where only absolute homonyms are involved as belonging to the same part of speech, being capable of occurring in similar constructions, and belonging to close semantic spheres, making them capable of occurring in the similar contexts.

3 Homonymy in Italian Dictionary entries

For the Italian language general evaluations on absolute homography estimated on lemma list samples were give by De Mauro (De Mauro 1994). Looking more closely at absolute homography in GRADIT, *Grande Dizionario Italiano dell'uso* (De Mauro 1999), the largest existing lexicographic work on the Italian language, we find that about 2,6% homographic entries of all 260,709 lemmata are homographic, a smaller part of which are absolute homographs.

Vocabulary	Lemmata (L)	L. with H. homograph	%
GRADIT	260,709	6,765	2.6%
Common Lexicon[1]	60,333	4,384	7.2%
VDB (Basic Dictionary)	6,696	961	14.3%

Table1: Homograph incidence at lemma list level

In Table 1 percentage of incidence of lemma having at least one homographic entry is presented. All data is derived from GRADIT. As we can see the further we proceed to the core of most common and frequent word the higher the incidence of homographic lemmata we observe. But observing properties concerning frequency of use only on a flat lemma list is obviously only a partial portrait of absolute homography.

[1] Common vocabulary is a section of the global lexicon containing the most common 60,000 words of Italian, known by all native speakers having completed junior high school. Basic Vocabulary is the core of the language containing the top 5.,000 frequent words in frequency lists of spoken and written language, and about 1,700 lemmata called of high availability (as in the French terminological tradition (Michéa 1953).

4 Homonymy and the New Basic Vocabulary of Italian (NVDB)

The necessity of providing a frequency list of basic vocabulary lemmas has conducted to the need of differentiating homographic entries not only for grammatical or relative homographs, but also for absolute homographs (leading to the differentiation of lemmata by exponent numbers aligned to GRADIT).

The frequency list that generated the first two layers of basic vocabulary (fundamental and high frequency) containing the top-rank 5.000 lexemes in the list were manually disambiguated for absolute homonyms using corpus evidence and further aligned with the homonym tagging scheme of largest dictionary of Italian language (*Grande Dizionario Italiano dell'Uso*, GRADIT, De Mauro (1999). Absolute homonymy poses peculiar problems from a computational and linguistic point of view, since it involves homographic entries belonging to the same part of speech but having divergent etymologies and usage (e.g. *riso* "rice" and "laughter", *calcio* as "calcium" and "soccer", etc., and words that are homographic only in singular or in plural, as in *testi* being plural of *testo* "text" and of *teste* "witness"; *sale* singular for "salt" and plural of *sala* "room").

We chose to address the problem manually because we needed a reliable list that could also act as a golden standard for the development of automatic tools. The paper problematizes the enquiry on homography and offers the results of different types of absolute homographs and their incidence in the basic vocabulary and in the 18 million corpus and its subcorpora. While homography in citation forms (relative and absolute) of the 5.000 top-ranked lemmata has an incidence of about 5.515.930 (27.6% of the corpus) reaching up to five homographic lemmata attested, the evaluation of absolute homography involves more than 400 of 5.000 top lemmata (8% of the list) and reaches 866.180 occurrences, with coverage of 4.3% of the corpus.

Furthermore absolute homographic entries exhibit different attestation behaviours in the corpus (most entries are attested with large divergence in the whole corpus and in different subcorpora, only a few are evenly distributed) giving the possibility of evaluating different homographic classes, sense distribution and characteristic grammatical and collocational patterns on the base of usage. The analysis of different typologies of absolute homographs provides insight on the phenomenon of homography at corpus level and also gives interesting feedback to the description of the specific lexicographic entries and their ordering.

The general incidence of absolute homography in reference corpora (at least for Italian), especially at level of the top ranking lemmata, seems to suggest the necessity of its signalling in frequency based list and dictionaries, contrary to the observation of Wang and Nation for the academic word list (Ming-Tzu and Nation 2004).

References

Buyssens, E. 1960. "Le signe linguistique". *Revue belge de Philologie et d'Histoire.* 38: 705-17.

Buyssens, E. 1965. *Linguistique historique: homonymie, stylistique, semantique, changements phonetiques*. Bruxelles Paris, Presses Universitaires de Bruxelles / Presses Universitaires de France.

Chiari, I. and T. De Mauro 2012. "The new basic vocabulary of Italian: problems and methods". *Rivista di statistica applicata – Italian Journal of Applied Statistics* 22: 21-35.

Cruz Cabanillas, I. de la. 1997. "The Conflict of Homonyms Revisited". *Studia Anglica Posnaniensia* 32: 101-13.

Cruz Cabanillas, I de la. 1999. The Conflict of Homonyms: Does It Exist? *Cuadernos de Investigation Filologica* 25: 107-16.

De Mauro, T. 1980. *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*. Editori Riuniti, Roma.

De Mauro, T. 1994. "Quantità-qualità: un binomio indispensabile". *Capire le parole*. Laterza, Bari: 97-106.

De Mauro, T. 1999. *Grande dizionario italiano dell'uso*. UTET, Torino.

Doniyor, A. 2009. *Homonyms in English and their specific features*. Diplomnaja Rabota. Gulistan State University.

Edmonds, P. and A. Kilgarriff. 2002. "Introduction to the special issue on evaluating word sense disambiguation systems". *Natural Language Engineering* 8: 279-91.

Ide, N. and J. Véronis. 1998. "Introduction to the special issue on word sense disambiguation: the state of the art". *Computational Linguistics* 24: 2-40.

Jespersen, O. 1928. "Monosyllabism in English." *Proceedings of the British Academy*. Humphrey Milford.

Kang, B. 2005. Aspects of the Use of Homonyms. *Language Research* 41: 1-29.

Katz, J. J. 1972. *Semantic theory*. Harper & Row, New York.

Martinet, A. 1974. Homonyms and polysemes. *La Linguistique* 10: 37-45.

Menner, R.J. 1936. The conflict of homonyms in English. *Language* 229-44.

Messelaar, P. A. 1985. Polysemy and Homonymy in

Lexicographers: A Plea for Greater Systematization. *Cahiers de Lexicologie* 46: 45-56.

Michéa, R. 1953. Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage. *Les langues modernes* 47: 338-44.

Ming-Tzu, K.W. and P. Nation. 2004. "Word Meaning in Academic English: Homography in the Academic Word List". *Applied Linguistics* 25: 291-314.

Mojela, V. M. 2007. "Polysemy and Homonymy: Challenges Relating to Lexical Entries in the Sesotho sa Leboa-English Bilingual Dictionary". *Lexikos* 17: 433-39.

Moldovan, V. 1981. "Semantic Homonymy in Soviet Lexicography". *Analele Universitatii din Timisoara, Seria stiinte filologice* 19: 127-33.

Parent, K. 2012. The Most Frequent English Homonyms. *RELC Journal* 43: 69-81.

Using lockwords to investigate similarities in Early Modern English drama by Shakespeare and other contemporaneous playwrights

Jonathan Culpeper
Lancaster University
j.culpeper@lancaster.ac.uk

Jane Demmen
Lancaster University
j.demmen@lancaster.ac.uk

Shakespeare's plays occupy a uniquely prominent position in English drama, though the dialogue spoken by his characters has much in common with dramatic dialogue in plays by other contemporaneous playwrights. In this paper, we identify and discuss some similarities between Shakespeare's plays and plays by a range of his peers, using lockwords (Baker 2011). Our data comes from a corpus of Shakespeare's First Folio (adapted from Mike Scott's Shakespeare corpus), and a specialised parallel reference corpus of other Early Modern English ("EModE") plays (constructed by Demmen 2013 from digitised play-text files on *Early English Books Online*).

Keywords and other key language structures (i.e. those which occur with comparatively low or high frequency, statistically) are now well established in corpus linguistics as a way of investigating language styles in literary texts (as well as in other genres). Useful corpus stylistic research has been carried out into Shakespeare's plays using keyness, by, for example, Archer and Bousfield (2010), Archer et al. (2009), Culpeper (2002, 2009), and Scott and Tribble (2006). This has revealed new, empirically-based insights into character construction and characterisation in a single play (e.g. Archer and Bousfield 2010 and Culpeper 2002, 2009), and into the language styles of different plays (e.g. Scott and Tribble 2006). Keyness research has provided some valuable new, quantitatively-based perspectives to complement the vast body of mainly qualitative research into Shakespeare's plays in the literary critical tradition.

However, as Baker (2004:349) argues, keywords highlight only the differences between texts. Similarities are also important, to provide a wider context in which to see how language in particular texts, genres or periods is typically characterised. Yet language similarities have received much less attention than language differences in some areas of corpus linguistics, and it is only recently that methods specifically aimed at investigating similarities have begun to appear (see Taylor 2013 for a more detailed discussion). Corpus stylistic studies which use keywords to investigate language

style differences have, to date, focused almost exclusively on language differences. Ho's (2011) comparison of earlier and later versions of John Fowles' novel *The Magus* is a notable exception, in which she identifies language features which the novelist chose not to change, as well as those he did, with the aid of specialist software for comparing multiple versions of texts in addition to corpus linguistic keyness tools.

In our study, we aim to add some balance to existing keyness research by focusing on similarities between language styles in EModE plays. To do so, we apply Baker's (2011) lockwords concept, which he argues as being "the opposite of Scott's (2000) concept of keywords" (2011:73). Lockwords are high-frequency words which occur with the most similar frequency, statistically, in two texts or corpora. In his (2011) study, Baker identifies lockwords in four diachronic corpora of 20th century British English, and finds that the word *money* remains statistically significant over time. In contrast, we use lockwords with synchronic corpora, to investigate similarities (a) between Shakespeare's plays and plays by other dramatists of the same historical period, and (b) in the language styles of characters in one Shakespearean play.

Most existing corpus stylistic research into EModE drama, such as the keyness research mentioned above, does not extend beyond Shakespeare's plays to plays by other dramatists. Culpeper's (2011) study is an exception, in which he uses a large reference corpus of EModE drama, wide-ranging in date and author, to conduct an initial comparison of Shakespeare's language style in plays compared to those by other playwrights. This has been taken up in greater detail by Demmen (2013 and in preparation). She uses a parallel reference corpus of EModE plays, which balances Shakespeare's First Folio in size, dating, and genre components (comedy, tragedy and history), to identify similarities as well as differences in authorial styles, using the dual methods of keyness and Baker's locking concept.

Our paper demonstrates that focusing on language similarities through statistical "locking" is of potential interest and use at (a) the macro level: between Shakespeare and a range of his peers, by indicating preferences for language features which Shakespeare and other popular dramatists of his day also shared, and (b) the micro level: between a group of individual characters in a single play, by highlighting preferences for language features shared by characters. First, we present an analysis of lockwords (words with the most similar high frequency) in Shakespeare's plays and a range of plays by other popular and successful playwrights of his day. This highlights a number of shared preferences for language features in dramatic dialogue. Second, we examine lockwords in the dialogue of six characters in Shakespeare's tragedy *Romeo and Juliet*. This extends Culpeper's (2009) keyness study, which investigates characterisation through language differences using keyness.

Our findings add to what is known about the language that is typical of the text-type of EModE drama, and about Shakespeare's style in comparison to that of other contemporaneous playwrights. Our research is of potential interest to scholars working with keyness tools in corpus linguistics, and those working with EModE drama in linguistics and in other disciplines.

References

Archer, D. and Bousfield, D. 2010. "See better, Lear'? See Lear better! A corpus-based pragma-stylistic investigation of Shakespeare's *King Lear*". In D. McIntyre and B. Busse (eds.) *Language and Style*. Basingtoke: Palgrave Macmillan, 183-203.

Archer, D., Culpeper, J. and Rayson, P. 2009. "Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies". In D. Archer (ed.) *What's in a Word-List? Investigating Word Frequency and Keyword Extraction.* Farnham: Ashgate, 137-157.

Baker, P. 2004. "Querying keywords. Questions of difference, frequency, and sense in keywords analysis". *Journal of English Linguistics* 32 (4): 346-359.

Baker, P. 2011. "Times may change, but we will always have *money*: Diachronic variation in recent British English". *Journal of English Linguistics* 39 (1): 65-88.

Culpeper, J. 2002. "Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*". In U. Melander-Marttala, C. Ostman and M. Kytö (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium*, Association Suedoise de Linguistique Appliquée (ASLA), 15. Universitetstryckeriet: Uppsala, 11-30.

Culpeper, J. 2009. "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*". *International Journal of Corpus Linguistic*s 14 (1): 29-59.

Culpeper, J. 2011. "A new kind of dictionary for Shakespeare's plays: An immodest proposal". In M. Ravassat and J. Culpeper (eds.) *Stylistics and Shakespeare's Language. Transdisciplinary Approaches.* London: Continuum, 58-83.

Demmen, J.E.J. 2013. *A corpus stylistic investigation of the language style of Shakespeare's plays in the context of other contemporaneous plays.* PhD thesis, Lancaster University, U.K.

Demmen, J. In preparation. "From keywords to lockwords: Investigating similarities in the dialogue of plays by Shakespeare and other dramatists of his day".

Early English Books Online, 1475-1700. ProQuest LLC. 2003-2013. See http://eebo.chadwyck.com (last accessed 30.05.2013).

Ho, Y. 2011. *Corpus Stylistics in Principles and Practice. A Stylistic Exploration of John Fowles' The Magus.* London: Continuum.

Scott, M. 2000. "Focusing on the text and its key words". In L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective. papers from the third international conference on Teaching and Language Corpora.* Vol. 2. Frankfurt am Main: Peter Lang, 103-121.

Scott, M. and Tribble, C. 2006. *Textual Patterns. Key words and Corpus Analysis in Language Education.* Amsterdam: John Benjamins.

Shakespeare Corpus. Mike Scott. See http://www.lexically.net/wordsmith/support/shakespeare.html (last accessed 30.05.2013).

Taylor, C. 2013. "Searching for similarity using corpus-assisted discourse studies", *Corpora* 8 (1): 81-113.

Not all keywords are created equal: How can we measure keyness?

Václav Cvrček
Charles University in Prague, Czech Republic
vaclav.cvrcek@ff.cuni.cz

Masako Fidler
Brown University USA
masako_fidler@brown.edu

1 Introduction

Keyword analysis has become over the last two decades one of the most popular starting points in corpus-based analysis of parole. Keywords (i.e. the output of the analysis) are word forms obtained statistically by comparing the target text(s) (the focus corpus) with a larger corpus that reflects the general linguistic patterns of the language (the reference corpus) (Scott 1999, Baker and Ellece 2011). They are word forms that occur in a text more frequently than expected by chance alone and are often closely connected to the overarching themes and genre of a text or set of texts.

It is a wide-spread assumption among many linguists that the main task of statistical tests, which compare the difference in frequencies (e.g. chi-square, log-likelihood tests), is to **identify** keywords. However, what is less explored but equally important is to find a method for **ranking keywords**. The latter is especially crucial when dealing with large data-sets, which may yield thousands of statistically significant keywords with almost no chance to examine each one of them thoroughly and carefully. In such cases, if we are to examine only the highest ranked keywords (as it is common practice), the metrics (i.e. the method of ranking or sorting) we choose becomes of great importance.

This paper is an attempt to further advance the research of C. Gabrielatos and A. Marchi, which was presented at the CADS conference in Bologna 2012. In their search for appropriate metrics for keyword analysis they argued in favour of a measure, which they called %Diff. It is essentially the difference of normalised frequencies (in the target text and the reference corpus) divided by normalised frequency of the word-type in the reference corpus:

$$\%Diff = (norm.fq(target) - norm.fq(ref)) / norm.fq(ref)$$

Using convincing examples, the authors argued against using log-likelihood or chi-square tests (instead of e.g. %Diff) to measure the extent of the difference between the target and reference corpora. Although these statistical tests reveal the

significance level of the difference (in fact, they tell us only if the difference is based on a sufficient amount of data), they do not express the **effect size**: the degree of the difference, which would be comparable among all keywords and could serve as a basis for their ranking.

Similarly, Adam Kilgariff (2009) argued in favour of "simple math" against statistical tests for independence (not for methodological but for practical reasons). His approach to comparing two corpora stems from a simple idea that what we really want to find is the proportion of the two relative frequencies. Unlike Gabrielatos and Marchi, however, he uses the "add X" method to avoid dividing by null in cases where a word-type is present only in one of the corpora. This approach can be considered as controversial especially with regards to the value of X, which, added to both frequencies entering the formula, yields different results (if we use 1, 10, 100 for X, we obtain different rankings which might not be the result we hoped for).

2 Objectives

In this paper we propose different metrics for ranking by combining the best of the two approaches mentioned above. We modified **Dice's coefficient**, which is traditionally used for comparing different sets. For those words that are marked by a statistical test of significance (e.g. log-likelihood) as keywords, we count the following index, according to which we sort the results:

Dice = 2 * (norm.fq(target) – norm.fq(ref)) / (norm.fq(target) + norm.fq(ref))

In other words, Dice represents the proportion of the difference of relative frequencies to their mean. Its value ranges from -2 (when a word is present only in the reference corpus) to +2 (when a word is present only in the target corpus). The main advantage of this method is that we do not have to account for the null in denominator.

3 Methodology

To demonstrate the importance of distinguishing between significance level (expressed by log-likelihood for instance) and the effect size of the difference (represented here by our variation of Dice coefficient), we conducted series of tests on the New Years addresses (NYAs) by Gustáv Husák, the last communist president of Czechoslovakia (this is part of a larger project entitled *A Needle in a Haystack* that examines the limits and possibilities of keyword analysis). Husák's NYAs from the mid-1970s through the late 1980s serve as a sufficiently challenging material since they are relatively short and appear on the surface to be perfectly "flat" (the texts repeat the same clichés, winding ritualistic sentences, and appear to contain identical themes such as the five-year plans, the capitalist imperialism, and the leading role of the USSR). We have shown, however, that keywords indicate subtle changes in the society and politics.[1]

Each of these texts consists of approximately 1,500 words, yielding 30-50 keywords. This is a sufficient and yet manageable amount to carefully examine the difference in ranking by Dice and log-likelihood. Our methodology will be roughly two-tiered: we will (1) examine the possible correlation between the frequencies of the keywords in NYAs and the frequencies of these words in the reference corpus on the one hand and Dice and log-likelihood on the other, and (2) examine in detail those keywords that come out very differently in Dice and log-likelihood to see the relationship between these discrepancies and their discourse-semantic functions.

We have preliminary results to show that Dice is likely to be more efficient in differentiating the informational weight between the grammatical words that are normally uninteresting for keyword analysis and lexically richer topic words in each text (Dice seems to outperform here log-likelihood significantly).

Dice has been already implemented in the *KWords* tool, an application created for the *Haystack* project. [2] To illustrate the difference in ranking according to Dice and log-likelihood, Dice ranked a word form 'of peace (adjective)' much higher than log-likelihood in the NYA delivered in January of 1981. This is the year that ended in Martial Law in the neighbouring Poland (December), which was immediately followed by US economic sanctions against Poland and subsequently by the growing tension between the Eastern Bloc and the USA. Similarly, the word form 'disarmament' received a much higher ranking than log-likelihood in the NYA in January of 1985; this precedes Mikhail Gorbachev's ascension to power and the start of serious negotiations on disarmament. In other words, Dice can rank word forms that presage societal and political changes in this genre.

References

Baker, P. and Ellece, S. 2011. *Key terms in discourse analysis.* London: Continuum.

The Czech National Corpus (Available online at http://www.korpus.cz)

[1] Cf. our conference presentations: http://trost.ff.cuni.cz/keywords/texty/KANSAS-VCMUF.pdf; http://www.aatseel.org/program/2013-aatseel-conference-program/, http://trost.ff.cuni.cz/keywords/texty/cads2012.pdf.

[2] http://kwords.korpus.cz/.

Gabrielatos, C. and Marchi, A. 2012. *Keyness: Appropriate metrics and practical issues.* CADS International Conference, Bologna, Italy, 13-15 September 2012 (Available online at: http://www.gabrielatos.com/Presentations.htm)

Kilgarriff, A. 2009. Simple Maths for Keywords Proc. Corpus Linguistics, Liverpool. (Available online at: http://www.kilgarriff.co.uk/Publications/2009-K-CLLiverpool-SimpleMaths.doc)

KWords (Available online at: http://kwords.korpus.cz/)

Scott, M. 1999. *WordSmith tools help manual, Version 3.0.* Oxford: Mike Scott and Oxford UP.

Context-based approach to collocations: the case of Czech

Václav Cvrček
Charles University
vaclav.cvrcek@ff.cuni.cz

Anna Čermáková
Charles University
anna.cermakova@ff.cuni.cz

Lucie Chlumská
Charles University
lucie.chlumska@gmail.com

Renata Novotná
Charles University
renata.novotna@ff.cuni.cz

Olga Richterová
Charles University
richterova.olga@gmail.com

1 Introduction

This paper presents some preliminary results of our current research-in-progress: a context-based approach to collocations in Czech. Collocation is an old concept, which has received substantial attention in corpus linguistics and has become one of its central concepts. The importance of the concept is undisputed but its practical data-driven description and definitions of the term vary. McEnery and Hardie (2012: 122-123) sum up the idea in the following way: "the term *collocation* denotes the idea that important aspects of the meaning of a word (or other linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which it frequently co-occurs".

Most linguists would agree with this basic idea. From here onwards, however, the views differ widely. At one end is the theoretical concept of collocation to which the most central contribution has been the work of John Sinclair. At the other end are the n-grams, clusters or lexical bundles. The varied and extensive research so far confirms the importance of the concept both for applied and theoretical linguistics. However, in practical terms, there is a clear need for an operational definition of collocation and some kind of reliable automatic detection procedure from a large corpus.

2 Looking for collocations

Collocations, as various patterns of co-occurrence (definitions vary), are being automatically identified in the corpus with various statistical association measures, most frequently MI-score, t-score,

LogDice, salience [1]. All of them are based on frequency of occurrences of words, on probability of their co-occurrence and/or size of the corpus. From the point of view of a linguist examining corpus data, each of the association measures has some advantages and disadvantages and all of them produce quite a bit of "noise" while at the same time leaving out some desirable results. There is no perfect measure nor is there any agreement on what exactly counts as collocation. The quest for collocations remains an open task.

3 Word Sketches

Word Sketches, now part of the Sketch Engine (Kilgarriff et al 2004), are today one of the most popular programmes used for looking up collocations. Word Sketches were first developed for English and presented at the Euralex conference in 2002. Today they handle a number of languages. Word Sketch is a short, automatic, corpus-based summary of a word's grammatical and collocational behaviour (Kilgariff and Rundell 2002).

The program is now being used by professional lexicographers, most notably in English lexicography (they were used for the first time for the Macmillan English Dictionary). The importance of the concept of collocation for e.g. learners is recognized also by lexicographers and there are two specialized English collocation dictionaries: Macmillan Collocations Dictionary (2010) and Oxford Collocation Dictionary (2002).

4 Czech situation

For Czech, however, the situation is rather different. Not only is there no Czech collocation dictionary, Czech also lacks a modern monolingual dictionary. The Czech language has about 10 million speakers and dictionaries have not been the domain of commercial publishers but rather have been traditionally produced by the lexicographic department of the Czech Academy of Sciences. However, its newest dictionary is one published in 1994, which is a second edition of a 1978 dictionary.[2]

On the other hand, Czech has a unique, large dictionary of phraseology and idiomatics in four volumes (similes, nominal idioms, verbal idioms, and sentence idioms), an area, which certainly overlaps to a degree with the concept of collocation. There is also the Czech National Corpus available with several large, representative corpora mapping written as well as spoken Czech.

5 New concept: P-collocations

We have recently embarked on a research project of Czech collocations with the aim of compiling a collocation dictionary of Czech. Czech belongs to Slavic languages, has free word-order and is highly flective. As such it requires a specific approach to the identification of collocations. We have been extensively testing the Word Sketches programme for this purpose. The programme produces promising results for nouns but with adjectives and verbs the results are significantly weaker.

Consequently, we started testing a new method of identifying collocations without need of applying grammar rules. The so-called "p-collocations" (as in proximity, see below) are extracted from the corpora with the use of a new programme, developed at the ICNC for this purpose. The method is based on two measures, obligatoriness and proximity (Cvrček forthcoming).

6 Obligatoriness and proximity

These two measures are derived from the quantitative properties of a context of language units. Immediate context positions of a word (first positions to the left and right of the node word) tend to have the lowest variability (i.e. a small number of different word-types). This is explained by the semantic constraints the keyword puts on its immediate context, which can be thus seen as a source of the most relevant information about its syntagmatic properties including local grammar. Thus the average distance between words (word understood in the most basic sense as a textual unit divided by spaces) is very relevant. Although it might sound rather simplistic, interim results show that the closer the words are, the more meaningful their combination is likely to be. From this basic assumption we derive the above mentioned variable 'proximity', which is defined as the average distance of two words in a span of seven words and, unlike other common association measures, is based solely on the mutual position of the two words within a given span.

Proximity is complemented by the second variable "obligatoriness", which is defined as the higher value of two fractions: frequency of the first word to the frequency of the combination as a whole and frequency of the second word to the frequency of the combination. The research on proximity so far suggests that proximity is a reliable indicator of a syntagmatic relation of two words while obligatoriness shows how stable the word combination is.

[1] See Evert's summary available at www.collocations.de.

[2] There is also a new, smaller dictionary of Czech (30 thousand headwords), published by a commercial publisher in 2011, but this misses the concept of collocation and corpus-based lexicography in general almost entirely.

7 Comparison of Word Sketches and P-collocations

This study will compare the output of Word Sketches for selected Czech and English words with the output of the new method introduced above. It will also argue in favour of a two-dimensional chart of proximity and obligatoriness (representing the syntagmatic aspect of collocations as well as their mutual semantic interconnection) as a suitable starting point for any collocational analysis.

References

Cvrček, V. (Forthcoming): *Kvantitativní analýza kontextu (Quantitative analysis of context)*. Praha: Nakladatelství Lidové noviny.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Euralex Proceedings,* Lorient, France, pp. 105-116.

Kilgarriff, A. & Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications – A Case Study. In *EURALEX Proceedings*, Copenhagen, Denmark, pp. 807-818.

Macmillan Collocations Dictionary (2010). Oxford: Macmillan Education.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics.* Cambridge: Cambridge University Press.

Oxford Collocation Dictionary (2002), Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Sinclair, J., Jones, S., Daley, R. & Krishnamurthy, R. (2004). *English Collocational Studies: The OSTI Report.* London: Continuum.

Slovník české frazeologie a idiomatiky I-IV. (2009). Ed. Čermák, F. Praha: Leda.

Slovník spisovné češtiny pro školu a veřejnost (1978, 1994). Praha: Academia.

Slovník současné češtiny (2011). Praha: Lingea.

A corpus-based study on the relationship between word length and word frequency in Chinese

Deng Yaochen
Dalian Maritime University
Deng_yaochen@163.com

Feng Zhiwei
Hangzhou Normal University
zwfengde2010@hotmail.com

1 Introduction

As tools for "conveying meanings in order to achieve objects" (Zipf 1949:20), words play a prominent role in verbal communication. With the recognition of the significance, a great number of studies have been conducted to investigate different aspects of vocabulary in the past two decades. While inspiring results have been achieved in terms of word formation, textual behaviour of lexical units, vocabulary acquisition and psychological processing, the dynamics of lexicon in language use has not received the attention it deserves. As a result, we know little about the principles and regularities governing the interaction between different properties of lexical units as well as the interaction between lexicon and system requirements.

The present study, based on data from modern Chinese corpus, intends to investigate the dynamic behaviour of lexical units in a quantitative linguistic paradigm. Particular attention is paid to the dynamics of lexicon in language use as well as the synergetic nature of language as a self-organising and self-regulating system. The results are interpreted within the theoretical framework of Synergetic Linguistics, aiming to reveal the universal nature of this linguistic phenomenon.

Synergetic Linguistics is an interdisciplinary approach to the modelling of certain dynamic aspects of the language system. It is theoretically founded on the view of language as a psycho-social phenomenon and a biological-cognitive one at the same time. The fundamental axiom of this theory is that "language is a self-organizing and self-regulating system – a special kind of dynamic system with particular properties" (Köhler 2005:760). So, in Synergetic Linguistics, language is characterized by the presence of cooperative and competitive processes, just like other self-organizing systems (Köhler 1993:41). The result of these processes is an optimal steady state of the language systems and an optimal adaptation to its environment.

The primary goal of Synergetic Linguistics is to systematize quantitatively the self-regulation in

language, a system which is seemingly disordered, and to capture linguistic laws by means of mathematically expressed dependencies (Feng 2012:267). Guided by the principles and research aims of Synergetic Linguistics, the present study, adopting a corpus-based approach, endeavours to explore the dynamics of lexical units in Chinese. Special efforts were made to reveal and model, in a mathematical way, the regularity of the influence from word length on word frequency. The dependence between these two lexical properties was investigated not only in different parts of speech but also in different registers of modern Chinese discourse. Specifically, the following questions are addressed: (1) How does word length influence word frequency in Chinese? (2) Does word length exert similar influence on functional words as on notional words in Chinese? (3) What are the mathematical models which can best capture the regularities revealed in (1) and (2)? (4) How do the parameters of the mathematical models vary with the change of registers of discourse?

2 Methods

All the data come from the Lancaster Corpus of Mandarin Chinese (LCMC) and the Spoken Corpus of Mandarin Chinese (SCMC), which are comparable in size but representative of standard written and spoken Chinese, respectively. Both LCMC and SCMC are complete homogenous corpus. A series of computer programs were specifically written for corpus processing and data collection, including tokenization, segmentation of word tokens and POS-tags, calculation of word frequency, measurement of word length, computation of the number of word types with the same word frequency and word length.

SPSS (v20) was employed to model the relationship between these two variables.

A Chinese word can include one or more Chinese characters. In Chinese, a Chinese character is generally a syllable. Therefore, the word length of Chinese can be calculated by Chinese character number which is included in the word.

3 Results and discussion

Panel A and Panel B in Figure 1 indicate the dependence of word frequency (FREQ) on word length (WL) in written and spoken Chinese, respectively.

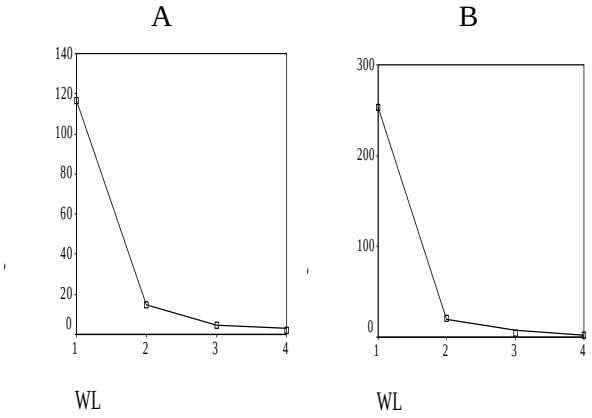


Figure 1. Dependence of word frequency on word length in written and spoken Chinese

Panel C and Panel D in Figure 2 show the dependence of word frequency on word length in notional words and function words in written Chinese, respectively.

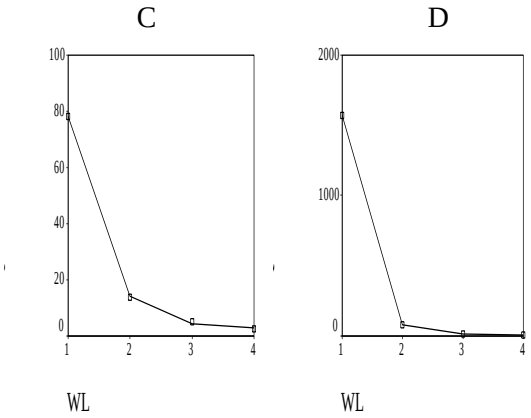


Figure 2. Dependence of word frequency on word length in notional and function words

As the results presented in Figure 1 and Figure 2 show, there exists a high dependency of word frequency on word length. The longer a word is, the less frequently it is used in discourse. The results suggest an inverse relation between these two properties. The power model $y=ax^b$ is proved to fit best the data and capture this regularity.

The results further indicate that the dependence is also present in the parameter of this model, *a*, on the register of discourse. The value of *a* for speech is significantly higher than that of writing. It is also observed to approximate the frequency of words with single morpheme. Hence, the parameter of this model, *a*, is powerful in distinguishing the texts of different styles.

The results of the study not only complement the current theories on the relationship between word length and word frequency, providing new evidence for the relationship as a linguistic universal, but also offer a new paradigm for style identification and text classification.

References

Feng, Z. 2012. Studying language by quantitative methods. *Foreign Language Teaching and Research* (bimonthly). Beijing. 256-269.

Köhler, R. 1993. Synergetic Linguistics. In Köhler, R. & Rieger, B. (eds.) *Contributions to quantitative linguistics.* Dordrecht: Kluwer Academic Publishers. 41-52.

Köhler, R. 2005. Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) *Quantitative Linguistics. An International Handbook*. Berlin: Walter de Gruyter. 760-774.

Strauss, U., P. Grzybek & G. Altmann. 2007. Word length and word frequency. In P. Grzybek (ed.) *Contributions to the Science of Text and Language*. Dordrecht: Springer. 277-294.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

"Anyway, the point I'm making is": relevance marking in lectures

Katrien Deroey
Ghent University
katrien.deroey@Ugent.be

Drawing on the British Academic Spoken English (BASE) [1] lecture corpus, this paper presents an overview of how important and less important discourse is marked lexicogrammatically (cf. Deroey and Taverniers 2012a; Deroey and Taverniers 2012b). Such markers of (lesser) relevance (e.g. *anyway*, *the point is*) are metadiscursive devices which combine discourse organization with evaluation along a 'parameter of importance or relevance' (Thompson and Hunston, 2000: 24). Relevance marking can help students discern the relative importance of points and so may aid comprehension, note-taking and retention. However, until recently very little was known about this feature of lecture discourse and the few markers that can be found in educational literature and most English for Academic Purposes (EAP) listening materials seem based on intuitions rather than corpus linguistic evidence.

Both studies are based on a close reading of 40 lectures to identify candidate markers which were then retrieved from the whole corpus of 160 lectures using Sketch Engine. In addition, for the study on relevance markers results were supplemented by items from the BASE word list and previous lecture research (Swales and Burke 2003; Crawford Camiciottoli 2004); markers discovered in the co-text of concordances were also added, as were words derived from or synonymous with all lexemes found through the above procedures. Interestingly, the manual analysis of 40 lectures yielded the vast majority of all markers.

The research on relevance markers revealed a wide variety of markers, the most frequent of which are not amongst those which may intuitively come to mind or which are typically included in EAP materials. The markers could be classified into different lexicogrammatical patterns based mostly on nouns (e.g. *the important point is*, *the thing is*), verbs (e.g. *remember*, *let me just emphasise*) and adjectives (e.g. *it is important to note*, *this is absolutely crucial*). Adverb patterns are extremely rare (e.g. *importantly*), as are expressions referring

[1] The BASE corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. The corpus is available from the Oxford Text Archive http://ota.ox.ac.uk/headers/2525.xml.

to assessment (e.g. *it is something that you can be examined on*). The verb pattern 'V clause' (e.g. *remember slavery had already been legally abolished*) and the noun pattern 'MN v-link', a metalinguistic noun with a link verb (e.g. *the point is*) are the predominant types of relevance markers.

Markers of lesser relevance were classified into five broad types according to how they signal lesser relevance: (i) message status markers assign a negative value in terms of relevance to part of the lecture message (e.g. *not pertinent, joke*) or signal transitions between more and less relevant discourse (e.g. *anyway*); (ii) topic treatment markers (e.g. *briefly, not look at, for a moment*) indicate limited discourse or time is devoted to a topic; (iii) lecturer knowledge markers (e.g. *not know, not remember*) suggest the lecturer has imprecise or partial knowledge about the topic; (iv) assessment markers (e.g. *not examine, not learn*) indicate what information will not be examined; and (v) attention- and note-taking markers (e.g. *ignore, not copy down*) direct students not to pay attention to or take notes of what is presented. Most denote partial relevance (e.g. *detail, in passing, briefly*) rather than irrelevance (e.g. *not matter, trash*) and some markers appear pragmaticalized in certain contexts. For instance, markers denoting limited coverage (e.g. *briefly, quickly, a little bit*) can serve as mitigation devices. As most markers require some or substantial interpretation to achieve their relevance marking effect, an understanding of the main characteristics and purposes of the lecture genre as well as co-textual, visual, non-verbal and prosodic clues seem particularly important in identifying the function of these lexicogrammatical items but poses a challenge to quantification. Indeed, Hunston's observation that 'much evaluative meaning is not obviously identifiable, as it appears to depend on immediate context and on reader assumptions about value' (2004: 157) is particularly pertinent here.

The research presented here should interest anyone interested in spoken (academic) discourse, evaluative language, identifying discourse functions in corpora, and EAP course design for lecture listening and delivery.

References

Deroey, K. L. B., & Taverniers, M. (2012a). 'Just remember this': Lexicogrammatical relevance markers in lectures. *English for Specific Purposes, 31* (4), 221-233.

Deroey, K. L. B., & Taverniers, M. (2012b). 'Ignore that 'cause it's totally irrelevant': Marking lesser relevance in lectures. *Journal of Pragmatics, 44* (14), 2085-2099.

Crawford Camiciottoli, B. (2004). Audience-oriented relevance markers in business studies lectures. In Del Lungo Camiciotti, G., & Tognini Bonelli, E. (Eds.), *Academic discourse: New insights into evaluation* (pp. 81-98). Bern: Peter Lang.

Hunston, S. (2004). Counting the uncountable: Problems of identifying evaluation in a text and in a corpus. In Partington, Morley, A. & Haarman, L. (Eds), *Corpora and discourse* (pp. 157-188). Bern: Peter Lang.

Swales, J. M. & Burke, A. (2003). "It's really fascinating work": Differences in evaluative adjectives across academic registers. In Leistyna P., & Meyer, C. F. (Eds.), *Corpus Analysis: Language structure and language use* (pp. 1-18). Amsterdam: Rodopi.

Thompson, G., & Hunston, S. (2000). Evaluation: An introduction. In Hunston, S., & Thompson, G. (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 1-27). Oxford: OUP.

Visualizing *chunking* and *collocational networks*: a graphical visualization of words' networks

Matteo Di Cristofaro
Lancaster University
m.dicristofaro@lancaster.ac.uk

The notions of *chunking* and *collocational networks* are central to linguistics (e.g. Bybee 2010; McEnery 2006, 2012); *chunking* has been described as follows: "When two or more words are often used together, they […] develop a sequential relation, […known] as 'chunking' […]. The strength of the sequential relations is determined by the frequency with which the two words appear together. […] The frequency with which sequences of units are used has an impact on their phonetic, morphosyntactic and semantic properties." (Bybee 2010, 25). The notion of c*ollocation* has been defined in many ways (McEnery and Hardie 2012, 122-123); in general "[…]the term *collocation* denotes the idea that important aspects of the meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which it frequently co-occurs" (ibid.).

When dealing with corpus data, the analysis of the aforementioned features can prove to be difficult as the results can depict a dense network of relations between words. Although visual representations of such networks have been attempted (see e.g. McEnery 2006), these methods had to deal with a series of limitations such as the media (printed paper) and quantity of data. Furthermore they tend to be more descriptive than analytical. What I propose is a method to visualize these networks that not only allows for descriptive graphical representation, but that can also be used to conduct a more detailed analysis of the data. In order to do so, I use the open source software Gephi (https://gephi.org/), a tool for visualization and exploration of any kind of networks. I demonstrate how the results drawn from corpus analysis can be "converted" and used as input for the software in order to obtain 2D or 3D graphs of the relations between a (virtually infinite) number of words. Furthermore the resulting graph is interactive, and therefore allows for the quick highlighting and isolation of specific word(s) and its/their relations, or for filtering of the data on the basis of specific statistical parameters. Any corpus data can be converted into a compatible Gephi format. As a case-study to show the basic steps of the method, I will show the results of the analysis I carried out for my PhD thesis on the study of Italian taboo language and taboo language constructions. The data was first retrieved and analysed through the web-interface SketchEngine (https://the.sketchengine.co.uk), and the results (i.e. the saliency scores provided by SketchEngine, based on logDice: see Kilgarriff 2012) were then converted into a compatible Gephi format through an ad-hoc script and loaded into the software. This allowed me to display networks with hundreds of "nodes" (each word is counted as a node in the network) and then to visually analyse the presence of chunking and/or (bidirectional) collocations by means of the different tools and filters available in the software (fig.1-2)

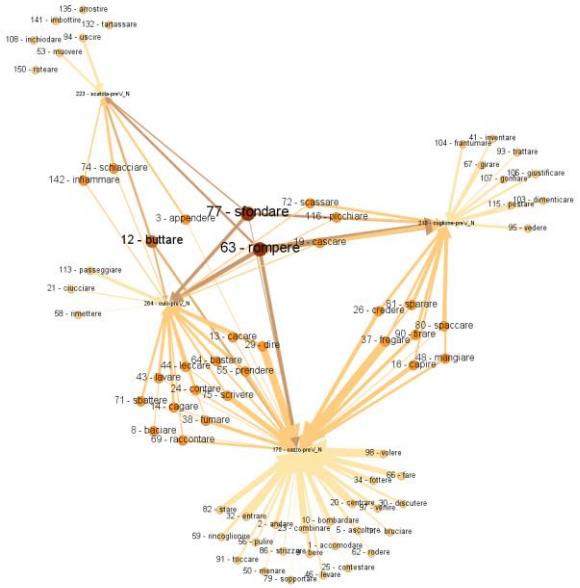


Figure 1: Visual rendition of collocational networks

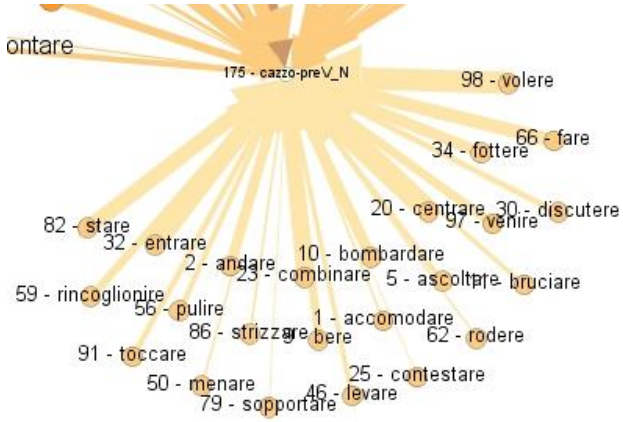


Figure 2: Details of the collocations of "cazzo" (lit. dick) in pre-verbal position.

Among the different tools, it is possible to set a threshold so that only those words with a minimum or maximum value (in my case this is the saliency score) are shown; or a word can be isolated in order to show only the relations to or from that specific word. Furthermore, the different nodes in the network are distributed on the basis of a chosen

parameters, so that the grouping of the different nodes can show specific features based on the raw data. This set of features has proven to be useful when determining the role of the word in the network, and its relations with the rest of the words. For my research I analysed the "status" of a word used in the corpus through the visual representation. I propose in my thesis that a distinction between *euphemisms* and *dysphemisms* (i.e. the contrary of *euphemisms*) can be drawn on the basis of the relation of the analysed word with the words with which it is used. The use of the method I propose has not only allowed me for the identification of *sequential relations* (see definition of *chunking* provided at the beginning) between words, and of *collocational networks*. It has also allowed me to conduct a series of more detailed statistical analysis on words which do not "stand out" among the data, but show an interesting behaviour to my research such as groups of words which are strictly related to each other so that they visually appear as a single node. By looking at the raw data for the words in these groups it was possible to establish that they visually appear together as they do share the same meaning when used with the word they all relate to. Features such as this would require a longer and more elaborate process to be spotted if the data was displayed through more "traditional" methods of representations (e.g. lists such as keywords lists). As the method I outlined is able to put together large sets of data and to triangulate the details of each word with the details of all the other words, it is possible to analyse large quantities of data just by looking at the visual rendition.

References

Bybee, J. 2010. *Language, Usage and Cognition*. CUP.

Kilgarriff, A. 2012 *Statistics used in the Sketch Engine*, available online at http://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

McEnery, T. 2006. *Swearing in English. Bad language, purity and power from 1586 to the present.* Routledge.

McEnery, T., Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. CUP

Using learner corpus tools in second language acquisition research: the morpheme order studies revisited

Ana Díaz-Negrillo
Universidad de Granada
anadiaznegrillo@ugr.es

Cristóbal Lozano
Universidad de Granada
cristoballozano@ugr.es

1 Revisiting the morpheme order studies

The so-called Morpheme Order Studies (MOS) conducted in the 70s have been crucial in our understanding of learner language (i.e., interlanguage) in the second language acquisition (SLA) of English. Researchers found a remarkably consistent sequence: some morphemes were produced/acquired earlier than others in L2 English (Table 1), independently of the learners' mother tongue (L1), age and learning environment. Results were also similar to what had been previously observed in child L1 English acquisition (for overviews, see Hawkins and Lozano 2006; Kwon 2005). This topic has received considerable interest recently in SLA research (Goldschneider and DeKeyser 2001; Kwon 2005; Luk and Shirai 2009).

Sequential order	Morpheme
1	progressive *–ing*
2	contractible copula –'s
3	plural *–s*
4	articles *a(n)/the*
5	contractible auxiliary (be) –'s
6	irregular past
7	regular past *–ed*
8	3rd person singular *–s*
9	possessive *–s*

Table 1: Sequence of L2 English morpheme acquisition

SLA researchers have typically used (quasi)experimental methods in the MOS, but the necessary next step is the use of learner corpus analysis. By triangulating data from previous studies (experimental data) with new corpus data (naturalistic data) and corpus-based tools (fine-grained annotation) we can get a fully-rounded picture of the acquisition of L2 English morphemes. We set off from the assumption that replication in SLA is a necessary condition to (dis)confirm previous findings and to eliminate possible biases in the research method (Porte 2012).

2 Methodological limitations of previous research

Morpheme acquisition has been measured in previous MOS with different methods: Suppliance in Obligatory Occasions (SOC), Target-Like Use (TLU) and score-based SOC (Ellis & Barkhuizen 2005 for overviews). One of the limitations of previous research is that experimental data are typically drawn from small learner samples under controlled conditions. In addition, SOC and TLU methods are coarse-grained in their analysis of leaner data since they do not fully explore the subtypes of errors produced by learners (**stealed, *stoled, *foots, *feets*, etc.).

3 MOS meet LCR

We replicate previous MOS findings but within Learner Corpus Research (LCR) since synergies between corpus linguists and SLA researchers are essential to fully understand interlanguage processes (Tono 2003). In so doing, we aim to compensate some limitations of MOS and LCR by combining the methodological strengths of LCR and the theoretical explanatory power of SLA in MOS. This requires, on the one hand, using larger amounts of naturalistic data (learner corpora). On the other hand, this also means broadening common research practice in LCR in two respects:

- First, while LCR has been mostly hypothesis-finding, resulting in largely descriptive accounts of learner performance (Myles 2005, 2007, Granger 2009), we attempt at hypothesis-testing, which involves setting off from a clearly defined theoretical basis with a view to empirically explaining developmental processes in SLA.
- Second, the methodological approaches in LCR have been mostly based on the contrastive analysis of learners' performance (commonly known as Contrastive Interlanguage Analysis, CIA), and the holistic, coarse-grained analysis of learners' errors (Computer-aided Error Analysis, CEA) (Granger 2008). This paper suggests a wider approach to the analysis of learner corpus data by studying not just problematic areas, but also correct uses in order to arrive at a better understanding of developmental processes in SLA. This must be examined with fine-grained, purposed-oriented annotation.

4 Towards fine-grained annotation

Corpus annotation is used to identify our units of analysis (grammatical morphemes). Annotation increases the potential of the corpus as it uncovers linguistic properties which can be then searched and quantified using corpus software. Various types of corpus annotation are possible in learner corpora, ranging from automatic annotation of the grammatical categories (POS tagging) to semi-automatic annotation of learner language properties. The latter has mostly been based on the identification and general descriptions of a variety of learner's errors (see Díaz-Negrillo & Fernández-Domínguez 2006 for an overview of error-annotation schemes).

We build on common annotation practices in learner corpora. But we argue for a type of annotation that can disclose a wider picture of specific features of learner's interlanguage, that is, tagging that (i) is purpose-oriented, (ii) is fine-grained and (iii) describes not just learners' subtle errors but also their correct uses.

This detailed analysis overcomes the second limitation of MOS cited above: coarse-grained measurement of morphemes. We do so by presenting a more fine-grained, ad-hoc tagset for each morpheme that takes into account the learners' morphological interlanguage processes (e.g., Figure 1 illustrates the tagset template with examples from irregular past). Our tagset builds on previous studies that have taken into account target-like use and non-target-like use, and, in some cases, overuse. But we refine non-target-like uses (underuse, misselection, misrealization [single marking, double marking] and overuse), which would potentially allow researchers to investigate well-known interlanguage processes such as the Dual Mechanism of irregular morphology processing (4) vs. (5), which are generated by different mechanisms.

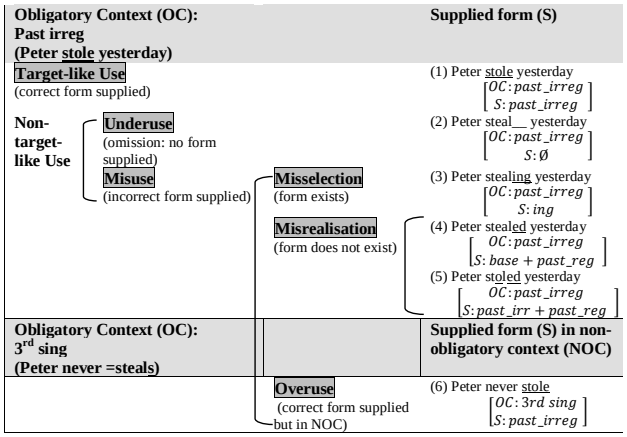


Figure 1: Tagset for irregular past

5 Our corpus analysis

To illustrate this, we briefly explore morphemes in the COREFL, a on-going corpus of L1 Spanish-L2 English as a Foreign Language in secondary schools

at several proficiency levels (A1, A2, B1, B2, C1), amounting to around 100,000 words. The grammatical morphemes in Table 1 were tagged in our corpus using the multi-layered annotation software UAM Corpus Tools (O'Donell 2009), where frequency analyses were performed.

Morphemes have been previously investigated in LCR by Tono's (2000) pioneering study on the acquisition morphemes in an L1 Japanese-L2 English corpus (JEFLL corpus) (see also McEnery, Xiao and Tono's 2006). Unlike previous MOS that underplayed the role of the L1, Tono found a strong L1 influence in the production certain morphemes (e.g., articles) (Figure 2). Our preliminary findings from COREFL for all proficiency levels (Figure 3) indicate that 3rd person singular *–s* is particularly problematic for L1 Spanish-L2 English learners, while other morphemes that were shown to be acquired early in the MOS (e.g., progressive *–ing*) are acquired later than expected. Additionally, unlike Tono's (2000) learners, our learners' accurate production of articles is higher. Additional findings reveal that certain morphemes are problematic across proficiency levels (though their production rates vary due to proficiency).

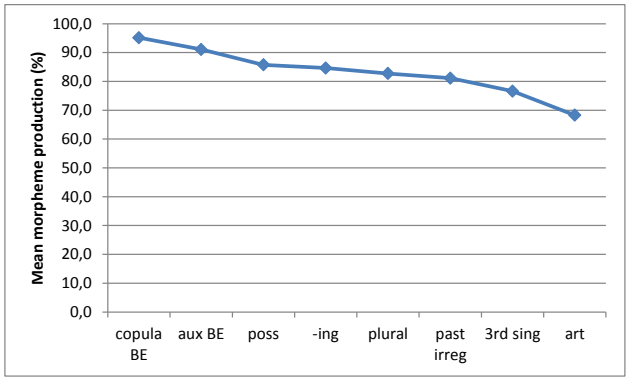


Figure 2: Tono (2000) (all proficiency levels)

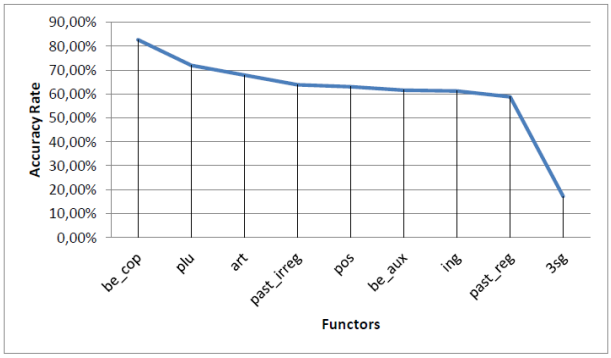


Figure 3: Our preliminary results (all proficiency levels)

Overall, this brief sample of our findings partially supports previous MOS results, but also diverges in systematic ways. The next step is to take a hypothesis-testing approach to LC to provide an SLA explanation of the observed accuracy profiles. In this way, we will arrive at a better understanding of the interlanguage processes underlying the acquisition of L2 English morphology.

References

Díaz-Negrillo, A. & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *RESLA*, 19: 83-102.

Ellis, R., & Barkhuizen, G. P. (Eds.). (2005). *Analyzing Learner Language*. Oxford University Press.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 259–275). Berlin: Mouton de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1), 1–50.

Hawkins, R. & Lozano, C. (2006). Second Language Acquisition of Phonology, Morphology and Syntax. In: K. Brown (ed). *The Encyclopedia of Language and Linguistics* (2nd Edition) (pp. 67-74). Oxford: Elsevier.

Kwon, E.-Y. (2005). The "Natural Order" of morpheme acquisition: A historical survey and discussion of three putative determinants. *Columbia University Working Papers in TESL & Applied Linguistics*, 5(1), 1–21.

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order or grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -s, articles, and possessive 's. *Language Learning*, 59(4), 721–754.

McEnery, T., Xiao, R., & Tono (eds) (2006). *Corpus-Based Language Studies: An Advanced Resource Book.* Routledge.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373–391.

Myles, F. (2007). Using electronic corpora in SLA research. In D. Ayoun (Ed.), *Handbook of French Applied Linguistics* (pp. 377–400). Amsterdam: John Benjamins.

O'Donnell, M. (2009). The UAM CorpusTool: Software for corpus annotation and exploration. In C. M. Bretones & et al (Eds.), *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Actual: Comprendiendo El Lenguaje y La Mente* (pp. 1433–1447). Almería: Universidad de Almería.

Porte, G. (Ed.). (2012). *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.

Tono, Y. (2000). A computer learner corpus based

analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 123–132). Peter Lang.

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the 2003 Corpus Linguistics Conference* (pp. 800–809). UCREL, Lancaster University: UCREL Technical Paper number 16.

Risk, chance, hope – the lexis of possible outcomes and infertility

Karen Donnelly
Lancaster University
k.donnelly@lancaster.ac.uk

1 Introduction and rationale

There is currently a large body of scholarly work exploring the semantics of risk (Lupton, 2004), including studies using corpus methodology to compare the semantic preference associated with risk (Hamilton et al, 2007, Hardy & Columbini, 2011). Risk is frequently associated with a medical prosody and it plays a central role in studies of health communication. The growing emphasis on choice in health care and a move to "consumer led" model leads to a heightened sense of the need for knowledge of potential risk to be shared. According to Fillmore and Atkins (1994), risk "belongs to the group of English words whose semantic descriptions share some reference to the *possibility* of an *unwelcome outcome*", which is often the case in communications on health. However there is so far a dearth of work on other linguistics expressions of uncertainty (i.e chance), where in essence there may be possibility of a *welcome* outcome.

This study uses corpus methodology to explore the semantic prosody of both risk(s) and other lexis of possible outcomes which were found to be unusually frequent in one or more specialist corpora of texts on infertility. Infertility was selected as a topic for this study due to its problematic position as both a social and medical issue (Greil et al., 2010) and the prevalence of media, medical and personal texts which proliferate around it.

2 Methodology

The data for this study comprises three specially built corpora of texts on infertility including; UK newspaper articles from 2006 – 2012 containing the term infertility (NEWS – 5, 259, 717 tokens), websites for fertility clinics from 2012 (CLINIC – 1, 277, 736 tokens) and UK blogs written by people experiencing infertility from 2006 – 2012 (BLOG – 1, 604, 725 tokens).

Initial analysis was carried out using Wordsmith Tools to elicit the top 100 lexical keywords from each corpus, when compared to UKWaC 10m subset. These keywords were then grouped thematically in order to allow the comparison of particular themes and preoccupations across the corpora and guide the selection of lexis for further study using collocations and concordance lines.

During the initial keyword analysis the terms *risk* and *risks were* found to be "key" in both the NEWS corpus and CLINIC corpus, however, these terms were noticeably absent from the BLOG corpus. Examining other keywords which carried a prosody of possible and uncertain outcome both NEWS and CLINIC corpora contained *chance* and *chances* as keywords which were also absent from the BLOG corpus keyword list. A further examination of the BLOG keywords showed that the term *hope* as a linguistic marker of possible outcomes and was selected for further analysis along with *risk(s)* and *chance(s).*

This further analysis looked at collocates (top 50 which occurred at least 5 times in 5L and 5R span of node word with MI 3 and LL 15.3)and concordance lines of the selected keywords in order to determine what Partington's (2004) account of semantic prosody describes as the "resulting affective meaning" of the words. This analysis focused on the negative or positive semantic prosody of each of the node words, the collocates used to describe possible outcomes and the actors who are involved in these outcomes.

3 Frequency

Initial frequency counts of the node words showed a difference in lexical choices , with a varied distribution of frequency/million words according to text type, with the NEWS corpus using *risk* most frequently, the CLINIC corpus showing a preference for both *risk* and *chance* whilst the BLOG corpus had the highest frequency of *hope* (see Table 1 for full breakdown).

	NEWS	CLINIC	BLOG
Risk	772.86	820.20	110.30
Risks	219.21	555.67	33.65
Chance	463.33	739.59	386.36
Chances	297.35	350.62	117.15
Hope	333.67	248.88	853.73

Table 1. Frequency of node in each corpora (per million words)

4 Semantic prosody

The top 50 collocates of the node words revealed a cline of negative/positive semantic prosody with *risk* carrying the most negative prosody with high levels of collocates of illness i.e. *cancer, heart disease, diabetes,* problematic reproduction i.e. *miscarriage, ectopic* and also more abstract negative terms such as *harm, complications, issues* and *defects.* Comparatively the collocates of risks tended to the management and knowledge about possibly negative outcomes i.e. *outweigh, aware, minimise. Hope* carries a less negative prosody as it embraces the possibility of a welcome outcome, however, it also carries a prosody of desperation through modifiers such as *just, only* and *really*. *Hope* if realised i.e. *given* or *offered* it is positive but if it is not realised then the prosody is more negative. Moving along the cline *chance(s)* is located at the most positive end with strong collocates such as *good, best,* and *success* suggesting a possibility of a welcome outcome, however a potentially negative prosody of *chance(s)* is carried in collocates such as *denied, robbed* and *get* suggesting that this chance can be given but can be taken away. It is also suggested that *chance* is finite as in *last chance saloon.*

5 Welcome/unwelcome outcomes

One of the strongest L1 collocate of all node words is *of,* therefore the concordance lines of the compounds *risk(s) of, chance(s) of, hope of* were studied to uncover possible welcome/unwelcome outcomes related to the node words.

The outcomes of risk(s) show the most variation, particularly in the NEWS corpus, ranging from generic medical; *cancer, heart disease,* reproductive; *infertility, miscarriage, stillbirth, multiple pregnancy,* to lifestyle; *divorce, financial problems.* The general health risk collocates are far more likely to occur in the NEWS and BLOG corpus, whilst unsurprisingly the CLINIC corpus focuses on risks related to reproduction, and more specifically reproductive technologies. The outcome of *risk* also co-occurs to the category of "problematic reproduction" encompassing such terms as *birth defects* and *inherited conditions,* suggesting that infertility as a problem is semantically linked with anything other than the birth of a "perfect" baby, the ultimate desirable outcome.

In the case of *chance(s)* by far the most frequent outcome described is *pregnancy* and related items such as *getting pregnant* and *having a baby,* which are also captured within the generic term *success.*

While the BLOG texts do show a co-occurrence between welcome outcomes such as *pregnancy* and *success* with chance and chances, both relate to how unlikely these are through terms such as *slim, low* and *f*ck all.* In contrast both NEWS and CLINIC emphasise the *good, high* and *increase* in terms of the chance and chances of a welcome outcome (successful pregnancy).

The BLOG corpus again shows strong collocation between *hope of* and having a child, however through collocates such as *miracle* demonstrate the view that there is a low possibility of achieving this outcome. The NEWS corpus also shows the a significant co-occurrence of the *hope of having a baby/child/children* with negative terms such as *low,*

last, and *gave up*, thus a positive outcome which is unlikely carries the same air of negativity as the high likelihood of an unwelcome outcome. Although, the CLINIC corpus contains the same negatively loaded terms such as *last*, *only* and *given up* with the desirable outcome of *having a baby* and *getting pregnant*, these negative terms are mitigated through the suggestion that this hope can be realised through the reproductive technologies offered by the clinics – which can *give people real hope.*

6 Conclusion

All the node words chosen fall within a category of uncertainty about a future outcome with an additional prosody of desirable or undesirable working on a cline from risk through hope to chance, deviating from general usage in which hope would more often be expected to carry a positive prosody.

The node words do not frequently co-occur with *infertility* as an explicitly undesirable outcome but rather the unwelcome outcome of pregnancy failing to happen i.e. miscarriage, or the lowering of the likelihood that the welcome outcome of pregnancy and having a baby will be achieved.

This welcome outcome of pregnancy is consistent across corpora, however the possibility of this outcome is approached differently in the different sets of texts, for example, although bloggers are less likely to mention *risk*, the possibility of a welcome outcome through *chance* and *hope* is also minimised potentially relating to the management of expectations.

To carry out a nuanced analysis of these lexical choices it is not enough to look at more/less frequent items in comparable corpora but one must go deeper and look at detailed concordances to uncover the semantic prosody of these choices.

References

Fillmore, C. J. & Atkins, B. T. S. 1992. "Toward a frame-based lexicon: The semantics of RISK and its neighbors". In A. Lehrer & E. Feder Kittay (Eds.), Frames, Fields, and Contrasts: *New Essays in Semantic and Lexical Organization.* Hillsdale, New Jersey: Lawrence Erlbaum, 75–102.

Greil, A, Blevins-Slauson, K and McQuillan, J. 2010. The experience of infertility: A review of recent literature. In: *Sociology of Health & Illness* 32:1 pp. 140–162

Hamilton, C., Adolphs, S. & B. Nerlich. 2007. The meanings of 'risk': A view from corpus linguistics. *Discourse and Society*, 18(2), 163–181.

Hardy, D. & Colombini, C. 2011. A genre, collocational, and constructional analysis of RISK.. *International Journal of Corpus Linguistics*, 16(4), 462–485.

Lupton, D. (1999). *Risk.* London: Routledge

Partington, A. 2004. "Utterly content in each other's company". *International Journal of Corpus Linguistics* 9:1, 131–156

Scott, M., 2008, *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.

Scots online: Linguistic practices of a distinctive message forum

Fiona M. Douglas
University of Leeds
f.m.douglas@leeds.ac.uk

This paper presents quantitative and qualitative corpus-based analysis of a distinctive and long-standing message forum. The forum is largely written in dialect – in a variety of urban vernacular Scots to be precise. But although mostly written in dialect, this forum is not hosted by a language activist group or cultural association, nor does it self-consciously badge itself as being linguistically politicised.

It is rooted in a specific geographical location where localness is paramount, and this paper argues that the forum linguistically constructs and maintains a sense of local and national identity. Its distinctive use of dialect unites but also bounds the virtual community it serves – a community that manages to be simultaneously local, diasporic, and global in its membership.

There are surprisingly few studies of CMC dialect data, notable exceptions being Kelle (2002); Androutsopoulos and Zeigler (2004); Siebenhaar (2006); Vandekerckhove and Nobels (2010), and fewer still that exploit corpus methodologies to do so. This paper uses corpus methods to investigate the nature of the dialect language used, forum posting behaviour and interactions, and the virtual community within which the forum functions.

The corpus is compiled from postings made between Jan 2007 and Aug 2011, and so comprises some four years and eight months worth of data – over 800,000 running words (pre-processing). Although in usual corpus-terms this seems small, it must be remembered that individual 'texts' in such corpora tend to be short. The corpus collected compares favourably with those of Fägersten (2006) (another message forum study with a corpus of 102,343 words) and Montero et al. (2007) (83,061 words from discussion forums), and is comparable with the corpus sizes for individual regional channels discussed in Siebenhaar (2006) as a basis for dialect study. It is also worth noting that, as Claridge (2007) and Beiβwenger and Storrer (2008) point out, internet message forum data require extensive pre-processing before anything approaching a searchable corpus can be achieved and, this can be time-consuming and tricky.

The message forum was established in 2000, is still going strong, and fulfils many of Herring's (2004) criteria for a virtual community. Although usually described as an example of 'new media', nowadays message forums may be seen, especially by a younger demographic, as fairly 'old technology', and to some extent, they have been superseded by the popularity and enhanced functionality of social networking sites. But the fact that this message forum survives, shows that it is seen as a valuable commodity and communication channel by its members. There are over 140 contributors, but there is a core membership of about a dozen posters who are very frequent and loyal contributors. The paper will offer analysis of overall trends in the forum alongside more detailed examination of the linguistic behaviour and posting relationships maintained by this central cohort.

The language used by forum posters is an interesting example of dialect in the online environment, a context which perhaps allows more freedom of linguistic expression. The paper investigates the extent to which the forum has established its own linguistic norms, and whether these are congruent with its localised virtual location and across speakers. It also assesses whether frequent contributors have consistent idiolectal preferences. Although there are some spelling conventions for Scots, is there any evidence that posters are adhering to these, and if not, are orthographic forms perhaps related to pronunciation – a sort of eye-dialect (c.f. Herring 2011)?

Comparisons are drawn with the varieties of Scots used in other online situations, but where the framing context is the more conventional Scots cultural or language group, and where arguably the writers are using Scots for more obviously ideologically motivated reasons. Keywords analysis (using ukWaC as a comparable reference corpus) yields interesting information about the most frequently occurring Scots lexical items and the density of markedly Scots forms.

Forum posts are identifiable by member, date and topic, so it is possible to carry out detailed analyses of: individual ethnographies and/or online identities; member solidarity, rules of engagement, and conflict resolution; thematic trends in topic posts; individuals who are dominant posters, and individual preferences for topic initiator and/or replier roles. The paper offers in-depth profiles of individual members and how they 'behave' in the forum, and of the make-up of the virtual community as a whole.

Finally, it considers the extent to which the forum and individual members might be considered to be 'language-aware', and deliberately or sub-consciously exploiting Scots linguistic features as an act of identity. Has the message forum developed its own 'linguistic sub-culture' (Sebba 2007) in which membership of the (virtual) discourse community is dependent on linguistic performance?

References

Androutsopoulos, J. and Zeigler, E. 2004. "Exploring language variation on the Internet: Regional speech in a chat community." In B. Gunnarsson, L.Bergström, G. Eklund, S. Fridell, L.H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren and M. Thelander (eds.) *Language Variation in Europe.* Uppsala, Sweden: Uppsala University Press.

Androutsopoulos, J. 2006. "Introduction: Sociolinguistics and computer-mediated communication". *Journal of Sociolinguistics* 10 (4): 419-438.

Beiβwenger, M. and Storrer, A. 2008. "Corpora of Computer-Mediated Communication." In A. Lüdeling and M. Kyto (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

Claridge, C. 2007. "Constructing a corpus from the web: message boards." In M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web.* Amsterdam/New York: Rodopi.

Herring, S.C. 2004. "Computer-mediated discourse analysis: An approach to researching online communities." In S.A. Barab, R. Kling and J.H. Gray (eds.) *Designing for Virtual Communities in the Service of Learning.* Cambridge: Cambridge University Press.

Herring, S.C. 2011. "Grammar and electronic communication." In C. Chappelle (ed.) *Encyclopedia of applied linguistics.* Oxford: Wiley-Blackwell.

Kelle, B. 2002. "Virtual dialect areas in the internet: 'Townchats'". In W. Gaul and G. Ritter (eds.) *Classification, Automation, and New Media*. Berlin: Springer-Verlag.

Montero, B., Watts, F. and García-Carbonell, A. 2007. "Discussion forum interactions: Text and context". *System* 35: 566-582.

Sebba, M. 2007. *Spelling and society: The culture and politics of orthography around the world.* Cambridge: Cambridge University Press.

Siebenhaar, B. 2006. "Code choice and code-switching in Swiss-German Internet Relay Chat Rooms". *Journal of Sociolinguistics* 10 (4): 481-506.

Vandekerckhove, R. and Nobels, J. 2010. "Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers". *Journal of Sociolinguistics* 14 (5): 657-677.

Linking adverbials in the academic writing of Chinese learners: a corpus-based comparison

Du Peng
Heriot-Watt University
mistadu@yahoo.com.cn

With an increasing number of international students coming to western universities, how to help them adapt to academic discourse is an issue which needs more attention. Much research has recently been conducted to identify the different uses of linking devices between native speaker (NS) and non-native speaker (NNS) written English. For example, Milton and Tsang (1993) identified an overuse as well as a misuse of logical connectors in Hong Kong students' essays in comparisonto the Brown and the LOB corpora, suggesting that the pedagogical impact on the learners would be the origin of the problems.Grange and Tyson (1996), who compared French learners' writing with a NS corpus by using both quantitative and qualitative analyses, had a similar observation.

Unlike previous research, Bolton et al. (2002) made a comparison across three corpora consisting of writing from Hong Kong learners, NS students and published academic journals. They found that both NS and NNS students overused connectors compared with professional writers. They also argue that the study of connectors should take sentence rather than word as the basic unit of analysis. Chen (2006) further developed Bolton et al.'s (2002) methodology by adopting both a word-based and a sentence-based approach, revealing that the Taiwanese students slightly overused conjunctive adverbials on the word level. However, the results were against the overuse hypothesis from the sentence-based perspective.

This dissertation reports on the use of linking adverbials (LAs) in Chinese learners' academic writing based on a corpus-based comparison. The learner corpus contains 30 assignments from 15 Chinese MSc TESOL students in the United Kingdom. The reference corpus, the British Academic Written English, consists of 2,761 assignments collected from 1,039 native university students. The hypothesis is that Chinese learners tend to overuse LAs in their writing compared with their British counterparts. The research question of this study is: to what extent and in what ways do Chinese learners of English overuse LAs in their academic written English.

In order to answer the research question appropriately, a corpus-based approach with both quantitative and qualitative analysis is adopted.

From the quantitative analysis, Chinese writers are found to overuse LAs both on the word level and the sentence level (see Table 1 and Table 2).

	NNS Corpus	NS Corpus
Word count	126,952	6,506,995
Raw frequency of LAs	2,016	85,910
LAs/10,000 words	159	132

Table 1. Overall frequencies of LA usage per 10,000 words

	NNS Corpus	NS Corpus
Sentence count	5,661	269,413
Raw frequency of LAs	2,016	85,910
LAs/1,000 sentences	356	319

Table 2. Overall frequencies of LA usage per 1,000 sentences

The log-likelihood test is adopted to assess the significance of the difference between frequency scores. The higher the value is, the more significant the difference is. It is important to mention that the log-likelihood value is always positive, however, the plus or minus symbol before the log-likelihood value indicates, respectively, the overuse or underuse in the NNS corpus compared with the NS corpus. As Table 3 shows, the learners show a striking preference forusing enumerative and additive LAs with a log-likelihood value of +90.4 (see Table 3).

Types of LAs	RF[1] (NNS)	RF (NS)	LL[2]
Enumeration & addition	670	23,114	+90.4*[3]
Summation	23	740	+4.2
Apposition	154	6,575	+4.74
Result/inference	535	26,876	+0.21
Contrast/concession	579	25,538	+12.18
Transition	55	3,067	-0.39

Table 3. The log-likelihood values of LAs by category

On the qualitative dimension, the additive LA *in addition* is found to be used redundantly. Furthermore, *besides*, which is used as an additive connector by the Chinese student writers, is the most overused individual LA (LL: +179). This usage, however, should not be encouraged, since *besides* is supposed to be confined to oral discourse. This study also reveals that the transitive LA *meanwhile* is misused by some Chinese learners. Finally, *now* (LL: -18.67) and *for example* (LL: -17.95) are identified as the two most underused LAs in Chinese students' writing.

Based on the discussion of research findings, it is argued that Chinese learners' overuse of LAs may be due to superficial understanding of cohesion and coherence, word count pressure, first language (L1) transfer and cultural background impact.

As for the pedagogical implications, this study indicates that it is essential for teachers to improve their understanding of cohesion and coherence, which is one of the preconditions for students to restrict their overuse of LAs. To develop students' ability to outline essay structures and organise ideas as part of a logical writing process would contribute more to coherence compared with the simple adoption of LAs.

In addition to enhancing students' organisational abilities, teachers may also need to work on enabling students to identify register differences. Students would need to be informed that some LAs, *besides*, for example,are not appropriate for academic writing. Another crucial matter students need to be aware of is that redundant and misused LAs, such as *in addition* and *meanwhile* identified in this study, would reduce the quality of their writing and even make it incoherent.

Furthermore, the corpus-based approach used in the current study could be applied in pedagogy to help students achieve better coherence in their writing. With regard to the use of LAs, students can discover the patterns of LAs in authentic text when examining the concordance lines from a collection of NS writing. It could be more convincing to learners when they are shown how accomplished writers use LAs, and how their own LA usage differs from standard academic writing (Milton and Tsang 1993).

As for future research, more attention could be paid to how word count pressure and L1 transfer impact on learners' LA usage, which may not adopt corpus-based comparison as the sole approach for the research. Another future direction would be the study of different argumentative styles, which may lead to different emphases on LA usage. It is also worth exploring what a role cultural background could play in these differences. Despite focusing on the appropriate use of LAs, more pedagogical research can be conducted to investigate how to help learners effectively enhance the ability to achieve better cohesion and coherence in their writing. It is hoped that this dissertation does not only present a picture of the Chinese TESOL students' LA usage, but also provides inspirations for relevant research in the future.

[1]RF: raw frequency.
[2]LL: log-likelihood value.
[3]*: a log-likelihood value of 15.13 or higher is significant at the level of $p < 0.0001$ (Rayson et al. 2004).

Acknowledgement

The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

References

Bolton, K., Nelson, G., and Hung, J. 2002. "A Corpus-Based Study of Connectors in Student Writing: Research from the International Corpus of English in Hong Kong (ICE-HK)". *International Journal of Corpus Linguistics*7(2): 165-182.

Chen, W. Y. 2006. "The Use of Conjunctive Adverbials in the Academic Papers of Advanced Taiwanese EFL Learners".*International Journal of Corpus Linguistics* 11(1): 113-130.

Granger, S., and Tyson, S. 1996."Connector usage in the English essay writing of native and non-native EFL speakers of English". *World Englishes* 15(1): 17-27.

Milton, J., and Tsang, E. S. 1993."A corpus-based study of logical connectors in EFL students' writing: Directions for future research". In R. Pemberton and E. S. Tsang (eds.)*Studies in lexis* (pp. 215-246). Hong Kong: HKUST Language Centre.

Rayson, P., Berridge, D., and Francis, B.2004."Extending the Cochran rule for the comparison of word frequencies between corpora". In G, Purnelle., C, Fairon., and A, Dister (eds.) *Le poids des mots* (pp. 926-936). Belgium: Universitaires de Louvain Press.

Public apologies and press evaluations: a CADS approach

Alison Duguid
University of Siena
duguid@unisi.it

'Journalists love the performatives of politics where something happens through someone saying it.' Bell (1991:207) A particularly resonant performative is that of the apology which can be defined as an own-face-threatening act involving an explicit expression or acknowledgement of responsibility and regret.

Speech events can of course be reported in a variety of ways. Over recent years the public apology has had a certain amount of prominence and has been seen as a strategic move in self-representation which gathers comment and discussion in the media. The political apology is often part of conflict and controversy and the Centre for Conflict Resolution claims that a sincere apology is a powerful tool to bring peace, stop arguments and restore broken relationships, so it is understandable that it is part of a repertoire of political choices. However they also warn that 'bad apologies on the other hand can strain relationships and cause bitterness to remain', as has been seen with the recent case of bankers' choice of apology strategies (Hargie *et al* 2010).

High profile apologies receive a lot of coverage in both old and new media, and reactions and evaluations of the perceived quality of the apology are broadcast widely. For example the two recent apologies by UK politicians, that of David Cameron after the Bloody Sunday report and that made by Nick Clegg on tuition fee rises, met with very different reactions and evaluations in both new and mainstream media.

It is possible to apologise using a range of strategies and linguistic forms. Many earlier studies are based on the analysis of forms elicited as a response to simulated situations and do not use naturally occurring data; most deal with an analysis of speaker intuitions about relatively informal private apology situations where issues of politeness are at stake (Blum-Kulka and Olshtain 1984; Meier 1998; Lakoff 2001; Kampf 2009). Previous literature on political apologies, includes Harris et al (2006) who analysed the political apology as a speech event in pragmatic terms and identified the salient characteristics of different types of political apology. In particular they underlined how one of its characteristics is the highly mediated nature of the event, thus differentiating the political from the informal and interpersonal apology in that we only have access to public apologies via the refracting

and framing representations by the press and other media. The Harris et al. (2006) study used a discourse analysis approach selecting data from a few high profile political apologies, considering the reactions to them as well as the forms they took. The present case study employs a corpus-assisted discourse studies approach or CADS (Partington 2004; Partington et al. forthcoming) to media evaluations of the public apology. This study is a further contribution to a number of corpus based studies dealing with aspects of pragmatics (McEnery et al. 2000; Partington 2003, 2006; Culpeper 2008; Jucker et al. 2009; Archer & Culpeper 2009; Taylor 2009, 2011 some employing the theoretical framework of (im)politeness in combination with corpus linguistics.

Public apologies are performed with a third party audience of press and public so, while the present study does analyse the varieties of form and components used by those performing apologies, the main focus of the paper is on how such public apologies are treated and evaluated in the media; in particular, lexical items and phraseologies related to *apology*, *sorry*, *regret* are examined as are the patterns used which evaluate the apologies. Evaluation here is considered as being the indication that something is good or bad (Hunston and Thompson 2001, and for corpus based studies of evaluation see Bednarek 2006; Morley and Partington 2010; Hunston 2011; Partington et al. forthcoming)and, as Labov (1976) reminds us, is often the main aim of the discourse. Admissions of culpability have to be balanced with the need to present a perceived identity as that of a competent, ethical and just individual; issues of control and lack of control (Duguid 2011; Partington et al. forthcoming) can be wielded as defence and positive face issues abound. The media evaluations bear this out both quantitatively and qualitatively. The lexis used in public apologies is often taken up by press and public and its ambiguities discussed. We find evaluations on a number of parameters: of timeliness, sincerity, spontaneity and, tellingly, what might be called the humiliation factor, the tabloids and television in particular showing a preference for this parameter with lexis such as *abject*, *grovelling*, *humiliating* as collocates of *apology*.

Another interesting issue is that of the blurred boundaries between public and private; theoretically only the apologizer can know private feelings of penitence and the expressions used in a public apology are not always a guide to how it will be perceived and judged by the public and media. The peculiar nature of Twitter means that a private sentiment is broadcast publicly and many participants are having to come to terms with this. We examine the Mc Alpine affair, in which a prominent UK politician was wrongly accused of child abuse, to see how this blurring had repercussions. It seems that, the evaluators of public apologies do indeed consider the act to be a strategy of self deprecation and deliberate self-positioning.

For the study a number of previously compiled corpora were interrogated, namely, the SiBol corpus comprising c.300,000,000 words of UK broadsheet newspaper texts, a corpus of White House briefings (c.1,500,000 words), a TV news corpus (c. 600,000 words). Also used for the analysis were an ad hoc search-word generated corpus of tabloid newspapers in their online form with *apology* as part of the search terms and a corpus of items using *Twitter apology* as a search term for press coverage of a number of examples of high profile apologies beyond the existing corpora. Wordsmith tools (Scott 1998) was used to interrogate the corpus.

This approach to the discourse of public apology allows us to examine very large amounts of data, to see if there have been changes over time, to compare data sets and to see how different discourse types typically represent the performance of public apology. Patterns and phraseologies are revealed showing how public expectations are being met or frustrated by public apologies.

References

Archer, D. and Culpeper, J. 2009. " Identifying key socio-pragmatic usage in plays and trial proceedings (1640-1760): An empirical approach via corpus annotation". *Journal of Historical Pragmatics* 10(2): 286-309

Bell, A. (1991) *The Language of News Media*. Oxford: Blackwell

Bednarek, M. 2006. *Evaluation in media Discourse.* London:Continuum.

Blum-Kulka, S. and Olshtain, E. 1984. "Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP)". *Applied Linguistics* 5: 196–213.

Culpeper, J. 2008. " Reflections on impoliteness, relational work and power". In D. Bousfield and M. Locher (eds) *Impoliteness in Language*. Berlin: Mouton de Gruyter.

Duguid, A. 2011. "Control: a semantic feature in evaluative prosody" Corpus Linguistics Conference 2011. Birmingham, UK.

Hargie, O., Stapleton, K. and Tourish, D. 2010. "Interpretations of CEO public apologies for the banking crisis: attributions of blame and avoidance of responsibility". *Organization* 17: 721-742.

Harris S., Grainger, K. and Mullany, L. (2006) 'The Pragmatics of Political Apologies', *Discourse & Society* 17: 715–37.

Hunston, S. 2011**.** *Corpus Approaches to Evaluation: Phraseology and Evaluative Language.* London/New York: Routledge.

Jucker A. ,Schreier, D, Hundt, M. (eds) 2009. C*orpora: Pragmatics and Discourse*. Amsterdam: Rodopi.

Kampf, Z. 2009. "Public (non-) Apologies: The Discourse of Minimizing Responsibility". *International Journal of Applied Linguistics* 8

Lakoff, R. 2001. "Nine Ways of Looking at Apologies: The Necessity for Interdisciplinary Theory and Method in Discourse Analysis", in D. Schriffin, D. Tannen and H. Hamilton (eds) *The Handbook of Discourse Analysis*. Oxford: Blackwell.

McEnery, T., Baker, P. & Cheepen, C. 2002. " Lexis, indirectness and politeness in operator calls". In P. Peters, P. Collins, and A. Smith (eds) *Language and Computers, New Frontiers of Corpus Research* . Amsterdam/New York: Rodopi.

Meier, A. J. 1998. "Apologies: What Do We Know?". *International Journal of Applied Linguistics* 8.

Morley, J. and Partington, A. 2009. "A few Frequently Asked Questions about semantic – or evaluative – prosody". *International Journal of Corpus Linguistics 14* (*2*): *139-158*.

Partington, A. 2003. *The Linguistics of Political Argument: The Spin-doctor and the Wolf-pack at the White House*. London: Routledge.

Partington, A. 2006. *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-talk.* London: Routledge.

Partington, A., Duguid, A., Taylor, C. (forthcoming) *Patterns and meanings in Discourse Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam and New York: John Benjamins.

Taylor, C. 2009. "Interacting with conflicting goals: Impoliteness in hostile cross-examination". In J. Morley and P. Bayley (eds.). *Corpus Assisted Discourse Studies on the Iraq Conflict: Wording the War*. London: Routledge.

Taylor, C. 2011. "Negative politeness features and impoliteness functions: A corpus-assisted approach". In Davies, B., A. Merrison and M. Haugh (eds.). *Situated Politeness*. London: Continuum.

Thompson, G. and Hunston, S. 2001. *Evaluation in Texts*. Oxford : OUP.

Using reference corpora for discourse analysis research: the case of class

Rosa Escanes Sierra
University of Sheffield
r.escanes.sierra@sheffield.ac.uk

This paper discusses the potential use of large reference corpora as a tool to critically explore the stereotypical construction of social class. The study draws from previous work within this yet to be widely developed methodological line of research (Mautner, 2007; Baker, 2006; McEnery and Xiao, 2004), which considers this type of corpus a *'rich repository of cultural information of society as a whole*' (Hunston, 1995, cited in Hunston, 2003, p. 117).

From a 'transdisciplinary' perspective (Fairclough, 2005), I briefly discuss the fact that class lies uncomfortably within inequality and identity debates. This is due to the fact that class is intrinsically linked to economic distribution (Mulderrig, 2007) and that a more subjective (Lawler, 2005) and moral (Sayer, 2005) approach to its analysis gives way to class-based stereotypes which tend to stigmatise lower classes.

To give insight into the linguistic representation of this phenomenon, the study explores the terms *working class*, *middle class* and *upper class* in the British National Corpus (BNC). Looking at the occurrences per million words of the three keywords, analysing their collocation profile and concordance lines, the study aims to answer the following questions: what patterns can be observed in the use of these three node words? What are the most statistically prominent collocates? In which contexts do they appear most commonly? What patterns regarding semantic prosody (Stubbs, 1996) can be found?

A diachronic analysis (by comparing the three periods 1960-1974, 1975-1984 and 1985-1993) of the occurrences per million words point to a decrease in class references (table 1). This implies a gradual process of sociolinguistic change in which class seems to be progressively less significant in contemporary British English. Even though interpretations of these kind of data are difficult, one could argue that this seems to match the progressive 'relativisation' of class as a source of inequality and identity determination (Pakulski and Waters, 1996).

Overall, *working class* is the most commonly used term compared to *middle class* and *upper class*. In other words, *working class* seems to be considered a 'deviant form' (Hunston, 2003, p. 66), further away from 'norm', since it is referred to much more

often[1]. *Upper class* is noticeably absent in the BNC (with only 4.8 occurrences per million words).

	Working class	Middle class	Upper class
Period 1960-1974	69.3	91.9	11.6
Period 1975-1984	100.1	51.2	20.1
Period 1985-1993	40.3	26.7	4.3
Overall (synchronic analysis)	39.8	26.8	4.8

Table 1. Occurrences per million words of *working class*, *middle class* and *upper class* in the BNC

The study also reveals some intriguing patterns regarding collocation profiles. For instance, *working class* was more frequently associated with criminality than *middle class* and *upper class*. There also seems to be a lack of statistically relevant collocates referring to economic differences. Moreover, specific gendered constructions of *middle class* and *working class* as well as class-based differences in the representation of youth were found. Finally, *upper class* was often part of an anachronistic and superficial contextualisation. We can observe some of these contrastive patterns in table 2.

Working class	Middle class	Upper class
Women, children, people, youth(s), white, men, (house)wife/ wives, boys, girls, male, black, adolescents, crime, respectable	Women, white, children, young, English, middle-aged, housewife, respectable	Accent, culture, century, English, White, British

Table 2. Statistically relevant collocates (according to MI-score and t-score) of *working class*, *middle class* and *upper class* in the BNC.

The significance of these collocation patterns are then put into perspective by analysing the concordance lines using the pertinent collocates as a filter. This qualitative examination points to a more negative semantic prosody surrounding *working class*, when compared to equivalent analyses of *middle class* and *upper class* concordances.
Moreover, when using semantically positive collocates such as *respectable*, class-based differences are again apparent: this collocate commonly appears between inverted commas when referring to *working class*, but not when referring to *middle class*.

I suggest then that, considering the limitations imposed by the necessary contextualization of the BNC, there seems to be a tendency to find working-class people entangled in 'negativity' and 'problematised' prosodies more often than middle-class or upper-class people. These discursive patterns, together with the seemingly lack of prominence in references to class-based economic differences, could contribute to the development of prejudices and stereotypes in inequality debates. In other words, I claim that there is a need to find an *'alternative discourse to create alternative ways of thinking about poverty and inequality'* (Marston, 2008, p. 360).

To conclude, I claim that the representative nature of reference corpora and their potential to uncover the most common uses of certain socially relevant keywords (Baker, 2006) is highlighted in this study, which contributes to the development of this line of investigation within interdisciplinary fields such as sociolinguistics and critical discourse analysis. Finally, I also point to the fact that the relevance of this study could be enhanced by comparing its results to more contemporary data (LexisNexis, BE06) and by considering alternative lexical items (*elite*, *chav*) that might be used to refer to class indirectly.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis.* London: Continuum.

Fairclough, N. (2005). Critical Discourse Analysis. *Marges Linguistiques , 5*, 76-94.

Hunston, S. (2003). *Corpora in Applied Linguistics.* Cambridge: CUP.

Hunston, S. (1995). Grammar in Teacher education: the Role of a Corpus. *Language Awareness* , 4: 15-32.

Lawler, S. (2005). Introduction: Class, Culture and Identity. *Sociology , 39*, 797-806.

Marston, G. (2008). A War on the Poor: Constructing Welfare and Work in the Twenty-first Century. *Critical Discourse Studies* , 5: 359-370.

Mautner, G. (2007). Mining Large Corpora for Social Information: the Case of Elderly. *Language in Society* , 36: 51-72.

McEnery, A., & Xiao, Z. (2004). Swearing in Modern British English: The Case of Fuck in the BNC. *Language and Literature , 13* (3), 235-268.

Mulderrig, J. (2007). *Equality and Human Rights: Key concepts and Issues.* Edinburgh: University of Edinburgh.

[1] In the same way that we would find more instances of left-handed than right-handed (Hunston, 2003, p. 66).

Pakulski, J. and Waters, M. (1996). *The Death of Class*. London: SAGE.

Sayer, A. (2005). Class, Moral Worth and Recognition. *Sociology , 39*, 947-963.

Stubbs, M. (1996). *Text and Corpus Linguistics.* Oxford: Blackwell.

Statistical modelling of natural language for descriptive linguistics

Stefan Evert
University of Erlangen
stefan.evert@fau.de

Gerold Schneider
University of Zurich
gschneid@es.uzh.ch

Hans Martin Lehmann
University of Zurich
Hmlehman@es.uzh.ch

1 Introduction

In this paper we argue against classic single feature analyses of relative and/or proportional frequencies in corpora. We show that at least the factors determining the sampling frame need to be included in the analysis. We discuss the consequences of omitting such external factors and the advantage of including them on the basis of several case studies using data from the Brown, LOB and ICE corpora. In a second step we explore strategies for dealing with missing external factors. Specifically we explore unsupervised methods like distributional semantics for inferring such factors.

2 Motivation

Traditionally, most quantitative research in corpus linguistics builds on a comparison of corpus frequencies using statistical hypothesis tests (e.g. Oakes 1998; McEnery and Wilson 2001). This approach has been criticised in recent years for several reasons:

1) Natural language data do not satisfy the randomness assumption underlying the kind of hypothesis tests commonly used for frequency comparisons (in particular, Pearson's chi-squared test and other methods for analysing contingency tables), which may lead to severely inflated significance values (Church 2000, Kilgarriff 2005, Evert 2006). Note that this is not an issue of asymptotic normality assumptions made by the chi-squared test: other methods such as likelihood-ratio tests (Dunning, 1993) or Fisher's exact test (recommended e.g. by Stefanowitsch and Gries 2003) are equally affected.

2) Hypothesis tests usually focus on p-values indicating the amount of evidence supporting a result, rather than effect size (which is linguistically

more relevant). As a result, trivial frequency differences may appear to be highly significant if the analysis is based on very large corpora (Gries, 2005).

3) The common practice of pooling data from all texts in a corpus (i.e. comparing overall corpus frequencies rather than frequencies in individual texts) ignores linguistic variation, i.e. the variability of linguistic phenomena within a language (Gries 2006). As a form of non-randomness, this may lead to spuriously inflated significance; on the other hand, language-internal variation might mask differences between language varieties (or between different phenomena in the same language if their variational patterns differ).

4) Significance testing is typically applied to one factor only. Such an approach is only sufficient if the factor under consideration has a dominant influence on the relevant frequency. If it is strongly correlated with another factor (and perhaps causally dependent on it), or has strong interactions with factors that are not considered, the significance tests become unreliable and there is a considerable risk of type-I errors (reporting insignificant differences as significant) for variationist or sociolinguistic differences. For certain types of tests, there is also a high risk of type-II errors (failure to detect true differences).

Gries (2006) argues that a methodologically sound approach to corpus linguistics needs to take linguistic variation into account explicitly, i.e. it must endeavour to explain the variation of relative frequencies across individual texts. In this paper, we propose that statistical regression models – in particular, generalised linear models (GLM, Dobson 1990) – are very well suited for this purpose and address all four criticisms raised above.

3 Case studies

In several case studies, we illustrate the unreliability of traditional significance tests and show how such problems can be addressed with the help of regression models. In a study dealing with passive voice variation, Evert (2006) shows that the assumption of randomness is not met by corpus data.

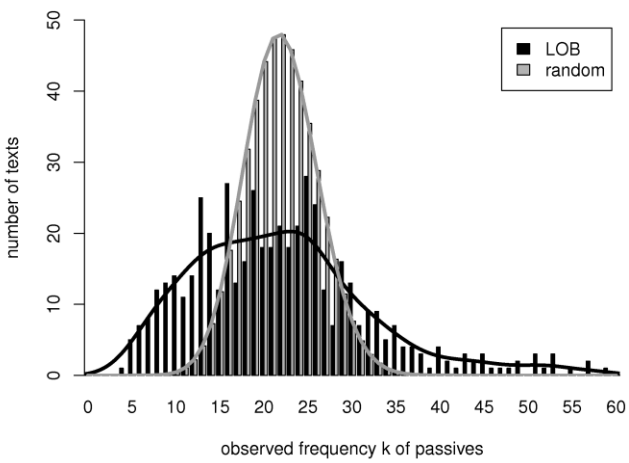


Figure 1. Distribution of the frequency of passive forms across texts in the Brown corpus. The grey bars and contour show the distribution expected if the randomness assumption were satisfied.

If passive forms were indeed spread randomly throughout the LOB corpus, the distribution of passive counts across the individual texts in the corpus would follow a binomial distribution, as indicated by the grey bars and contour line in Figure 1. This is the amount of variability assumed by a chi-squared test applied to the pooled frequency data from the entire corpus. However, the true variability of per-text frequencies is much larger, shown by the black bars and contour line in Figure 1. Accordingly, the chi-squared test will overestimate the significance of observed differences, which might in fact be explained by the large amount of variability of the true distribution.

One may be tempted to apply Student's t-test to per-text relative frequencies, instead. However, the t-test assumes a Gaussian distribution, which is clearly not met by the corpus data in Figure 1. In our case study, this results in a type-II error, making the t-test excessively conservative. The large amount of variability between texts hides actual differences between language varieties when comparing the LOB and Brown corpora.

The problem that content words are not randomly spread throughout a corpus, but tend to cluster in certain documents is well known as the *London bus problem* or the *Noriega* problem (Church 2000). But it partly also applies to grammatical forms such as passives and functions words. In our case study on LOB versus Brown passives, it turns out that register, style and the individual document have much greater influence than regional variation. A carefully designed regression model which includes semantic and stylistic factors shows that regional variation is indeed significant.

Similarly, in our case studies of ICE past perfect and modals, the genre and the individual document are more significant factors than the variety.

Concerning past perfect, Sedlatschek (2009) claims that past perfect form is significantly more frequent in Indian English than British English, using chi-square. He uses his own corpus and attains significance just below the 5% level, so he also advises to treat the result with caution. Our ICE data does not support his result. Figure 2 shows that the randomness assumption is not met in this case, either. We explore regression models including additional factors.

In the case of modal verbs, Nelson (2003) claims that the modal verb *should* is significantly more frequent in East African English than in British English, based on a chi-squared test. Nelson has used ICE data, too, and his findings agree with our observations.

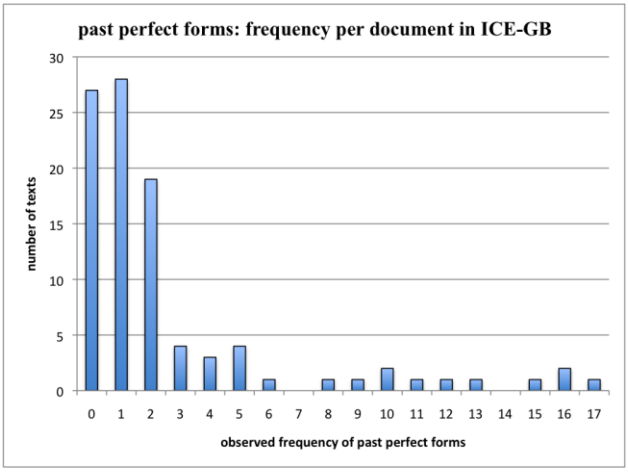


Figure 2. Distribution of the past perfect form in a large subset of ICE-GB written

But the randomness assumption is not met, as Figure 3 shows. The influence of other factors: genre, text, semantics, is strong. In a sense, only individual texts can be treated as independent tokens. In order to account for this fact, we use regression models that include more fine-grained document-type information.

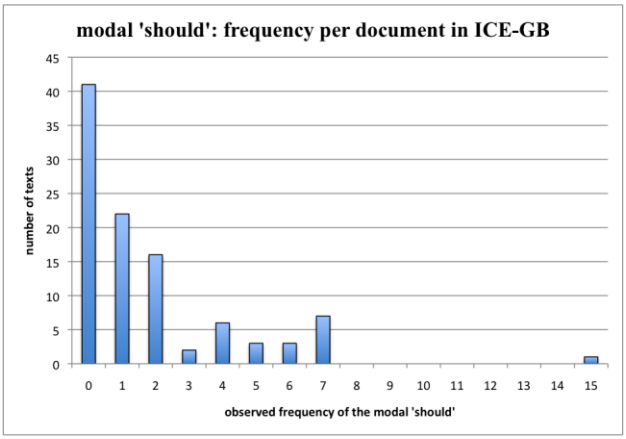


Figure 3. Distribution of the modal verb *should* in a large subset of ICE-GB written

To further illustrate these points, we use the Saxon genitive as a test case on a wide range of corpora and with data sets ranging from sparse to abundant, using relative and proportional frequencies.

4 Conclusion

When applied appropriately, GLM can give a satisfactory account of language-internal variability and separate the factor under investigation from all the other factors underlying language variation. Previous attempts to use GLM and similar regression models in corpus linguistics have been hampered by the fact that most of the relevant factors are not accessible to the model: LOB and the ICE corpora have meta-information on genre (and modality in ICE), but do not include other factors (speaker/author, style, topic, etc.) that would account for differences between texts from the same genre. We use unsupervised methods to infer such factors from language-internal correlation patterns, namely multidimensional register analysis in the spirit of Biber (1988) and distributional semantic models (see e.g. Turney and Pantel, 2010).

We believe that statistical regression models should be used consistently in Corpus Linguistics for the analysis of frequency data: they help to avoid distortion of significance values due to non-randomness, they can account for variability of frequency data, they produce estimates of effect size in addition to p-values, they can take multiple factors into consideration and allow us to study their interactions, and they give predictive models which can be tested against novel data. Thus, this approach opens up entirely new perspectives for Corpus Linguistics research.

References

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Church, K. 2000. "Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p^2." In *Proceedings of the 17th conference on Computational linguistics*, pages 180–186.

Dobson, A. J. 1990. *An Introduction to Generalized Linear Models.* Chapman and Hall, London.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 19 (1), pages 61–74.

Evert, S. 2006. "How random is a corpus? the library metaphor". *Zeitschrift für Anglistik und Amerikanistik* 54 (2).

Gries, S. T. 2005. "Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff". *Corpus Linguistics and Linguistic Theory* 1 (2), pages 277–294.

Gries, S. T. 2006. "Exploring variability within and

between corpora: some methodological considerations". *Corpora* 1 (2), pages 109–151.

Kilgarriff, A. 2005. "Language is never ever ever random". *Corpus Linguistics and Linguistic Theory* 1 (2), pages 263–276.

McEnery, T. and Wilson, A. 2001. *Corpus Linguistics: An Introduction, 2nd edition*. Edinburgh University Press, Edinburgh.

Nelson, G. 2003. "Modals of obligation and necessity in varieties of English". In Peters, P., editor, *From Local to Global English*, pages 25–32. Dictionary Research Centre, Macquarie University, Sydney.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. EUP, Edinburgh.

Sedlatschek, A. 2009. *Contemporary Indian English: variation and change.* Varieties of English around the world. John Benjamins, Amsterdam / Philadelphia.

Stefanowitsch, A. and Gries, S. T. 2003. "Collostructions: investigating the interaction between words and constructions". *International Journal of Corpus Linguistics* 8 (2), pages 209–43.

Turney, P. D. and Pantel, P. 2010. "From frequency to meaning. Vector space models for semantics". *Journal of Artificial Intelligence Research* 37, pages 141–188.

Literature and statistics – a corpus-based study of endings in short stories

Jennifer Fest
RWTH Aachen University
Jennifer.Fest@rwth-aachen.de

Stella Neumann
RWTH Aachen University
Neumann@anglistik.rwth-aachen.de

1 Research Background

Although endings constitute an important function in texts, more detailed analyses of them as structural elements are rare both in linguistics and in literary studies. Within the framework of the former, and especially in the field of discourse studies, some suggestions have been made as to what might indicate endings in different types of texts (Werlich 1976; Schegloff and Sacks 1999), yet no empirical evidence has been collected so far to test the validity of these models. In literary studies, the situation is similar; while some categorisations of possible markers for text endings have been suggested, only Krings (2003) followed a corpus-based approach in order to investigate the empirical suitability in a corpus of literary texts.

It is especially in literary studies that scholars agree on the fact that an ending constitutes not only an important, but elementary part in a text. Its functions can be manifold; it can serve to sum up the story that has just been told (Rabinovitz 2002), it can help the author to make or emphasise a point (Torgovnick 1981) and can even be a pivotal, if not turning point in a narrative (Eichenbaum 1971; Sternberg 1976). Furthermore, it is what the reader will most likely keep in mind (Torgovnick 1981); when reading a story, we expect the plot to come to a conclusion and very often, we intuitively notice when the end of a story is drawing near.

The study at hand is a collaborative project including researchers from linguistics and literary studies in order to make use of different perspectives and methods (Toolan 2008). Taking the theoretical backgrounds as well as methodological approaches from both fields into account, it aims at getting to the bottom of this intuition. It tries to understand which features in a text trigger the sense of an approaching ending in a reader and give the ending its impact and importance. The features and categories which have been suggested in the literature so far serve as a rough basis for the analysis; in general however, the texts used for the work are examined statistically and without previous limitations as to what could be expected and searched for.

2 Data and method

The corpus that has so far been compiled for the analysis consists of 130 short stories by thirteen different authors. It covers both American and British authors and all in all contains 592,224 words. The first texts have been collected via Project Gutenberg, which allows free access to stories for which the copyright has expired, and at a later stage more texts were digitalised and added to the corpus. All files are enriched with basic HTML mark-up and were tagged for parts of speech and sentence IDs to allow more detailed queries using CQP workbench. For very specific queries, scripts based on python code were used.

With these tools at hand, queries were constructed to cover those features which in literary studies are considered markers of an ending (Krings 2003; Wenzel 2011). To achieve this, different approaches were combined; while for instance the marker of sentences starting with "And" could be analysed rather easily, the creation of closural allusion by a density of words from the same semantic field required not only a much more sophisticated query, but also a definition of this field and a categorisation of words. Finally, different methods had to be applied for analyses concerning features of the text structure, such as sentence length or number of sentences and paragraphs.

3 Research questions / findings

Complicated and challenging though it was to constitute useful and concise methods, they offered the chance to investigate many different research questions and issues. On a micro-level, the frequencies of individual signals of endings were identified and put into comparison; in this context, it for instance showed that most authors make use of very short last sentences, while a repetition or resumption of the title, which is considered a framing mechanism, is less frequent. On a macro-level, these results can be compared to identify particularities of certain literary epochs or traditions. Moreover, texts of individual authors can be contrasted and thus mirror a change in writing habits or the development a writer might have undergone in the course of their life. Also, it becomes apparent that although most of the features suggested by literary scholars can indeed be traced statistically, most of the analysed authors do not make use of all of them but restrict themselves to certain mechanisms or a combination of few.

Since the first analyses rendered very promising results, many more research questions have sprung up in the course of this work. One question which is still under debate and will need many more analyses to answer is the issue of where an ending really starts. Furthermore, while on the one hand an enlargement of the corpus and an advancement of the scripts and mark-ups are currently in progress, the concept of analysing endings has on the other hand been transferred to a broader range of topics, including other types of texts as well as films, digital media and psychological frameworks.

References

Eichenbaum, B. 1971. "O. Henry and the Theory of the Short Story". In L. Matejka and K. Pomorska (eds.) *Reading in Russian Poetics. Formalist and Structuralist Views*. Cambridge: MIT Press.

Krings, C. 2003. *Zur Typologie des Erzählschlusses in der englischsprachigen Kurzgeschichte*. Frankfurt am Main: Lang.

Rabinovitz, P. 2002. "Reading Beginnings and Endings". In B. Richardson (ed.) *Narrative Dynamics: Essays on Time, Plot, Closure, and Frames*. Columbus: Ohio State University Press.

Schegloff, E. and Sacks, H. 1999. "Opening up Closings". In A. Jaworski and N. Coupland (eds.) *The Discourse Reader*. London: Routledge.

Sternberg, M. 1976. "Temporal Ordering, Modes of Expositional Distribution, and Three Models of Rhetorical Control in the Narrative Text". *PTL* 1: 295-316.

Toolan, M. 2008. "Narrative Progression in the Short Story: First Steps in a Corpus Stylistic Approach". *Narrative* 16 (2): 105-120.

Torgovnick, M. 1981. *Closure in the Novel*. Princeton: University Press.

Wenzel, P. 2011. "Schlüsse über Schlüsse: Zur Struktur der Schlussgebung in Literatur und Kultur". In K. Herrmann (ed.) *Neuroästhetik: Perspektiven auf ein interdisziplinäres Forschungsgebiet*. Kassel: Kassel University Press.

Werlich, E. 1976. *A Text Grammar of English*. Heidelberg: Quelle & Meyer.

Corpus Linguistics and English for Specific Purposes: Which unit for linguistic analysis?

Lynne Flowerdew
Hong Kong University of Science and Technology
lclynne@ust.hk

There is now a substantial body of research on corpus linguistics and English for Specific Purposes (ESP) and edited volumes have recently appeared or are in preparation (e.g. Boulton et al. (eds.) 2012; Gotti & Giannone (eds.), forthcoming). Some ESP corpora such as the 2.6-million word MICUSP (Michigan Corpus of Upper-student Papers) contain a variety of different text types, while other ESP corpora are quite specialised, e.g. the 500,000-word Guangzhou Petroleum English Corpus (Zhu 1989), one of the first specialised corpora to be compiled (see Flowerdew 2004, 2011, 2012 and Warren 2010b, for more details on ESP corpora). ESP corpora can thus loosely be categorised into those for English for General Academic Purpses (EGAP) as in the case of MICUSP, or English for Specific Academic Purposes (ESAP) if they concentrate on a particular discipline.

The vast majority of research on ESP corpora commences from a bottom-up perspective in which lexis or some kind of lexico-grammatical unit is taken as the starting point for analysis, which is then often categorized according to rhetorical function; in contrast, in the top-down approach the functional components of a genre are determined first and then all the texts in a corpus are analysed in terms of these components (see Biber et al. 2007 for more information on top-down and bottom-up approaches). The following bottom-up linguistic units as an entry point to the corpus analysis have been identified in the literature on corpus-based ESP. These are, to a large extent, also driven by the software used.

- Frequency lists
- Key words and key-key words
- Lexical bundles (n-grams)
- Phrase frames
- Concgrams
- Collocational frameworks
- 'small words'
- Semantic sequences

Frequency lists can be generated by a variety of software, with many ESP corpus studies making use of the freely-available *AntConc* software (see Anthony 2007). Arguably, the most well-known corpus-derived frequency list of core academic vocabulary is Coxhead's (2000, 2011) Academic Word List (AWL), although Hyland and Tse (2007) dispute the concept of such core vocabulary, arguing for specificity in disciplinary discourses. Key words are also a type of frequency list, reflecting words of unusually high frequency in a specialist corpus when compared with a larger-scale reference corpus (see Scott and Tribble 2006). Key words form the backbone of a number of contrastive ESP corpus-studies as they reflect disciplinary epistemologies (see Malavasi and Mazzi 2010). Moreover, it should be noted that most studies examine key words from a phraseological perspective in line with Sinclair's 'extended units of meaning'.

One type of phraseology is lexical bundles, which are also a well-researched feature in different types of ESP corpora (Biber 2006; Hyland 2008); these, like key words, are a useful indicator of disciplinary variation. Phrase frames (Fletcher 2007) are similar to lexical bundles but allow for an internal variable slot in the n-gram. Phrase frames have been the subject of investigation in MICUSP by Ädel and Römer (2012), who found the frame *with the * of* to be common in academic prose having top variants of *idea, use, help*. Concgrams (Greaves 2009) can be seen as developing from phrase frames, but are more flexible as they allow for constituency and positional variation. Cheng (2009) has investigated congrams in financial text and Warren (2010a) their behavior in engineering texts. Warren presents concordance output to illustrate the two-word congram *design/structural*, although it is still up to the researcher to examine each concordance line to see whether the concrams are meaningfully associated, e.g. *…framing plan (d) represents the final **design** of **structural** framing…*, or just simply co-occur, e.g. *…reduced occupied **structural** space, and shorter **design** time, have been realized* (p. 116). Collocational frameworks (see Renouf and Sinclair 1991), which have some similarity to phrase frames, but unlike phrase frames constitute meaningful units, have also been investigated in ESP texts. For instance, in a 300,000-word corpus of 100 medical papers Marco (2000) found the following frames to be the most frequent: *the…of; a…of; be…to*, which she then classified in notional terms. For example *the…of* was found to be the preferred pattern for nominalisations, a framework also noted in biology texts, e.g. *the cloning of, the efficacy of* (Hyland 2009). The framework *a…of* was found with processes of quantifying and categorizing, e.g. *a total of*; *a* (*large/small*) *number of* and the framework *be…to* to signal relational processes, e.g. *be similar to*. Other ESP researchers have commenced with 'small words' as a lead-in to lexico-grammatical patterns. 'Small words',

according to Gledhill's (2000, 2011) definition, are high-frequency, closed-class grammatical items such as prepositions. Using the *AntConc* software, Gledhill (2011) searched on the string ***of* was *(-ed)*** to find processes of Biochemical entities, yielding instances such as the following:

> *In our case, the optimum content **of** acetonitrile **was** found to vary between 25 and 30% depending on the column efficiency.*
>
> *The efficacy **of** zidovudine **was** shown to reduce risk of transmission by 66% in the treated group.*
>
> *The prevalence **of** restraint **was** found to be 68% (n+69).*

Likewise, Groom's (2010) starting point is also with prepositions as a probe for investigating 'semantic sequences', which are repeated sequences of meaning which may be realized through a range of different grammatical forms (Hunston 2010). For example, one common semantic sequence identified by Groom in a 3.2-million-word corpus of journal articles representing the disciplinary discourse of history was that of 'conceptualization' + *of* + 'phenomenon', as illustrated below in Figure 1:

Conceptualisation	*of*	**Phenomenon**
the institution	*of*	Mother's Day
the complementary component of	*of*	cultural stewardship
The cases	*of*	(ex socialist) Marcel Déat and (ex communist) Jacques Doriot

Figure 1: Conceptualisation + of + Phenomenon (Groom 2010: 68)

It can be seen that the above ESP studies commence from a lexical base, which tends to be the case as far as small, specialised corpora are concerned. While ESP studies commencing with these lexically-oriented bottom-up linguistic units provide valuable insights into lexis or lexical patterning, they primarily focus on syntagmatic relations. In this paper it will be argued that it would also be useful to put more attention on paradigmatic relations so that grammatical patterns can be more easily discerned, in order to provide a more comprehensive picture of language in specialised domains. Some suggestions for future analysis of ESP corpora will be provided.

References

Ädel, A. and , Römer U. 2012. "Research on advanced student writing across disciplines and levels. Introducing the *Michigan Corpus of Upper-level Student Papers*". *International Journal of Corpus Linguistics* 17 (1): 3-34.

Anthony, L. 2007. *AntConc 3.2.* Faculty of Science and Engineering. Waseda University.

Biber, D. 2006. *University Language. A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., Connor, U. and Upton, T. 2007. *Discourse on the Move*. Amsterdam: John Benjamins.

Bondi, M. and Scott, M. (eds.) 2010. *Keyness in Texts*. Amsterdam: John Benjamins.

Boulton, A., Carter-Thomas, S., and Rowley-Jolivet, E. (Eds.) 2012. *Corpus-informed Research and Learning in ESP*. Amsterdam: John Benjamins.

Cheng, W. 2009. "*Income/interest/net.* Using internal criteria to determine the aboutness of a text". In K. Aijmer (ed.) *Corpora and Language Teaching*, pp. 157-177. Amsterdam: John Benjamins.

Coxhead, A. 2000. "A new academic wordlist". *TESOL Quarterly* 34 (2): 213-238.

Coxhead, A. 2011. "The academic wordlist 10 years on: Researching and teaching implications. *TESOL Quarterly* 34 (2): 213-238.

Fletcher, W. 2002-2007. *KfNgram*. Annapolis, MD: United States Naval Academy.

Flowerdew, L. 2004. "The argument for using English specialised corpora to understand academic and professional language". In U. Connor and T. Upton (eds.) *Discourse in the Professions: Perspectives from corpus linguistics*, pp. 11-33. Amsterdam: John Benjamins.

Flowerdew, L. 2011. "ESP and corpus studies". In D. Belcher, A. Johns and B. Paltridge (eds.) *New Directions in English for Specific Purposes Research*. Ann Arbor, MI: University of Michigan Press.

Flowerdew, L. 2012. *Corpora and Language Education*. London: Palgrave Macmillan.

Gledhill, C. 2000. "The discourse function of collocation in research article introductions". *English for Specific Purposes* 19 (2): 115-135.

Gledhill, C. 2011. "The 'lexicogrammar' approach to analyzing phraseology and collocation in ESP texts". *Asp* 59, 5-23.

Gotti, M. & Giannoni, D. (eds.) (forthcoming) *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse.* Bern: Peter Lang.

Greaves, C. 2009. *ConcGram 1.0: A phraseological search engine*. Amsterdam: John Benjamins.

Groom, N. 2010. "Closed-class keywords and corpus-driven discourse analysis". In M. Bondi and M. Scott (eds.) *Keyness in Texts*, pp. 59-78. Amsterdam: John Benjamins.

Hunston, S. 2010. "Starting with the small words:

Patterns, lexis and semantic sequences". In U Römer and R. Schulze (eds.) *Patterns, Meaningful Units and Specialised Discourses,* pp. 7-30. Amsterdam: John Benjamins.

Hyland, K. 2008. "As can be seen: lexical bundles and disciplinary variation". *English for Specific Purposes*, 27(1): 4-21.

Hyland, K. 2009. *Academic Discourse*. London: Continuum.

Hyland, K. and Tse, P. 2007. "Is there an 'academic vocabulary'?" *TESOL Quarterly* 41 (2): 235-253.

Malavasi, D. and Mazzi, D. 2010. "History v. marketing: Keywords as a clue to disciplinary epistemology". In M. Bondi and M. Scott (eds.), pp. 169-184.

Marco, M.J.L. 2000. "Collocational frameworks in medical research papers: a genre-based study". *English for Specific Purposes* 19 (1): 63-86.

Renouf, A. and Sinclair, J.McH. 1991. "Collocational frameworks in English". In K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics*, pp. 128-143. Amsterdam: Rodopi.

Scott, M. and Tribble, C. 2006. *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Warren, M. 2010a. "Identifying aboutgrams in engineering texts". In M. Bondi and M Scott (eds.) *Keyness in Texts*, pp. 113-126. Amsterdam: John Benjamins.

Warren, M. (2010b) "Online corpora for specific purposes". *ICAME Journal* 34: 169-188.

Zhu, A. 1989. "A quantitative look at the Guangzhou Petroleum English Corpus". *ICAME Journal,* 13: 28-38.

Corpus frequency or the preference of dictionary editors and grammarians?: the negative and question forms of *used to*

Kazuko Fujimoto
Soka University
kazuko@soka.ac.jp

1 Introduction

Searching corpora enables us to find that there are some cases where descriptions in dictionaries and grammar books do not necessarily reflect the corpus-based frequency of words and structures. The negative and question forms of *used to,* which refers to a habitual action or state in the past, may be good examples to take.

Since *used to* is used as both an auxiliary and a lexical verb, opinions have long been divided among grammarians about which negative and question forms are acceptable or should be used. The variation in the forms is also observed in learners' dictionaries. This diversity of the forms in the learners' dictionaries motivated my corpus-based study of *used to*. The aim of this paper is to examine the frequency and the spoken/written distribution of the negative and question forms of *used to* in British/American English corpora. My corpus findings indicate that the dictionaries do not fully reflect the actual usage of *used to*. Generally they seem to reflect dictionary editors' and grammarians' attitudes towards it. It may be suggested that the forms prioritized in teaching should be reconsidered based on the frequency of the negative and question forms of *used to*.

2 Material

First, I will compare examples and usage notes in the latest edition of eight major advanced learners' dictionaries: *Cambridge Advanced Learner's Dictionary* (Third edition, 2008), *Collins COBUILD Advanced Dictionary of English* (Seventh edition, 2012), *Longman Dictionary of Contemporary English* (Fifth edition, 2009), *Macmillan English Dictionary for Advanced Learners* (Second edition, 2007), *Oxford Advanced Learner's Dictionary* (Eighth edition, 2010), *Collins COBUILD Advanced Dictionary of American English* (2007), *Longman Advanced American Dictionary* (Second edition, 2007), and *Merriam-Webster's Advanced Learner's English Dictionary* (2008). I will also refer to English grammar and usage books when it is necessary to compare some more information

especially for the forms about which the diversity of opinion is observed among the dictionaries. Next, the frequency and distribution of the negative and questions forms will be examined in three types of corpora: the Brown family of corpora (BFC),[1] the British National Corpus (BNC),[2] and the Corpus of Contemporary American English (COCA).[3]

The following negative and question forms presented in the learners' dictionaries will be examined: *did not/didn't use to, did not/didn't used to, used not to, never used to, use to not, usen't to,* and *usedn't to* for the negative forms, and *did . . . use to . . . ?, did . . . used to . . . ?,* and *used . . . to . . . ?* for the question forms.

3 Results and discussion

The main focus of the discussion in the learners' dictionaries and the grammar books examined is on the acceptability of the use of *used to* after *did*. Seven of the eight dictionaries have *did not/didn't use to,* and this is the first negative form presented in the five dictionaries. Only one dictionary has *did not/didn't used to*. It is notable that five of the eight dictionaries give *did . . . use to . . .?* for the first or only question form. None of the dictionaries gives *did . . . used to . . . ?*

I attempted to make a synchronic and diachronic analysis with BFC so that BrE and AmE would be compared by tracing the change of the usage of *used to* from 1931 to 2006. However, as a result of the corpus-family search, the small number of occurrences prevented me from making a meaningful comparison. In this paper the analysis will mainly be based on BNC and COCA, with which written and spoken language in BrE and AmE can be compared. The difference in frequency of each negative and question form between the written and spoken sub-corpora of BNC and COCA ([BNC_W], [BNC_S], [COCA_W], and [COCA_S] respectively) is examined by log-likelihood tests and %DIFF (the '% difference' of the normalized frequencies [per million words]).[4]

The frequency of *used to* is much higher in the spoken sub-corpora than in the written sub-corpora in BNC and COCA (and the difference is statistically significant at the level of $p<0.0001$). It is noticed that the frequency of the negative and question forms of *used to* is extremely low in both the written and spoken sub-corpora. In most cases in BNC and all the cases in COCA, the frequency per million words is less than 1.00. In BNC and COCA, all the negative and question forms except *did not use to* in COCA are more frequent in their spoken sub-corpora than in their written sub-corpora. For both the negative and question forms, after *did, used to* is more frequent than *use to* in BNC_S, COCA_W, and COCA_S (in BNC_W, they have equal frequency), though about the spoken sub-corpora it might be necessary to take into account the transcription problem of *use to* and *used to*.[5] The corpus findings also reveal that *never used to* is the most frequent of the negative forms in all the sub-corpora. However, this negative form can be found in only two of the eight learners' dictionaries.

4 Conclusion

The learners' dictionaries in general provide different results from my corpus findings. The dictionary editors seem to hesitate to approve *did . . . used to* because some grammarians regard this as less acceptable or not correct, though my findings show that it is more common than *did . . . use to*. Leech (2011: 13-14, 18, 27) emphasizes the importance of frequency information for "language learning and teaching purposes" (the principle "more frequent = more important to learn"). Since the frequency of the negative and question forms of *used to* is extremely low, the priority of teaching negative and question forms of *used to* needs to be reconsidered or different constructions or expressions may be recommended for its negative and question forms. Further observations about how the frequency of the forms has been changing will also be necessary.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English.* Harlow: Pearson Education Limited.

Carter, R. and McCarthy, M. 2006. *Cambridge grammar of English.* Cambridge: Cambridge University Press.

Gabrielatos, C. and Marchi, A. 2011. "Keyness: Matching metrics to definitions". Paper presented at *Corpus Linguistics in the South. Theoretical-methodological challenges in corpus approaches to discourse studies – and some ways of addressing them*, University of Portsmouth, 5 November 2011. Available online at http://eprints.lancs.ac.uk/51449/ (accessed 13 February 2012).

Gabrielatos, C. and Marchi, A. 2012. "Keyness: Appropriate metrics and practical issues". Paper

[1] The Brown Family (extended): *powered by CQPweb* was used. Available online at http://cqpweb.lancs.ac.uk/ (accessed 21 August 2012)

[2] The CQP-Edition of *BNCweb* was used. Available online at http://bncweb.lancs.ac.uk/ (accessed 22 August 2012)

[3] Since COCA is a monitor corpus, it is updated regularly. As of 22 August 2012, the latest update was in June 2012. Available online at http://corpus.byu.edu/coca/ (accessed 22 August 2012)

[4] Gabrielatos and Marchi (2011).

[5] See Hoffmann et al. (2008: 37-38) about transcription problems in BNC and the information on the COCA spoken transcripts available online at http://corpus.byu.edu/coca/.

presented at *CADS International Conference 2012. Corpus-assisted discourse studies: More than the sum of discourse analysis and computing?*, University of Bologna, 14 September 2012. Available online at http://repository.edgehill.ac.uk/4196/ (accessed 8 November 2012).

Garner, B. A. 2009. *Garner's modern American usage (Second edition).* New York: Oxford University Press.

Gilman, W. E. 1989. *Webster's dictionary of English usage*. MA: Merriam-Webster Inc.

Hands, P. (ed.) 2011. *Collins CUBUILD English grammar (Third edition).* Glasgow: HarperCollins Publishers.

Hands, P. (ed.) 2012. *Collins CUBUILD English usage (Third edition).* Glasgow: HarperCollins Publishers.

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund, Y. 2008. *Corpus linguistics with BNCweb – a practical guide.* Frankfurt am Main: Peter Lang.

Leech, G. 2011. "Frequency, corpora and language learning". In F. Meunier, S. Cock, G. Gilquin and M. Paquot (eds.) *A taste for corpora*, 7-31. Amsterdam: John Benjamins Publishing Company.

Leech, G, Hundt, M., Mair, C. and Smith, N. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Leech, G. and Svartvik, J. 2002. *A communicative grammar of English (Third edition).* Harlow: Pearson Education Limited.

Mair, C. and Leech, G. 2006. "Current change in English syntax". In B. Aarts and A. MacMahon (eds.) *The handbook of English linguistics*, 318-342. Oxford: Blackwell.

Peters, P. 2004. *The Cambridge guide to English usage*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language*. London: Longman.

Sinclair, J. (ed.) 2005. *Collins CUBUILD English grammar (Second edition)*. Glasgow: HarperCollins Publishers.

Swan, M. 2005. *Practical English usage (Third edition).* Oxford: Oxford University Press.

Discourse characteristics of English in news articles written by Japanese journalists: 'Positive' or 'negative'?

Fujiwara Yasuhiro
Aichi University of Education
fujiwara@auecc.aichi-edu.ac.jp

1 Research aim

This research project aims to detect, if any, distinctive discourse/pragmatic features in the English written by Japanese professional 'users' of English, that are journalists who contribute to English newspapers issued in Japan. The main research question of this study is, "What, if any, are differences in the use of English between Japanese professionals and native speakers of English, even though the Japanese writers have already acquired this language at the level of mastery, and use it in an international context?"

2 Background: learners or users?

This research focus is inspired by a recent debate, re-raised mainly by the 'English as a Lingua Franca' school (ELF, Jenkins 2000; Seidlhofer 2004), on the validity of 'native speaker' as the ultimate/ absolute goal in second language acquisition (SLA). In SLA, it has been the usual practice to compare the use of English by nonnative speakers (NNS) and that by native speakers (NS), finding out differences between them, and to conclude that there is still a need for NNSs/learners to acquire this language further (Cook 1999, 2002, 2007). In other words, it is 'native speakers' who are 'models' in language acquisition, and never-ending approximation to them is required on the part of learners, especially in the Expanding Circle (i.e., EFL countries such as Japan); Outer Circle Englishes (e.g., Indian English) are considered as 'institutionalized', and consequently given a similar status to that of Inner Circle Englishes (e.g., British English), which Kachru (1985) characterizes as 'established' (For a critical discussion on this distinction, see Hino 2009).

Such a tendency has been typically seen in learner corpus studies (LC, Granger 1998, 2002). For example, a well-known corpus of varieties of English, the International Corpus of English (Greenbaum 1996), covers only Inner and Outer Circle Englishes as legitimate varieties, while Expanding Circle ones are compiled with the label of 'learner' corpus (the International Corpus of Learner English). Along with a somewhat 'default' method of analysis called 'Contrastive Interlanguage

Analysis' (Granger 1998, 2002), this research direction leads researchers in LC/SLA to presuppose that all NNSs aspire to behave like NSs and therefore, the ultimate model of the use of the language is a 'native speaker'.

However, this NS-based orientation has been, in recent years, 'under fire' (Granger 2009, p. 18), mainly by some SLA researchers such as Cook (1999) and proponents of English as a lingua franca (Jenkins 2000; Seidlhofer 2004). What they argue is that not all NNSs hope to use this language like NSs, and more fundamentally, that a language learner in SLA, who has, by definition, a mother tongue, will be a bilingual user, not a monolingual speaker of a language. That is, a prescriptive norm of monolingual native speakers over bilinguals is highly questioned. As a counter movement against learner corpus studies, Seidlhofer compiled an ELF 'user' corpus, the Vienna Oxford International Corpus of English (Seidlhofer 2004; VOICE 2011). This corpus collected English within the Expanding Circle as a legitimated variety, and has been followed by other ELF corpus projects such as ELFA (A Corpus of English as a Lingua Franca for Academic Settings, Mauranen 2003, 2006, 2007) and ACE (the Asian Corpus of English as a Lingua Franca, Kirkpatrick 2010).

Although some distinctions between learner corpora and user corpora are guaranteed, such as whether a speaker is in a real context of language use, Expanding Circle Englishes (at the moment, mainly Europeans') are codified with the different labels (i.e., learners or users), resulting in a situation where leading scholars in each field admit that samples of a learner corpus is similar to those of a user corpus (Seidlhofer 2004; Granger 2009). In fact, even Prodromou (2006, 2008), an ELF researcher, casts doubt on the proficiency level of samples in the VOICE, and outlines some proposals concerning the required conditions of what he calls 'successful bilingual speakers'.

While considering some of the previous discussions and referring to several studies cited above, this author felt the increasing necessity of compiling a 'user' corpus, totally distinct from existing learner corpora in the Expanding Circle. To this end, the author, in 2005, launched a corpus compilation project called a 'Japanese User Corpus of English' (JUCE, Fujiwara 2007). This corpus, still small in size and limited in register, aims to compile data of the use of English by Japanese professional users of English. What the corpus intends to reveal is an ultimate attainable level of the second language, and whether or not professional users of English still transfer their L1 linguistic features or culture, even at the level of mastery.

Saito Hidezaburo, a distinguished early scholar of English linguistics in Japan, states; 'The mastery of a language has for its final object the expression of the exact light and shade of meaning conceived by the speaker. ... In short, the English of the Japanese must, in a certain sense, be Japanized' (Saito 1928, p. 5). Furthermore, Nishiyama Sen, a well-known simultaneous English translator in Japan, comments that "Japanese who speak English are likely to express themselves using a style and vocabulary originating in Japanese" (Nishiyama 1995, p. 1). Therefore, by analyzing English used by Japanese professional English writers and that by Inner Circle writers, this research attempts to identify differences between them in style or discourse.

3 Research

As mentioned above, this study aims to empirically detect discourse/pragmatic features of English written by Japanese professional users. To attain this object, the author compiled a Japanese User Corpus of English, currently an approximately-one-million-word corpus of various news articles written by Japanese writers, and a self-made corpus of *TIME* (hereafter, TIME) comparable to the JUCE in size, genre, text length, and so forth. Both corpora were grammatically annotated by CLAWS4, a built-in part-of-speech tagger in Wmatrix 3 (Rayson 2009), both developed at Lancaster University.

These two corpora, the JUCE and the TIME, are compared in the following two phases; 1) part-of-speech (POS) analysis as the exploratory approach and 2) keyword analysis focusing on a particular linguistic item as the confirmatory approach. Firstly, using the methods of multivariate analysis such as cluster analysis and correspondence analysis, it was found that there is much possibility to distinguish these two types of English writers, and that some POS tags characteristic of a seemingly Japanese variety of English are perhaps prepositions (PP), articles (AT), adjectives (ADJ), and common nouns (NN) (see Figures 1 & 2).

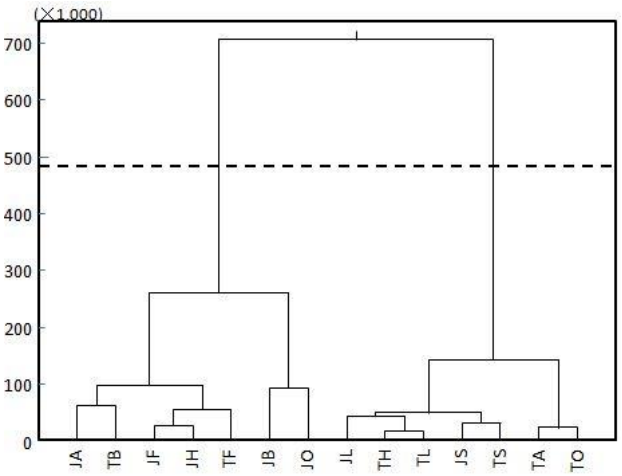


Figure 1. Results of cluster analysis

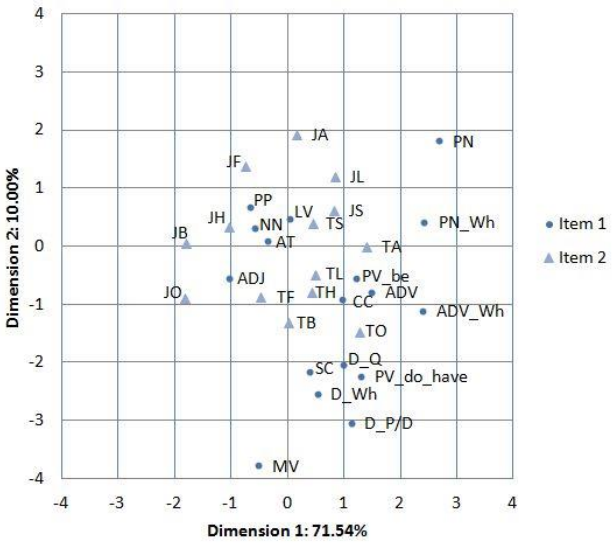


Figure 2. Plotted patterns obtained from correspondence analysis: Item 1 (POS), Item 2 (texts)

These parts of speech are normally employed to form noun phrases and much use of them naturally results in less use of pronouns. Then, this survey more closely analyzed the use of personal pronouns in the JUCE data, considering 'overuse' and 'underuse' of each pronoun, compared to the TIME and other Inner Circle English data. These comparisons revealed a clear tendency for Japanese writers of English to underuse nearly all personal pronouns, especially plural pronouns that can refer to people in general (e.g., we, you, and they).

Considering the obtained findings, this survey tentatively concludes that 'Japanese English', in reflecting Japanese language, potentially has some characteristics such as more use of content nouns with the definite articles, and less use of personal pronouns, especially ones that refer to people in general. In Japanese linguistics, the ratio of nouns to other parts of speech is said to be comparatively high, and compared to Chinese and German, the amount of nouns in ordinary Japanese is the most among the three (Muraki 1987). That is, Japanese language seems relatively noun-based. In addition to that, the use of personal pronouns in Japanese is linguistically marked and usually avoided, especially in the case of 'I' and 'you'. The reason for this phenomenon is, according to Suzuki (1996), that people are generally referred to by their position, occupation, and roles in family or society, and the use of 'I' and 'you' would in Japanese society imply the relation between opposed entities. By his detailed analysis, he even goes further, saying Japanese has 'no' equivalents of personal pronouns in western languages such as English. Furthermore, there is no similar usage referring to people in general by plural pronouns in Japanese.

That is, these features of English used by Japanese 'professional' writers can be regarded as a L1 transfer. Furthermore, in SLA, they are called even 'negative' transfers in that these are, to a significant degree, not correspondent to the NS use, and also 'fossilized' in that even after finalizing the process of acquiring this language, they still show 'deviant' use from the NS standard.

However, as mentioned before, it is apparent that the Japanese professionals intelligibly and effectively use it in an international context, conveying news to the world through English. Also, Biber et al. (1999) refer to the use of 'we' or 'you' as problematic since the referent is often obscured by the overuse of these pronouns. Thus, this study can provoke more discussions on the validity of 'native speaker' as the 'ultimate' goal; Is this a 'positive' or 'negative' transfer?

Acknowledgement

This research has been supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Young Scientists (B) 24720218, 2012-2013.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (Eds.). 1999. *Longman grammar of spoken and written English*. London: Longman.

Cook, V. 1999. "Going beyond the native speaker in language teaching". *TESOL Quarterly* 33: 185-209.

Cook, V. 2002. "Background to the L2 user". In V. Cook (ed.) *Portraits of the L2 user* (pp. 1-28). Clevedon: Multilingual Matters.

Cook, V. 2007. "The goals of ELT: Reproducing native-speakers or promoting multicompetence among second language users?" In J. Cummins & C. Davison (eds.) *International handbook of English language teaching* (pp. 237-248). New York: Springer.

Fujiwara, Y. 2007. "Compiling a Japanese User Corpus of English". *English Corpus Studies* 14: 55-64.

Granger, S. 1998. "Leaner English around the world". In S. Granger (ed.) *Learner English on computer* (pp. 13-24). Longman.

Granger, S. 2002. "A bird's-eye view of learner corpus research". In S. Granger, J. Hung, & S. Petch-Tyson (eds.) *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins.

Granger, S. 2009. "The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation". In K. Aijmer (ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam: John Benjamins.

Greenbaum, S. (ed.). 1996. *Comparing English worldwide: the International Corpus of English*. New York: Oxford University Press.

Hino, N. 2009. "The teaching of English as an international language in Japan: An answer to the dilemma of indigenous values and global needs in the Expanding Circle". *AILA Review* 22: 103-119.

Jenkins, J. 2000. *The phonology of English as an international language.* Oxford: Oxford University Press.

Kachru, B.B. 1985. "Standards, codification and sociolinguistic realism: The English language in the Outer Circle". In R. Quirk, & H. G. Widdowson (eds.) *English in the world: Teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Kirkpatrick, A. 2010. "Researching English as a lingua franca in Asia: the Asian Corpus of English (ACE) project". *Asian Englishes* 13(1): 4-19.

Mauranen, A. 2003. "The corpus of English as lingua franca in academic settings". *TESOL Quarterly* 37: 513-527.

Mauranen, A. 2006. "A rich domain of ELF – the ELFA corpus of academic discourse". *Nordic Journal of English Studies* 5(2): 145-159.

Mauranen, A. 2007. "English as an academic lingua franca – the ELFA project". *Nordic Journal of English Studies* 7(3): 199-202.

Muraki, S. 1987. "Gengo kan no imi bunya betsu goi ryo no hikaku – nihongo, chugokugo, doitsugo no baai" [A contrastive study between Japanese, Chinese, and German, in terms of semantic fields of vocabulary]. In The Editorial Committee of a Festschrift for Prof. Mizutani Shizuo (ed.) *Keiryo Kokugo Gaku to Nihongo Shori – riron to ouyou* [Computational Japanese Linguistics and Japanese Processing – Theory and Application]. (pp. 93-107). Tokyo: Akiyama Shoten.

Nishiyama, S. 1995. "Speaking English with a Japanese mind". *World Englishes* 14(1): 27-36.

Prodromou, L. 2006. "Defining the 'successful' bilingual speaker of English". In R. Rubdy, & M. Saraceni (eds.) *English in the world: Global rules, global roles* (pp. 51-70). London: Continuum.

Prodromou, L. 2008. *English as a lingua franca: A corpus-based analysis.* London: Continuum.

Rayson, P. 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University.

Saito, H. 1928. *Saito's Japanese-English Dictionary*. Tokyo: Nichieisha.

Seidlhofer, B. 2004. "Research perspectives on teaching English as a lingua franca". *Annual Review of Applied Linguistics* 24: 209-239.

Suzuki, T. 1996. *Kyoyou toshite no gengo gaku* [Linguistics as a field of Liberal Arts]. Tokyo: Iwanami Shinsho.

VOICE. 2011. What is VOICE? A computer corpus of English as a lingua franca. Available online at http://www.univie.ac.at/voice/page/what_is_voice

Negotiating TRUST during a corporate crisis: a corpus-assisted discourse analysis of BP's public letters after the Gulf of Mexico oil spill

Matteo Fuoli
Lund University
Matteo.Fuoli@englund.lu.se

1 Introduction

This paper examines the discourse strategies deployed by BP's CEO to restore stakeholders' trust after the 2010 Gulf of Mexico disaster. It focuses on the CEO's 'letter to shareholders' and 'letter to stakeholders' included in the company's annual report and social responsibility report, respectively. Considering the catastrophic consequences of the accident and the reputation damage suffered by BP due to its controversial management of the crisis, the research question I consider is the following:

> What discourse strategies does BP's CEO adopt in his public letters to restore stakeholders' trust?

To address this question, I compared the BP letters with those published by four other oil companies, before and after the accident. The analysis combines data-driven quantitative and qualitative methods. It draws on Appraisal Theory (Martin and White 2005) and Hyland's (2005) concept of metadiscourse to investigate how interpersonal language resources are deployed in the CEO's public letters to construct a trustworthy corporate identity. The results indicate that affect, evaluation and epistemicity play an important role in BP's trust-building communicative efforts.

2 Data and methods

Corpus design: In order to identify BP's trust-building discourse strategies I compared:

1. the BP letters with those of four other oil companies which were not directly affected by the disaster;
2. the letters published by BP and the other oil companies before and after the accident.

Thus, for this study I compiled a small-size specialized corpus of CEO letters published by BP and four other major oil companies in the years from 2009 to 2011 (see Table 1 for details).

All companies considered are privately owned, publicly traded and ranked among the world's 100

Companies	BP (UK), Chevron (US), ConocoPhillips (US), ExxonMobil (US), Royal Dutch Shell (UK)
Number of available letters per company	BP: 4 LSH[1], 2 LSK[2] Chevron: 4 LSH, 4 LSK ConocoPhillips: 4 LSH, 1 LSK ExxonMobil: 4 LSH, 4 LSK Royal Dutch Shell: 4 LSH, 4 LSK
Total number of letters	35
Mean letter size	LSH: 1,061 words LSK: 957 words
Total corpus size	35,811 words

Table1: Corpus details

largest corporations[3]. The corpus includes both the 'letter to shareholders' and 'letter to stakeholders' published by each company, where available. The letters are included in the companies' annual reports and corporate social responsibility reports, which are public documents that can be freely downloaded from their websites.

Framework of analysis: In order to operationalize the concept of TRUST for the analysis of the CEO letters, I adopted the model proposed by Mayer et al. (1995). The authors identify three fundamental personal attributes that can encourage trust, namely *ability*, *benevolence* and *integrity*. Simply put, someone is more likely to be trusted if he or she is perceived to be a) competent (ability), b) aligned with the interests and needs of the trustor (benevolence), and c) consistent, honest and sincere (integrity).

In line with the question raised at the beginning, the goal of my analysis is to determine how language resources are exploited in the CEO letters to convey these three attributes. I will systematically analyze how ability, benevolence and integrity are discursively constructed in the letters, using a set of analytical tools derived from Appraisal Theory and Hyland's metadiscourse framework. Both deal with the interpersonal functions of discourse and offer valuable insights into how identities and relationships are discursively negotiated. The set of categories and functional explanations offered by these two theories seem to be well suited to the analysis at hand. Ability can be communicated through positive appraisals of a company's performance (*This was an extraordinary response*), benevolence by demonstrating empathy and proximity through the use of affect terms (*We are deeply sorry*) and personal pronouns, integrity by underlining the reliability of the information provided through epistemic and modal markers (*You will see a continuing, relentless focus on safety*) and displaying ethical commitment by means of positive evaluative expressions (*Our refreshed values*).

Procedure: The language features connected to the three facets of trustworthiness were annotated in the letters using the UAM corpus tool (O'Donnell 2008). The annotation scheme comprises five main categories:

1. AFFECT: word and phrases denoting positive or negative emotions and states of mind, e.g. *We are excited about the emerging opportunities we see*.
2. EVALUATION: positive or negative explicit appraisals of a company's capacity, determination, ethical commitment, e.g. *With the talent and commitment of the people of ExxonMobil, we are strong, resilient, and well-positioned for the future*.
3. EPISTEMICITY: words and phrases expressing tentativeness/assertiveness and signaling dialogic engagement, e.g. *A subsea blowout in deep water was seen as a very, very low-probability event, by BP and the entire industry – but it happened.*
4. MODALITY: modals and semi-modals verbs of possibility, necessity, prediction, e.g. *We remain acutely aware that we must continue to address the challenge of climate change*.
5. PERSONAL PRONOUNS: e.g. *When I heard about the accident I could immediately picture how it might affect the people who live and work along that coast*.

Inter-coder agreement scores were used as a measure of reliability in the annotation process (Artstein and Poesio 2008; Fuoli 2012).

The distribution of the features listed above was compared across companies and through time. An inspection of standardized residuals based on a chi-square analysis was carried out to identify statistically significant differences in the frequency of the features considered between BP and the other oil companies, as well as between the letters published before and after the accident (see Figure 1 below). Quantitative findings were integrated with qualitative analysis of the discursive and rhetorical functions of the annotated features based on the

[1] Letters to shareholders
[2] Letters to stakeholders
[3] Financial Times Global FT 500 ranking: http://www.ft.com/intl/companies/ft500. Accessed 7 January 2013.

insights of Appraisal Theory and the metadiscourse framework.

3 Preliminary results

The mosaic plot below shows some preliminary quantitative results. The plot compares the letters to shareholders published by BP and the other oil companies before and after the accident. It schematically represents the frequency of the features annotated in the texts (box height) and marks any statistically significant deviation from the expected values through a shade/outline scheme.

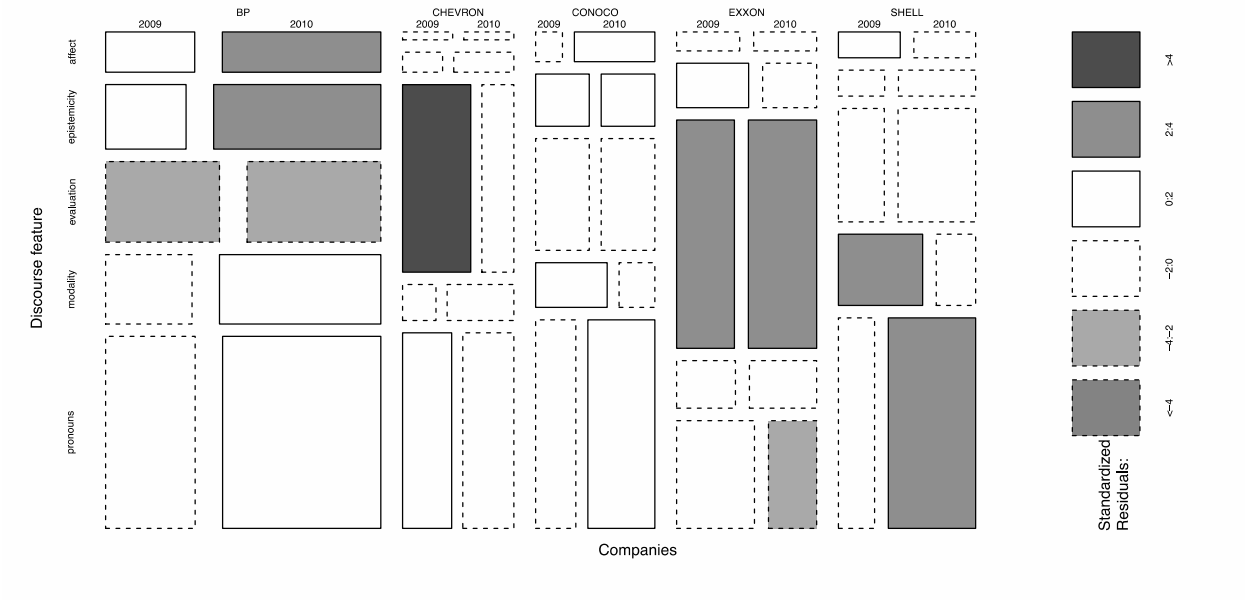


Figure 1: Oil companies' letters to shareholders before and after the disaster.

Shaded boxes with solid outline denote significantly overrepresented values, while shaded boxes with dashed outline correspond to significantly underrepresented values. The plot shows that BP's letter published after the accident contains a significantly higher number of expressions of affect (e.g. *shocked*, *close to my heart*) and epistemic/dialogic markers (e.g. *I believe it was the right thing to do*). Modals and personal pronouns are also overrepresented as compared to the previous year (solid outline), although this value is not statistically significant. Interestingly, compared to the other companies, BP's letters exhibit a significantly lower number of explicit positive self-appraisals (e.g. *ExxonMobil's 2010 results reflect the strength of our proven business model*). A qualitative analysis of the meanings that the annotated items convey and their interplay in discourse provides additional valuable insights. In BP's 2010 letter, for example, the accident is construed as an unforeseeable event through epistemic expressions such as *the unthinkable*, *previously unthinkable*, *very, very low-probability event*. Further, the first pronoun *I*, which is rarely found in the other companies' letters, is used here in the CEO's personal narrative of the accident, which, rhetorically, serves to demonstrate personal involvement and contributes to establishing empathy and solidarity with those affected by the disaster. Similarly, BP's 2010 letter includes a comparatively higher number of negative affect terms, e.g. *tragic*, *shocked*, *great sadness*, *grief*, *with a heavy heart*. These expressions occur at the beginning of the letter and can be seen to perform a similar empathy-building role.

The results of the quantitative and the qualitative analysis combined are interpreted in light of Mayer's model of TRUST. The frequent use of affect expressions and personal pronouns, for example, can be seen to contribute to communicating benevolence, i.e. the trustee's care for and alignment with the trustor. The frequent use of epistemic and dialogic markers also aids this purpose, by indexing openness to dialogue and a receptive attitude. Integrity is construed through positive appraisals of the company's ethical stand (e.g. *our priorities remained clear*, *renewed rigour*) and determination (e.g. *dauntless spirit*, *inner strength of BP and our people*) but, also, by representing the accident as unforeseeable, thereby obliterating the company's responsibility and negligence. Competence is mostly communicated in an implicit manner, through epistemic markers of certainty (e.g. *I am convinced*, *shows*) and the frequent use of the predictive modal *will*. Explicit positive appraisals of the company's capacity and performance are rare. Indeed, given the company's position, they could have come across as

inappropriate and insincere.

The analysis will be extended to the letters published in 2011 and will include a comparison of the two letter types included in the corpus, i.e. the letters to shareholders and letters to stakeholders.

References

Artstein, R. and Poesio, M. (2008). *Inter-coder agreement for computational linguistics*. Computational Linguistics, 34(4): 555–596.

Fuoli, M. (2012). "Assessing social responsibility: A quantitative analysis of appraisal in BP's and IKEA's social reports". *Discourse & Communication*, 6(1): 55–81.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London & New York: Continuum.

Martin, J. and White, P. (2005). *The language of evaluation: Appraisal in English*. London & New York: Palgrave Macmillan.

Mayer, R., Davis, J., and Schoorman, F. (1995). "An integrative model of organizational trust". *Academy of management review*, 20(3): 709–734.

O'Donnell, M. (2008). "Demonstration of the UAM corpus tool for text and image annotation". In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 13–16. Association for Computational Linguistics.

Using corpus analysis to compare the explanatory power of linguistic theories: A case study of the modal load in if-conditionals

Costas Gabrielatos
Edge Hill University
costas.gabrielatos@edgehill.ac.uk

1 Motivation, background and aims

Corpus-based examinations of the extent of modal marking (*modal load*) in *if*-conditionals in the written BNC (Gabrielatos 2007, 2010: 189-229, 266-295) have revealed that they have a significantly higher modal load (ML) than average,[1] as well as a higher ML than indirect interrogatives with *if* and *whether*, concessive conditionals with *even if* and *whether*, and non-conditional bi-partite constructions with *when* and *whenever*. Crucially, *if*-conditionals show a higher ML than conditionals with other subordinators (e.g. *assuming*). Also, the ML of *if*-conditional protases (i.e. their subordinate parts) is comparable to that of the baseline, despite their being already modally marked by *if*.[2]

Explanations for the emerging ML patterns are sought in the tenets of two recent linguistic theories: *Construction Grammar* (CxG) (e.g. Fillmore 1998) and *Lexical Grammar* (LG) (e.g. Sinclair 1996). The juxtaposition was motivated by the significant overlap in their tenets. Both theories take into account meaning (semantic and pragmatic) as well as lexical and grammatical factors. *Constructions* are symbolic units with particular features pertaining to their form and meaning. The former specify morphological, phonological and syntactic properties; the latter specify semantic and pragmatic attributes (e.g. Croft and Cruse 2004: 257-258). In LG, the unit is the *lexical item*, comprising the core (a word/phrase) and its collocates, semantic preference, semantic prosody and colligations (Sinclair 1996). The core difference between the two theories is that LG gives clear prominence to lexis over grammar (the component of colligation is optional), whereas CxG posits no distinction between them.

This study examines whether the ML of *if*-conditionals can be explained by the semantic preference of the word *if* (LG), or by the semantic component of conditional constructions (CxG).

[1] This was calculated using a random sample from the written BNC, the ML of which was used as the baseline.
[2] *If* was not included in the modal load.

2 Data and methodology

The study uses eleven random samples of 1,000 *s-units*[1] from the written BNC:

- All types of constructions, providing an indication of the average frequency of modal marking in written British English – which was used as the baseline.
- Non-conditional constructions, taken collectively.
- Conditional constructions with *assuming, if, in case, on condition, provided, supposing,* and *unless*.
- Conditional-concessive constructions with *even if* and *whether*.
- Indirect interrogative constructions with *if* and *whether*.
- Constructions with *when* and *whenever*, as they are seen as combining "condition with time" (Quirk et al. 1985: 1089), and are presented as synonymous with unmodalised *if*-conditionals (e.g. Palmer 1990: 174-175).

The methodology combined close analysis of the samples (in order to divide the constructions into clauses), manual annotation (for modal marking), and quantitative analysis. The ML was established through the interaction of two complementary metrics: modal density and modalisation spread (Gabrielatos 2010: 50-52). Modal density (MD) is the average number of modal markings per clause, and is expressed as the number of modal markings per 100 clauses. Modalisation spread (MS) is the proportion of constructions that carry at least one modal marking, and is expressed as the percentage of modalised constructions. MD helps comparisons between samples by normalising for the complexity of the constructions in each, while MS corrects for heavily modalised constructions in the sample (see Ball 1994: 297-300). The ML of constructions is represented graphically in a scatterplot as the interaction of MD and MS values (see Figure 1). The size of similarities/differences of the MD and MS values of the constructions in focus was also examined using hierarchical cluster analysis (Gabrielatos 2010: 52-54).

The empirical comparison of the explanatory power of the two theories was motivated by the following overarching hypothesis: If different types of constructions sharing the same subordinator (particularly if) show similar ML, then this could be seen as the result of the subordinator's semantic preference (SP), and would indicate support for LG. If constructions within the same family (particularly conditionals) show similar ML, irrespective of their subordinators, then this could be seen as the result of the construction's semantic component, and would indicate support for CxG.

More precisely, the baseline and non-conditional constructions provided initial reference points against which the ML of the constructions in focus could be compared. Comparisons between conditional constructions helped investigate a) whether all conditional constructions have comparable ML, and, when this was not the case, b) the extent to which the ML of a conditional construction could be attributed to its subordinator. Comparisons with even-if concessive-conditionals and indirect interrogatives with if helped investigate the extent to which ML was due to the nature of the conditional construction or the word if. Comparisons with conditional-concessives and indirect interrogatives with whether provided a reference point for conditional-concessive and indirect interrogative constructions respectively – while also providing further opportunities to examine the influence of subordinators on ML. The comparisons were carried out not only between whole constructions, but also between their subordinate parts. The latter comparison was deemed necessary, because the ML of subordinate parts can better reflect the SP of subordinators within the usual collocation span of 5 words.

The comparisons of MD and MS take into account the statistical significance of differences, using the log-likelihood statistic, with $p \leq 0.05$ as the threshold for statistical significance, and $p \leq 0.01$ indicating high statistical significance. The MD and MS values were also submitted to cluster analysis to reveal their progressive patterning.

3 Main results

The comparison of the ML of whole constructions (Figure 1)[2] and the cluster analysis (Figure 2) revealed patterns which seem to support an explanation of ML in terms of CxG:

- *If*-conditionals (*if*_cnd) have a much higher MD and MS than indirect interrogatives with *if* (*if*_q) ($p \leq 0.01$) – and the two constructions are in completely different clusters.
- The conditional-concessives with *even if* (*even-if*_cc) are in the same pre-final cluster with *if*_cnd, whereas *if*_q is found in the other pre-final cluster.
- The two indirect interrogative constructions (*if*_q and *whether*_q) cluster together, despite having different subordinators.

[1] An *s-unit* is a stretch of text delimited on either side by a sentence-boundary marker (e.g. full-stop, question mark) (Sperberg-McQueen and Burnard 2007).

[2] In Figures 1 and 3, dotted lines indicate the MD and MS of the baseline.

However,

- although most conditionals cluster together, two of them (*in case*, *on condition*) are in a different pre-final cluster;
- the two conditional-concessives (*even if* and *whether*) are in different pre-final clusters.

The examination of the ML of subordinate parts (Figure 3) and the clustering of their ML values (Figure 4), however, seem to indicate that the ML may be due to the SP of if – particularly if we consider that even if can be expected to have a different SP from if (e.g. Quirk et al. 1985: 1002).

- *if*_cnd and *if*_q have comparable ML, and cluster together .
- *even-if*_cc have much lower ML than *if*_cnd and *if*_q, and are in a different major cluster.

Still, another pattern seems to provide support for CxG:

- *whether*_q have comparable ML with both *if*_q, and share the same major cluster.

4 Brief discussion

At first glance, neither the SP of the subordinator, nor the type of construction, on their own, seem able to fully explain the results. What seems to best explain the ML patterns is their combined effect. In this light, CxG clearly demonstrates a stronger explanatory power, as a construction specifies morphosyntactic, lexical and semantic attributes (among others), which it also treats as having equal importance. This is also supported by a closer examination of issues pertaining to the ML patterns within the immediate co-text of the subordinator and the nature of LG.

The immediate co-text (subordinate parts) had to be defined grammatically, not lexically. This was because if is not a 'free agent'. On its own, if is found in two constructions (conditionals, indirect interrogatives), and in further two as part of a multi-word subordinator: conditional-concessives (even if) and comparison clauses (as if) (Quirk et al. 1985: 1110). Therefore, a collocational analysis of if would not reveal useful patterns, as it would provide a homogenised picture of its SP in the four constructions taken together − with prominence on its collocates in if-conditionals, as they account for an estimated 85% of its occurrences (Gabrielatos 2010: 194). Finally, a collocational analysis of if aiming at establishing its semantic preference would not be posited within LG, as it does not treat function words as cores of lexical items.

The combined influence of subordinator and construction type is fully consistent with the tenets of CxG. More so, CxG accounts for the interaction between all components of a construction through

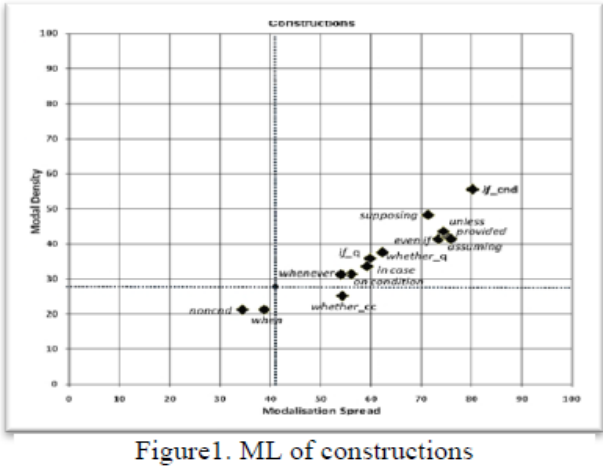


Figure1. ML of constructions

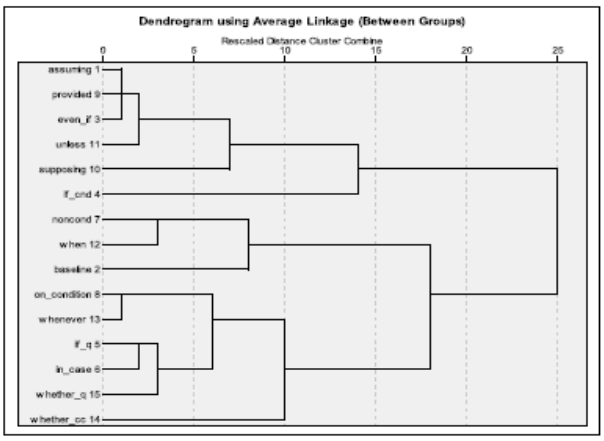


Figure 2. Clustering of ML (constructions)

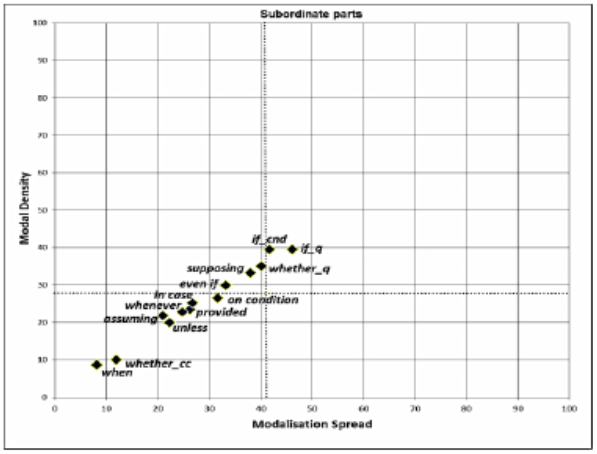


Figure 3. ML of subordinate parts

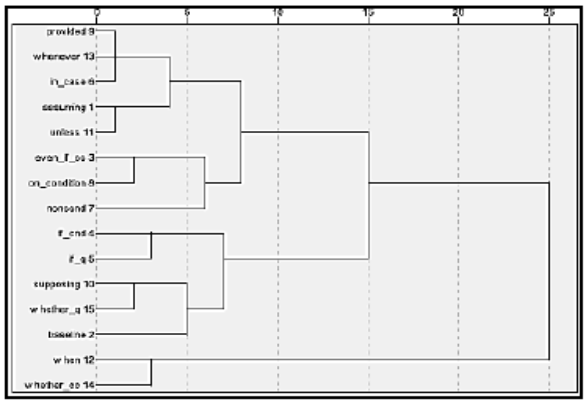


Figure 4. Clustering of ML (subordinate parts)

the "principle of no synonymy" (Goldberg, 1995: 67): morphosyntactic (including lexical) differences between constructions lead to semantic/pragmatic differences, and vice versa.

5 Conclusions

The analysis has provided strong indications that CxG rather than LG can account for the ML patterns examined here. It was also shown that ML patterns are sensitive to different combinations of constructional attributes, as it would be predicted by the principle of no synonymy. This suggests that subordinators, rather than being the core of a lexical item, are better seen as one of many components defining a construction. Consequently, if a semantic attraction of the subordinator can be posited, this has to be understood as being influenced by the type of construction that the subordinator is used in. In this light, semantic preference could be more usefully treated as part of a construction's semantic component.

References

Ball, C.N. 1994. "Automated text analysis: Cautionary tales." *Literary and Linguistic Computing* 9 (4): 265-302.

Croft, W. and Cruse, D.A. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Fillmore, C.J. 1998. "The mechanisms of "Construction Grammar"." In S. Axmaker, A. Jaisser and H. Singmaster (eds.) *General Session and Parasession on Grammaticalization.* Proceedings of the Fourteenth Annual Meeting of Berkeley Linguistics Society, February 13-15, 1998 (pp. 35-55). Berkeley: Berkeley Linguistics Society.

Gabrielatos, C. 2007. "If-conditionals as modal colligations: A corpus-based investigation." In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007*. Birmingham: University of Birmingham. Available online at bit.ly/ModalColligations

Gabrielatos, C. 2010. *A corpus-based examination of English if-conditionals through the lens of modality: Nature and types*. Unpublished PhD thesis, Lancaster University. Available online at bit.ly/CG-Thesis

Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Palmer, F.R. 1990. *Modality and the English Modals* (2nd ed.) Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair J.McH. 1996. "The search for units of meaning." *Textus* 9 (1): 75-106.

Sperberg-McQueen, C.M. and Burnard, L. 2007. *TEI P5: Guidelines for electronic text encoding and interchange*. The Text Encoding Initiative Consortium. Available online at http://www.tei-c.org/P5/Guidelines/AI.html

Digital corpora and other electronic resources for Maltese

Albert Gatt
University of Malta
albert.gatt@um.edu.mt

Slavomír Čéplö
Charles University
bulbul@bulbul.sk

1 Introduction

This paper describes the development and current status of digital corpora and other NLP resources for Maltese. After an introduction which briefly examines the linguistic situation and the history of corpus development for Maltese, the paper will focus on the efforts of two teams of researchers to create a digital corpus and set of related tools for contemporary and historical Maltese. These efforts, while largely independent at the outset, are in the process of being aligned.

2 Corpus linguistics and Maltese

The groundwork for current initiatives to collect machine readable data for Maltese was laid in the MaltiLex Project (Rosner et al., 2000). The stated aim of Maltilex was to construct an electronic Maltese lexicon, based on corpus data; however, the resulting corpus was of a relatively small size and lacked any meaningful structural or grammatical annotation. More recently, Ussishkin et al. (2009) used web resources to create a medium-sized corpus, primarily for use in the extraction of lexical resources to inform experimental work on Maltese lexical processing. Two recent European initiatives, Clarin and METANET4U have also provided impetus for further development of Maltese language resources: while work within Clarin was mainly focused on the digitisation of resources within the humanities, the METANET initiative aimed to build a common, Europe-wide infrastructure to accomodate corpora and text and speech processing tools.[1]

The present paper focuses on two more recent efforts to build corpora and related tools for Maltese, on a much larger scale. We summarize the challenges involved in data collection and text preparation as well as the choice of corpus management tool and related issues. We focus especially on issues related to opportunistic data collection from the web, and a description of the methods used to harvest text from web pages and online documents in various formats. The resulting corpora – the *MLRS Corpus*[2] running on the IMS Open Corpus Workbench, recently released in version 2.0 with part of speech tagging, and the beta version of the *bulbulistan corpus*[3] based on the NoSketchEngine – are then briefly introduced with some comments on the impact of *MLRS Corpus* as a linguistic resource on both researchers and the lay public. While *MLRS* and *bulbulistan* arose as separate initiatives, efforts are underway to make them compatible.

3 Data preparation and annotation

The next section goes on to describe the process of adaptation of the data in some detail. In particular, we describe

(i) the creation of a spell-checking dictionary through crowd-sourcing and its use in the creation of version 2.0 of the MLRS corpus;
(ii) the use of a dictionary-based algorithm to correct Maltese text written without diacritics (a common practice whereby, for example, *ħ* is written as *h*) and alternative ways of dealing with this problem;
(iii) ongoing efforts to harness these resources to develop better spell-checking algorithms;
(iv) the addition of new levels of annotation in the corpora.

Annotation is then discussed in some detail with special attention devoted to POS-tagging and related issues. These include the choice of tagset with regard to Maltese as a Semitic language with a hybrid morphology (including a highly productive, non-Semitic component), the choice of tagger and the process of POS-tagging with minimal available manually tagged data. The resulting versions of both corpora (the current, tagged version 2.0 of the *MLRS Corpus* and the newly released tagged beta version *bulbulistan corpus*) are then discussed. We also consider the benefits of a multi-level approach to POS annotation, whereby text is first tagged with basic, category-level information, with subsequent morphological analysis to include more fine-grained POS-level information (such as number, gender, and pronominal suffixes).

4 Balance and representativeness

One of the challenges with both *MLRS Corpus* and the *bulbulistan corpus* arises from their being opportunistic corpora. This is chiefly manifested in their composition where journalistic texts are

[1] Several tools developed for the Maltese Language Resource Server, one of the corpora discussed here, are now available as web services under the METANET4U framework, including a POS Tagger, tokeniser and sentence splitter. See http://metanet4u.research.um.edu.mt

[2] Maltese Language Resource Server, Malta, accessible at http://mlrs.research.um.edu.mt/.

[3] Bratislava / Prague, accessible at http://www.bulbul.sk/bonito2/.

overrepresented and some text types and genres are represented scarcely or not at all (cf. Table 1 and 2).

Text type	Number of tokens
Journalistic texts	68.800.000
Parliamentary debates	43.400.000
Belles lettres	375.000
Academic texts	170.000
Legal texts	4.800.000
Religious texts	403.700
Speeches	18.000
Web pages (blogs etc., including Maltese Wikipedia articles)	6.500.000
Miscellaneous other texts	123.000

Table 1: MLRS corpus

Text type	Number of tokens
Journalistic texts	80.000.000
Parliamentary debates	50.000.000
Belles lettres	400.000
Academic texts	100.000
Other (blogs, ads etc.)	50.000

Table 2: bulbulistan corpus

We discuss current efforts and future plans for transition from this model to a more balanced and representative one with some attention devoted to the bilingual nature of Maltese society and what this means for the ideas of representativeness and balance (considering e.g. the comparatively low proportion of texts in some subject areas such as economics or mathematics as compared to other languages of similar size and status). Various proposals for creating a balanced and representative subcorpus are introduced and the methods for their creation are proposed.

5 Diachronic dimension

In their current versions, both corpora are primarily synchronic, but efforts have been made to add a diachronic dimension by including older texts. We will discuss these efforts and the theoretical and practical challenges they pose. These involve diverse issues ranging from the definition of what counts as older (currently the focus is on the period between 19th century and the establishment of official orthography in 1921) through selection and collection of data to its annotation. This includes dealing with various systems of orthography which for the most part exhibit significant variation and inconsistency and in some cases use non-Latin characters, all of which raises interesting problems in text normalization and processing.

6 Beyond corpora

In the course of compiling the corpora, a number of related tools and resources have been created to aid in the computer-assisted processing of Maltese. The tools comprise a sentence splitter, chunker, tokenizer and POS-taggers, while the resources include full list of Maltese verbs adapted to the Semitic root structure (based on Spagnol 2011) and a library of out-of-copyright literary works in Maltese, all publicly available.

We also discuss existing and forthcoming work on solutions for computer-aided morphological and syntactic analysis, such as the inclusion of Maltese in the Grammatical Framework project (Dannélls and Camilleri 2010).

7 What's next

In lieu of a conclusion, we lay out a brief road map for further development of both corpora, including the gradual addition of more and more diverse texts and further levels of linguistic description (lemmatization, morphological analysis and rudimentary syntactic description) as well as the prospects of merging them into a single resource. We also discuss the development of other electronic resources for Maltese and the inclusion of Maltese corpus data in other projects, such as the InterCorp project and the SketchEngine resource.

References

Bovingdon, R. and Dalli, A. 2006. "Statistical analysis of the source origin of Maltese." In: A. Wilson, D. Archer and P. Rayson (eds.) *Corpus linguistics around the world*. Amsterdam: Rodopi.

Dannélls, D. and Camilleri, J.J. 2010. "Verb Morphology of Hebrew and Maltese – Towards an Open Source Type Theoretical Resource Grammar in GF." In: *Proceedings of Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects (LREC 2010)*. Malta, April 2010.

Rosner, M., Fabri, R. and Caruana, J. 2000. *Maltilex: A Computational Lexicon for Maltese*. Msida: University of Malta.

Spagnol, M. 2011. *A Tale of Two Morphologies: Verb structure and argument alteration in Maltese*. Unpublished PhD thesis, University of Konstanz. Available online at http://d-nb.info/1017360529/34.

Ussishkin, A., Francom, J. and Woudstra, D. 2009. "Creating a web-based lexical corpus and information-extraction tools for the Semitic language Maltese." In: *Proceedings of the SEPLN-SALTMIL 2009 Workshop: Information Retrieval and Information Extraction for Less Resourced Languages*, University of the Basque Country, 9-16.

The role of the speaker's linguistic experience in the production of grammatical agreement: A corpus-based study of Russian speech errors

Svetlana Gorokhova
St Petersburg State University
svetlana@sg13900.spb.edu

1 Introduction

While offering a variety of explanatory approaches, most current theories of grammatical agreement do not account for the role of the speaker's linguistic experience in adult agreement production. Studies of experimentally induced agreement errors have led different authors to conclude that agreement computation is affected by the syntactic structure of a phrase (Bock & Cutting 1992; Bock & Eberhard, 1993; Franck et al. 2006, 2010), by conceptual information such as imageability of the head and local nouns (Eberhard 1999; Humphreys & Bock 2005; Vigliocco & Hartsuiker 2002), by semantic information such as semantic integration of the head and local noun (Solomon & Pearlmutter 2004; Gillespie & Pearlmutter 2011), by whether retrieval mechanisms manage to identify the correct agreement source (Badecker & Kumianiak 2007), and by how wide the grammatical scope of planning an utterance is (Gillespie & Pearlmutter 2011). At the same time, a recent study by Haskell, Thornton, and MacDonald reveals that number agreement may be computed based on the speaker's linguistic experience (Thornton and MacDonald 2003; Haskell et al. 2010).

2 Data and methods

The goal of this paper was to study agreement errors in a highly inflected language like Russian and to find out whether agreement production *is* affected by the speaker's linguistic experience, i.e. whether the frequency of occurrence of linguistic constructions influences the occurrence of agreement errors.

Unlike most experimental studies of agreement, which investigated number agreement, the paper focuses on the production of Adj-N **case agreement** in Russian. 274 naturally produced case agreement errors recorded in Russian normal speech were analysed using frequency data from the disambiguated part of Russian National Corpus, which comprises about 6 mln. word tokens. The errors (commonly referred to as "slips of the tongue") were collected by tape-recording and digitally recording everyday conversations, telephone conversations, and live TV and radio programs such as talk shows and interviews.

In Russian, attribute adjectives must agree with the noun they modify in case, number, and gender and the adjective form has to be computed based on the head noun form. At the same time, some of the adjective case forms are homonymous.

The analysis involved naturally produced "reversed agreement" errors in modifier-head [Adj+N] constructions, when a speaker selects an irrelevant noun case form based on the case-ambiguous pre-modifier adjective form (while it is in fact the reverse that has to be done), e.g.

(a) **PL.LOC → PL.GEN**

ob	okončatel′n-**YX**	resul′tat-**AX**
about	final-PL.**GEN/LOC**	result-PL.**LOC**
→		
ob	okončatel′n-**YX**	resul′tat-**OV**
about	final-PL.**GEN/LOC**	result-PL.**GEN**

(It is too early to talk) about the final results.

(b) **PL.DAT → PL.GEN**

k	shest-I	document-**AM**
for	six-PL.**GEN/DAT/LOC**	document-PL.**DAT**
→		
k	shest-I	document-**OV**
for	six-PL.**GEN/DAT/LOC**	document-PL.**GEN**

(about fifty index cards) for six documents

(c) **F.SG.LOC → F.SG.GEN**

po	elektronn-**OJ**	počt-**E**
by	electronic-**SG.F.GEN/DAT/INS/LOC**	mail-SG.F.**DAT**
→		
po	elektronn-**OJ**	počt-**Y**
by	electronic-**SG.F.GEN/DAT/INS/LOC**	mail-SG.F.**GEN**

(I tried to send her some educational materials) by e-mail.

(d) **F.SG.GEN → F.SG.DAT**

u	et-**OJ**	kartin-**Y**
at	this-**SG.F.GEN/DAT/INS/LOC**	painting-SG.F.**GEN**
→		
u	et-**OJ**	kartin-E
at	this-**SG.F.GEN/DAT/INS/LOC**	painting- SG.F.**DAT**

This painting (can be interpreted in many ways).

(e) **F.SG.INS → F.SG.GEN**

so	svo-**EJ**	točk-**OJ**
with	own- **SG.F.GEN/DAT/INS/LOC**	point- SG.F.**INS**
→		
so	svo-**EJ**	točk-**I**
with	own- **SG.F.GEN/DAT/INS/LOC**	point- SG.F.**GEN**

(I could not see a person with his) own point of view

The examples seem to suggest that processing the adjective whose case inflection markers are homonymic, e.g. GEN/LOC, the production system is faced with ambiguous information and has to choose one of the several alternative noun case forms, which may result in the selection of a wrong form (e.g. GEN instead of LOC).

3 Results

Table 1 shows sample results of a comparison of target and error modifier-head [Adj+N] construction frequencies in the disambiguated part of Russian National Corpus.

[Adj + Noun] construction	Target form	Target frequency	Error form	Error frequency
eta kartina *this painting*	F.SG.GEN	1312	F.SG.DAT/LOC	1402
malen′kaja strana *small country*	F.SG.GEN	55	F.SG.DAT/LOC	73
ugolonye dela *criminal cases*	PL.GEN	28	PL.LOC	5
eti dokumenty *these documents*	PL.LOC	483	PL.GEN	1313
vse slučai *all cases*	PL.LOC	507	PL.GEN	1233
vtoraja polovina *second half*	F.SG.LOC	107	F.SG.GEN	119
ee rodstvenniki *her relatives*	PL.DAT	89	PL.GEN/ACC	553
elektronnaja počta *electronic mail*	F.SG.DAT	10	F.SG.GEN	39

Table 1. Sample frequencies of target and error modifier-head [Adj+N] constructions

The comparison between the frequencies of occurrence of target and error modifier-head [Adj+N] constructions in the Russian National Corpus reveals that speakers tend to substitute more frequent constructions for less frequent constructions (p (273) < 0.001).

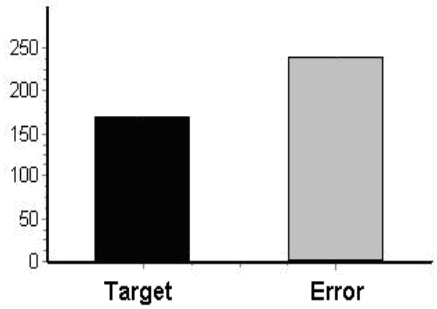


Figure 1. Modifier-head [Adj+N] constructions: Target frequencies vs. error frequencies

4 Conclusion

The result suggests that agreement production is shaped by distributional patterns in speakers' experience. The production mechanism makes use of the probabilistic information about a relevant modifier-head construction stored in long-term memory. The production of agreeing forms may thus be regarded as a lexical choice in which alternative agreeing forms compete for selection. The error construction may be a well-entrenched recurrent pattern, which a speaker, basing on their experience, tends to use as a default schema. Such lower-level agreement schemas may override generalized constructional schemas, causing agreement computation to derail. In this sense, grammatical agreement may be claimed to be stored rather than computed online.

References

Badecker, W., and Kuminiak, F. 2007. Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language* 56: 65-85.

Bock, K., and Cutting, J. 1992. Regulating mental energy: Performance units in language production. *Journal of Memory and Language* 31: 99-127.

Bock, K., & Eberhard, K. M. 1993. Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes* 8: 57-99.

Eberhard, K. M. 1999. The accessibility of conceptual number to the processes of subject-verb agreement in English. *Journal of Memory and Language* 41: 560-578.

Franck, J., Lassi, G., Frauenfelder, U., & Rizzi, L. 2006. Agreement and movement: A syntactic

analysis of attraction. *Cognition* 101: 173-216.

Franck, J., Soare, G., Frauenfelder, U., and Rizzi, L. 2010. Object interference in subject-verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language* 62: 166-182.

Gillespie, M., and Pearlmutter, N. J. 2011. Hierarchy and scope of planning in subject-verb agreement production. *Cognition* 118: 377-397.

Haskell, T.R., Thornton, R., and MacDonald, M.C. 2010. Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition* 114: 151–164.

Humphreys, K. R., and Bock, K. 2005. Notional number agreement in English. *Psychonomic Bulletin & Review* 12: 689-695.

Solomon, E. S., and Pearlmutter, N. J. 2004. Semantic integration and syntactic planning in language production. *Cognitive Psychology* 49: 1-46.

Thornton, R., and MacDonald, M. C. 2003. Plausibility and grammatical agreement. *Journal of Memory and Language* 48: 740–759.

Vigliocco, G., and Hartsuiker, R. 2002. The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin* 128: 442-472.

Keywords, lexical bundles and phrase frames across English pharmaceutical text types: A corpus-driven study of register variation

Łukasz Grabowski
Opole University
lukasz@uni.opole.pl

1 Introduction

The problem of linguistic variation in English used in various contexts of the use of medicinal products, in particular in terms of recurrent linguistic patterns, constitutes an under-researched area of linguistic investigation. This study conducted from a register perspective (Biber and Conrad 2009), which employs both quantitative and qualitative research procedures, aims to provide a corpus-driven description of vocabulary and phraseology, namely key words, 4-word lexical bundles, and phrase frames based on 4-grams in patient information leaflets (463 texts), summaries of product characteristics (146 texts) and clinical trial protocols (240 texts) written originally in English. In other words, keywords, lexical bundles and phrase frames are treated as linguistic markers of register variation.

The study is largely based on the methodology proposed by Biber (2006, 2009), Goźdź-Roszkowski (2011) and Roemer (2009), which enables one to explore lexico-phraseological profiles of three above-mentioned text varieties and to determine functions of lexical and phraseological items found therein.

2 Why study register variation across pharmaceutical texts?

The rationale behind this study is the idea of linguistic variation defined as variability in the choice of linguistic forms in different situational contexts of language use (Biber 2006). More specifically, it is hypothesized that three pharmaceutical text varieties – patient information leaflets (short 'PILs'), summaries of product characteristics (short 'SPCs') and clinical trial protocols (short 'CTPs') – found in different contexts of the use of medicinal products will prioritize different lexical and phraseological patterns and thus reveal a high degree of linguistic variation in terms of the use and function of vocabulary and phraseology (although these text varieties deal with roughly the same major theme, i.e. medicinal products).

Thus, the general aim of this study is to show that pharmaceutical text varieties can be distinguished not only on the basis of different communicative purposes, functions, situational contexts of use as well as production circumstances and typical users. On the contrary, the study aims to show that because of these differences PILs, SPCs and CTPs (treated as distinct pharmaceutical registers) reveal a high degree of linguistic variation in terms of the use and function of vocabulary and phraseology.

3 Previous studies on linguistic variation across pharmaceutical text types

Although medical language has been investigated from many different perspectives – ranging from discourse analysis (e.g. Cordella 2004; Gotti and Salager-Meyer 2006), modality (Vihla 1999), terminology (e.g. Holt et al. 2002; Worthen 2004) to applied linguistics (e.g. Hoekje and Tipton 2011) – there are virtually no studies explicitly addressing the problem of linguistic variation in English used in pharmaceutical contexts, in particular in terms of recurrent linguistic patterns. Most linguistically-oriented studies have either considered pharmaceutical text varieties to be parts of medical discourse (Gotti and Salager-Meyer 2006), or focused on a single text variety and a limited selection of linguistic features (e.g. Gledhill 1995a, 1995b, 1996; Paiva 2000). It appears, however, that there are virtually no studies aimed to show that language used in various contexts of the use of medicinal products varies depending on a text variety or discourse community. Consequently, there are no readily available descriptions of linguistic variation in a particular pharmaceutical text variety relative to other text varieties. The current state of affairs ignores heterogeneity and variability found in text varieties used in various contexts of the use of medicines.

4 Research material and study stages

The research material encompasses 463 PILs (474,458 tokens), 146 SPCs (670,907 tokens) and 240 CTPs (468,957 tokens) collected in three domain-specific custom-designed corpora. The study itself – largely based on the methodology proposed by Biber (2006, 2009), Goźdź-Roszkowski (2011) and Roemer (2009) – was conducted with the use of WordSmith Tools 5.0 (Scott 2008) and kfNgram (Fletcher 2007) in three stages. First, the key words were generated against a custom-designed pharmaceutical reference corpus, including, apart from PILs, SPCs and CTPs, samples of research articles and academic textbooks on pharmacology (c. 2.5 million word tokens). In the next stage, 50 top-frequency lexical bundles in PILs, SPCs and CTPs were identified and compared in terms of their functions. The analysis ended with a comparison of the use and function of phrase frames based on 4-word lexical bundles (short '4-p-frames'). Following Biber (2009) and Roemer (2009), the analyses deal with phrase frames based on lexical bundles with a variable slot in either initial, medial or final position, which differs from the approach used in one of the later studies conducted by Roemer (2010: 103) where only n-grams with an internal variable slot are treated as 'proper' phrase frames. Since this study is primarily a descriptive account of register variation across three pharmaceutical text types rather than a pedagogically-oriented attempt at development of a phraseological profile of a single pharmaceutical text variety, the more inclusive approach to the analysis of phrase frames – encompassing identical n-grams with an initial, medial or final variable slot – has been adopted.

5 Hypotheses and research questions

In accordance with the hypothesis adopted in this study, PILs, SPCs and CTPs prioritize different lexical and phraseological patterns because of varying discipline-specific practices associated with the situational contexts of their use. In order to test this hypothesis, the study has two specific aims operationalized in the form of the following research questions:

a) to identify 'register features', such as keywords, lexical bundles and phrase frames, typical of either PILs, SPCs or CTPs;

- Are there any keywords, lexical bundles or phrase frames that are used repeatedly in PILs, SPCs or CTPs?
- Which keywords, lexical bundles and phrase frames are register-specific?

b) to determine whether any similarities or differences in the use of vocabulary (keywords) and phraseology (lexical bundles and phrase frames) are contingent on situational contexts of the use of pharmaceutical text varieties;

- What are the functions of keywords, lexical bundles and phrase frames typical of PILs, SPCs and CTPs?
- Are the functions of keywords, lexical bundles and phrase frames associated with situational characteristics of PILs, SPCs and CTPs?

- What is the degree of overlap between three text varieties?

6 Results

The results revealed that patterns of language use differ across PILs, SPCs and CTPs, which confirms topic- and function-related differences between these three registers. Firstly, PILs have more keywords marking participation, specifying pharmaceutical form of medicinal products, as well as recommendation "advisory" key words and general language keywords. SPCs, on the other hand, have more keywords referring to names of chemical substances found in medicines, names of medical conditions and side-effects, names of procedures associated with the use of medicines as well as measurement keywords. Finally, CTPs are dominated by participation, institutional and procedural keywords. As regards top-frequency 4-word lexical bundles, the study revealed that PILs are dominated by stance bundles (epistemic stance bundles, obligation/directive bundles and desire bundles), SPCs have more referential bundles (identification/focus, temporal, terminological, procedure-related and measurement bundles) and in CTPs discourse-organizing bundles prevail. Finally, as regards 4-p-frame variation, it was revealed that 20 top-frequency 4-p-frames are specific either to PILs, SPCs or CTPs. Moreover, 20 top-frequency 4-p-frames are the most productive in SPCs followed by PILs; the former have 317 variants and the latter only 288, which also translates into higher variant/p-frame ratio (VPR) in SPCs, underlining their higher pattern variability. Interestingly, the study revealed that high frequency of 4-p-frames does not go hand in hand with their high pattern variability since the most productive 4-p-frames are not exactly the same as the most frequent ones. The functional analysis revealed that discourse-organizing (e.g. *if you * any, the * of the*) and referential 4-p-frames (e.g. *your doctor will *, tell your doctor **) prevail in PILs, stance 4-p-frames dominate the SPCs (e.g. *should not be *, dose should be **) and referential 4-p-frames (e.g. *status of the *, objective of * trial*) prevail in CTPs. In general, the functions of 4-p-frames result from the functions of the majority of their variants (or textual realizations), which are lexical bundles. To sum up, the results revealed that the observed differences are linked with the situational and functional characteristics of three pharmaceutical text varieties under scrutiny. With respect to the analysis of the use and function of 4-p-frames, the results showed that pattern variability is not only content-related – as it was hypothesized by Roemer (2009) – but it can also be function-related, which is shown more explicitly by the application of the methodology of register analysis proposed by Biber and Conrad (2009).

7 Suggestions for the future

Notwithstanding some preliminary attempts at developing a functional classification of phrase frames (e.g. Roemer & Brook O'Donnell, 2009; Roemer, 2010, 2011; Gerbig, 2011), their functional interpretation constitutes a rather unexplored research area and it is definitely worthwhile addressing this problem in the future (of the diversity of functions of different lexical bundles constituting a phrase frame, in particular).

Also, it is possible to extend the scope of description of a lexico-phraseological profile of different pharmaceutical registers by investigating semantic sequences (Hunston 2008) and/or concgrams (Cheng et al. 2006, Greaves 2009) – two recently proposed approaches to the analysis of contiguous and non-contiguous multi-word units, and thus to widen the scope of linguistic markers of register variation.

Finally, the methodology employed in this paper can be re-used in any future studies on register variation across text varieties found within other professional discourses and in languages other than English (e.g. German, Polish or Russian).

References

Biber, D. 2006. University Language. A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing". International Journal of Corpus Linguistics 14 (3): 275-311.

Biber, D. and Conrad, S. 2009. Register, genre and style. Cambridge: CUP.

Bouayad-Agha, N. 2006. The Patient Information Leaflet (PIL) corpus. The Open University, accessed 12 April 2012. http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL

Cheng., W, Greaves, C. and Warren, M. 2006. "From n-gram to skipgram to concgrams". International Journal of Corpus Linguistics 11 (4): 411-433.

Cordella, M. 2004. The Dynamic Consultation: A Discourse Analytical Study of Doctor-patient Communication. Amsterdam: John Benjamins.

Fletcher, W. 2007. KfNgram. Available online at http://www.kwicfinder.com/kfNgram/kfNgramHelp.

html. Annapolis: USNA.

Gerbig, A. 2011. “Key words and key phrases in a corpus of travel writing”. In M. Bondi and M. Scott (eds.), Keyness in Texts (pp. 147-168). Amsterdam: John Benjamins.

Gledhill, C. 1995a. Scientific Innovation and the Phraseology of Rhetoric. Posture, Reformulation and Collocation in Cancer Research Articles. Unpublished PhD thesis. The University of Aston, Birmingham.

Gledhill, C. 1995b. “Collocation and genre analysis. The discourse function of collocation in cancer research abstracts and articles”. Zeitschrift für Anglistik und Amerikanistik 1: 1-26.

Gledhill, C. 1996. “Science as a collocation. Phraseology in cancer research articles”. In S. Botley J. Glass, T. McEnery and A. Wilson (eds.) Proceedings of Teaching and Language Corpora. UCREL Technical Papers Vol. 9 (pp:108-126). Lancaster: UCREL.

Gotti, M. and Salager-Meyer, F. (eds.) 2006. Advances in medical discourse analysis: oral and written contexts. Frankfurt: Peter Lang.

Goźdź-Roszkowski, S. 2011. Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study. Frankfurt am Main: Peter Lang Verlag.

Holt, R., Stanaszek, M. and Stanaszek, W. 1998. Understanding Medical Terms: A Guide for Pharmacy Practice. London: Taylor & Francis

Hunston, S. 2008. “Starting with the small words: Patterns, lexis and semantic sequences”. International Journal of Corpus Linguistics 13 (1): 271-295.

Paiva, D. 2000. “Investigating style in a corpus of pharmaceutical leaflets: results of a factor analysis”. In Proceeding of. Annual Meeting of the ACL (pp. 52-59). Hong Kong.

Roemer, U. and Brook O’Donnell, M. 2009. “Positional variation of phrase frames in a new corpus of proficient student writing”. Paper presented at AACL conference. Edmonton, Canada, 9 Oct 2009. Available online at http://www.ualberta.ca/~aacl2009/PDFs/RoemerODonnell2009AACL.pdf

Roemer, U. 2009. “English in Academia: Does Nativeness Matter?” Anglistik: International Journal of English Studies 20 (2): 89-100.

Roemer, U. 2010. “Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews”. English Text Construction 3 (1): 95-119.

Roemer, U. 2011. “Corpora, phraseology and academic discourse”. Paper presented at ELC 2011 conference. Belo Horizonte, Brazil, 11 Nov 2011. Available online at http://www.letras.ufmg.br/linguisticacorpus2011/data1/arquivos/ELC2011presentationUteRoemer.pdf

Scott, M. 2008a. WordSmith Tools 5.0. Liverpool: Lexical Analysis Software.

Tiedemann, J. 2009. “News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces”. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing 5 (pp. 237-248). Amsterdam: John Benjamins,

Vihla, M. 1999. Medical Writing: Modality in Focus. Amsterdam: Rodopi.

Worthen, D. 2004. Dictionary of Pharmacy. London: Taylor & Francis.

Lexical density in writing assignments by university first year students

Carmen Gregori-Signes
Universitat de València
carmen.gregori@uv.es

Begoña Clavel-Arroitia
Universitat de València
begona.clavel@uv.es

Lexical density lexical diversity or richness (Daller et all. 2003) are terms which refer to statistical measures used to gauge the lexical wealth of texts but also to assess students' progress. As a rule, texts with a lower density are more easily understood, and spoken texts have lower lexical density levels than written texts (Ure 1971; Halliday 1985). However, as argued by Johansson (2008), a text may have high lexical diversity (i.e. contain many different word types), but low lexical density (i.e. contains many pronouns and auxiliaries rather than nouns and lexical verbs) or vice versa. The lexical richness of a text accounts for how many different words are used in a text, while lexical density provides a measure of the proportion of lexical items (i.e. nouns, verbs, adjectives and some adverbs) in the text (Johansson 2008). Both measures have been applied in computer analyses of corpus data.

Following the line of work proposed by Johansson (2008), this article analyses both features, lexical density and lexical diversity, in two corpora of 80 essays of 150 to 200 words; that is, two pieces of writing by each student (cf. Laufer and Nation 1995). The students are first year students at university who are officially registered in a Preliminary English Test (PET) course, after having achieved an A2 level in secondary school. Accordingly, the writing task have been designed to fit the requirements of the PET (B1 in the CEFR) official examinations. The first set of written work was collected during the second week of class; the second set a month and a half later, after the students had been exposed to regular teaching during which the first five-six units of a book level PET/B1 were completed.

The objective of the present analysis is to compare the degree of both lexical density and lexical richness at the beginning and the end of the teaching period with the intention of assessing their CEFR level at the beginning and the end of an instruction period of 3 h/week. One possible drawback of this study could be the short time period between the two writing tasks. However, pedagogical reasons support our choice, since it has often been argued that students at higher levels do not depend exclusively on classroom input but learn from other sources, and thus a three month period is sufficient to show some progress in the learner.

Students' written production has always been a central part in the assessment of student's linguistic competence. Thus, knowledge of lexical density or diversity obtained through reliable quantitative and qualitative measures may provide teachers with a more accurate picture of lexical progress, at least in the case of whole groups. This lexical awareness is in itself sufficiently interesting since it may help teachers to reflect on their teaching and the suitability of their teaching materials. In our view, too much attention has been paid to mistakes or errors made by students, and much less on the lexical richness (or poverty) of their oral or written production. In spite of the difficulties of implementation, lexical richness should be a relevant factor to be taken into consideration in material design and learner's assessment, those principles certainly justify the research being described here as the first step in a larger study, which will look at patterns in lexical development.

The analysis of lexical density was carried out applying the Type Token Ratio (TTR) method since the essays were approximately of the same length (cf. Johansson 2008). Two essays by the same learner were considered enough to assess lexical richness if we measure it through *the Lexical Frequency Profile*, as argued by Laufer and Nation (1995) "which looks at the proportion of high frequency general service and academic words in learners' writing". In this article, the referents for lexical frequency are the wordlists published by Cambridge English Vocabulary Profile for A2 and B1 (McCarthy 2012), which are made available through the Internet website for English Profile and through the collection of textbooks *English Unlimited.* The analysis involved two steps. First a lexical density test (TTR) was applied in order to compare both essays. Secondly, in order to assess lexical richness according to CEFR levels, a wordlist of each one of the assignments was elaborated and compared to the official wordlists published by the English Vocabulary Profile.

The results of the analysis show the actual level of the students right after finishing secondary school (i.e., do the students really have an A2 level?) and their level after three months university instruction at PET level. They also indicate whether they are closer to either level at the beginning and the end of the teaching period.

This study shows some preliminary results of a longitudinal, larger, research project which intends to study lexical development of the same students during four years, with the intention of finding out possible patterns in their progression. The results will be compared to see similarities or differences between the developmental patterns of lexical development that both lexical density and lexical richness indicate.

References

Daller, Helmut, Roeland van Hout & Jeanine Treffers-Daller. 2003. "Lexical richness in the spontaneous speech of bilinguals". *Applied Linguistics* 24 (2), 197-222.

Laufer, B. and Nation P. 1995. "Vocabulary Size and Use: Lexical Richness in L2 Written Production". *Applied Linguistics* 16 (3):307-322 doi:10.1093/applin/16.3.307.

English Unlimited and English Vocabulary Profile. Available online at http://www.cambridge.org/gb/elt/students/zones/custom/item6889739/2325594/Adult-English-Unlimited-and-English-Vocabulary-Profile/?site_locale=en_GB&currentSubjectID=2325594.

English Profile. http://www.englishprofile.org/

Halliday, M . A . K . 1985. *Spoken and written language.* Geelong Vict.: Deakin University.

Johansson, V. 2008. "Lexical diversity and lexical density in speech and writing: a developmental perspective". Lund University, Dept. of *Linguistics and Phonetics Working Papers* 53: 61-79.

Ure, J. 1971. "Lexical density and register differentiation". J. E. Perren, J. L. M. Trim (Eds.) *Applications of linguistics*. Cambridge: Cambridge University Press: 443–452.

Geographical Text Analysis Mapping and spatially analysing corpora

Ian Gregory
Lancaster University
I.Gregory@lancaster.ac.uk

Alistair Baron
Lancaster University
A.Baron@lancaster.ac.uk

Patricia Murrieta-Flores
Lancaster University
P.Murrieta-flores@lancaster.ac.uk

Andrew Hardie
Lancaster University
A.Hardie@lancaster.ac.uk

Paul Rayson
Lancaster University
P.Rayson@lancaster.ac.uk

1 Introduction

The over-arching aim of this paper is to draw together approaches from two different fields that have traditionally had little to do with each other namely Computational Linguistics and Geographical Information Systems (GIS). This will allow the geographies contained within corpora to be properly exploited and will have a wide range of applications across many disciplines. A GIS is effectively a database management system designed to manage spatially-referenced data – in other words data which use coordinates to provide a location on the Earth's surface for each item of data that they contain. This allows the data within the GIS to be mapped, structured and analysed using geography in addition to more conventional database operations (Chrisman et al 2002; Longley et al 2001).

Traditionally, the use of GIS has been restricted to cartographic sources and quantitative data such as censuses. The challenge that this paper addresses from a GIS perspective is to allow GIS to include unstructured texts, the most widely growing form of digital data. From a Computational Linguistics perspective the techniques we are devising will allow the geography within corpora to be exploited, something that has largely been ignored to date. From an applied perspective it allows many disciplines whose primary sources are textual but who are interested in geography an entirely new method of approaching these sources.

In recent years a number of authors have experimented with using Natural Language

Processing (NLP) techniques to extract place-names from corpora and allocate coordinates to these place-names using a gazetteer so that the places can be mapped (see, for example Grover et al. 2010; Gregory and Hardie 2011; Yuan 2010). This is known as geo-referencing the data as every place-name will have a coordinate allocated to it that provides it with a point location in a real-world coordinate system such as latitude and longitude or British National Grid. It is not the aim to explain these techniques in detail in this paper as these techniques only provide a starting point. Instead the aim here is to explore what can be done with a geo-referenced text, something that involves crossing the divide between Computational Linguistics and GIS.

The Histpop collection[1] is a corpus of around 12 million words that accompanied the printed reports of the census and the Registrar General from 1801-1937. The Registrar General published reports annually and decennially on births, marriages and – most importantly – deaths. Deaths are particularly interesting as this period marked the start of the public health movement when government monitoring of, and intervention into, health – and particularly mortality – started to draw links between factors such as poor water quality and overcrowded housing on the one hand, and diseases such as cholera and measles (which was frequently fatal to children in the nineteenth century) on the other (Szreter 1991). This material consists of both many tables of geographically disaggregated statistics and large amounts of textual reports that summarise and explore these statistics. The material was digitised by the University of Essex and geo-referenced by Claire Grover and her colleagues at the University of Edinburgh. We are primarily interested in the Registrar General's reports from 1851-1911 for England and Wales, a corpus of around 2.5 million words that covers the period when mortality started to decline dramatically.

2 *Where* is the corpus talking about?

In GIS terminology a dataset for which we have point locations, such as a text georeferenced in the manner described above, is called a point layer. Because it includes point locations for each feature – in this case place-name instance – it could be mapped as a dot map however dot maps are notoriously difficult to understand when there are more than a few dozen locations. Instead a technique called density smoothing can be used to make the pattern more comprehensible. An example of this is shown in figure 1 which shows all of the place-name instances from the 1850s volumes of the Registrar General's reports. Areas shade blue have a low density of place-name instances while those in red have increasingly high densities. As well as being a simple map, figure 1 also points to some of the analytic uses of geo-referenced data. The densities have been used to identify five clusters of locations that are particularly commonly mentioned in the corpus. These have been defined using a threshold density of more than one standard deviation above the mean density. Four of the five clusters are perhaps not surprising: the industrial north-east (1), Wakefield (2), south Lancashire including Liverpool and Manchester (3), and London (5). Cluster 4 is, however, more surprising stretching from Oxford to Bedford in a comparatively rural part of the country. The absence of clusters around, for example, the South Wales coalfield or Birmingham is also surprising.

Density smoothed surfaces also provide alternative ways of ranking the importance of place-name instances. Table 1 shows two different approaches to this. The left-hand column uses traditional word frequency counts to identify the five most commonly used place-names in the corpus. The problem with these is that they have no concept of nearby places or whether a place such as London consists not just of the name "London" but also a large number of other, more local, names. The right hand column of table 1 thus uses density scores to rank the importance of place-name instances based on the densities shown in figure 1. This gives a very different impression – London is clearly the most talked about place within the corpus as all the top ranked densities of instances lie within it.

Word Frequency Count	Kernel Density
London	London
Manchester	Holborn
Liverpool	Shoreditch
Nottingham	Vauxhall
Leicester	Kennington

Table 1: Ranking place-name instances.

3 *What* is the corpus saying about these places?

Simply identifying what places a corpus is talking about is interesting, however the real interest is in what is being said about these places. One way into this is to follow up on places identified as interesting in figure 1 using conventional concordances on, for example, all of the place-names that occur in cluster 4. This asks the question "what is being said about this location?"

[1] www.histpop.org

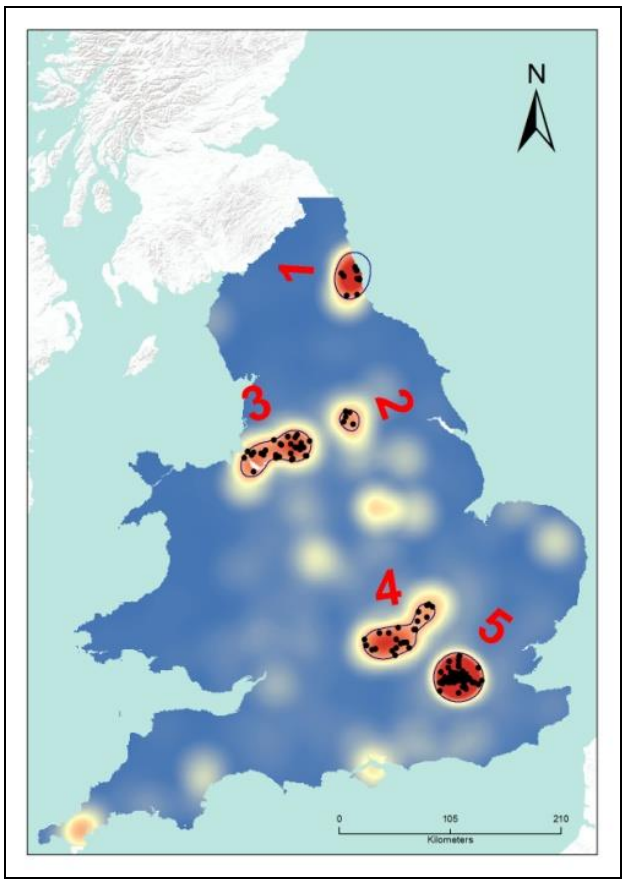


Figure 1: Density smoothed map of place-names instances from the Registrar General's reports for the 1850s.

An alternative approach would be to take a theme of interest such as a key word or a semantic tag and ask "which places are associated with this theme?" Here collocation-based approaches are needed. Take, for example, the phrase "At Royton, in Oldham, where the drainage was imperfect, typhoid fever was prevalent." This clearly associates Royton and Oldham with both typhoid fever and imperfect drainage. One issue is how near to a place-name a word has to be to collocate with it. The convention of within five words is probably too narrow in this context thus, for now, we are using within the same sentence however this requires further research.

4 Conclusions

Bringing together Computational Linguistics and GIS offers the potential to explore texts in an entirely new way. It can be used to summarise the broad geographies within a text, the geographies of particular themes, or the themes associated with particular places. Beyond the examples given here it could be used to explore, for example, whether newspapers reports of where crimes occur match actual crime statistics or whether government reports are biased towards or away from certain areas such as cities or deprived areas. We are at the early stages if this work but it has major implications. The geographical content of texts has traditionally been difficult to understand, these approaches have the potential to unlock them.

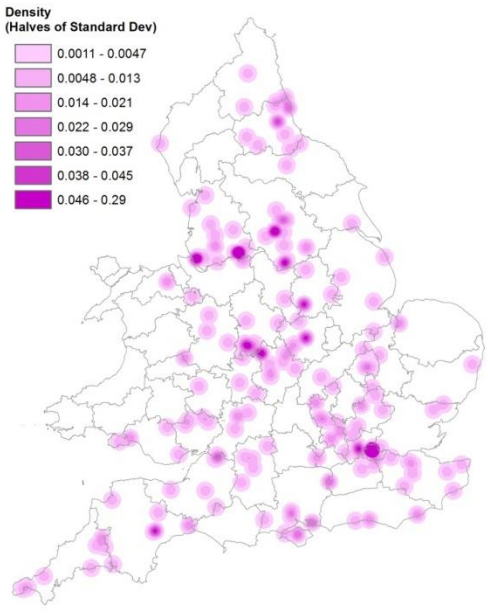


Figure 2: Places that collocate with 'measles'

Acknowledgments

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850). We are very grateful to Claire Grover (University of Edinburgh) for providing the georeferenced Histpop corpus to us, and the Richard Deswarte for his advice on this collection.

References

Chrisman, N. 2002. *Exploring Geographic Information Systems. 2nd edition.* New York: John Wiley

Gregory, I.N. and Hardie A. 2011. "Visual GISting: Bringing together corpus linguistics and Geographical Information Systems". *Literary and Linguistic Computing* 26: 297-314.

Grover, C., Tobin, R., Woollard, M., Reid, J., Dunn, S. and Ball, J. 2010. "Use of the Edinburgh geoparser for georeferencing digitized historical collections". *Philosophical Transactions of the Royal Society A* 368: 3875-3889

Longley P.A., Goodchild M.F., Maguire D.J. and Rhind D.W. 2001. *Geographical Information*

Systems and Science. New York: John Wiley.

Szreter, S. 1991. "The GRO and the Public Health Movement in Britain, 1837-1914". *Social History of Medicine* 4: 435-463.

Woods, R.I., Watterson, P.A. and Woodward, J.H. 1988. "The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part I" *Population Studies* 42: 343-366.

Woods, R.I., Watterson, P.A. and Woodward, J.H. 1989. "The causes of rapid infant mortality decline in England and Wales, 1861-1921. Part II" *Population Studies* 43: 113-132.

Yuan, M. 2010. "Mapping text." In D. Bodenhamer, J. Corrigan and T. Harris (eds.) *The Spatial Humanities: GIS and the future of humanities scholarship*. Bloomington: Indiana University Press. 109-123.

The role of phonological similarity and collocational attraction in lexically-specified patterns

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Over the course of the last 30 or so years, corpus linguistics has abandoned a sharp qualitative division between syntax on the one hand and lexis on the other. Rather, approaches such as Pattern Grammar (within corpus linguistics) and Construction Grammar (within cognitive linguistics) emphasize how the slots of more abstract elements such as patterns, or constructions, are filled with more concrete elements such as lexical items and what this reveals about the meanings, functions, and prosodies of linguistic elements. One interesting kind of elements on this scale from most concrete/constant (words) to most abstract (completely lexically-unfilled argument structure or linking constructions) are idioms and partially lexically-filled constructions. While there are always some obvious semantic restrictions, or preferences, on which lexical items may fill slots in patterns/constructions to maintain the intended/lexicalized meaning, recent exploratory work (Gries 2011) has shown that the both the lexicalization of idioms and the preferred fillers for constructional slots also appear to be affected by phonological characteristics. For instance, Gries (2011) showed that

in lexically-filled V-NP_{DO} idioms, the verb and the words in the DO-NP have a tendency to alliterate (relative to non-idiomatic V-NP_{DO} patterns);
in the partially lexically-filled *way*-construction (e.g., *he fought his way through the crowd*), the verb is more likely to begin with [w] relative to V-way_{DO} patterns that are not the *way*-construction.

In addition, the fillers in the idiomatic patterns in question were characterized by collocational attractions than the non-idiomatic patterns (according to both *MI* and *t*-scores). While this initial exploration yielded unexpected but significant and interesting results, there are several obvious ways in which Gries (2011) should be improved; the present paper explores two main ways to go beyond his initial analysis.

The first main way has to do with using more appropriate ways to study phonological/articulatory similarity. First, the measure of phonological similarity adopted by

Gries is somewhat crude and even for studying alliteration it is necessary to go beyond just the initial segment. Therefore, in this study, I will not just include the first sound of the verbs/nouns in question, but the complete onset. This will yield the same result for *bite the bullet* (since both onsets involve only the initial sound of the word), but recognize more precisely that *gain some ground* only shares the first /g/ but not the complete onset. (Time permitting, I will also discuss cases where the phonological similarity is not in the onset, as in [**meɪ**k hed**weɪ**].) Similarly, I will not just explore the sounds *per se* but also the articulatory similarity, which means that the partial similarity of *get the boot* will be recognized (/g/ and /b/ share +plosive and +voice). Second, if Gries's hypothesized cognitive explanation for the lexicalization preferences was correct, one would expect to find the same similarity effects in proverbs (as in [gɪv ðə **d**evɪl hɪz **d**ju:]). Gries mentions one example ([ðə **kæt** ɪz aʊtə ðə **bæg**]) but doesn't study this systematically; in this study, I will apply the above measures to a sample of proverbs.

The analysis of this data set poses interesting challenges in terms of how multiple levels of similarity can be incorporated into the analysis. Recent corpus-based work by Snider (2009) involves a cluster-analytic similarity measure that is useful in this context, the so-called Gower's metric, which is a measure of similarity that can quantify similarity based on multiple numeric and non-numeric indices and, thus, can integrate both the binary similarity of sharing or not sharing +voice and numeric measures such as Levenshtein's string edit distance.

The second main way of improvement over Gries (2011) has to do with the collocational attraction found in the idioms. Gries uses collocational measures that fail to capture any potential directionality effect (*MI*, *t*, $p_{\text{Fisher-Yates}}$) even though it seems likely that in some cases at least, say, the nounDO selects for a particular verb (or vice versa). Thus, in this study I will discuss the role of directionality of the collocational attraction using Ellis's (2007) Δ*P*, which in the lexically-filled V-NP_{DO} idioms allows to contrast the attraction from the verb to the head noun of the DO to the attraction from the head noun of the DO to the verb.

While these case studies – just like the one it intends to improve on – are still somewhat exploratory, the results speak to a greater relation between phonology and syntax than allowed for by traditional theories (cf. also Schlüter 2003) by illustrating how articulatory characteristics facilitate lexicalization. In addition, they showcase the power of a corpus-based approach even to lexically fully-specified expressions where little of the variability in slots that corpus linguistics usually study can be found. The data analyzed are from the Collins Cobuild Dictionary of Idioms, the CELEX database, and a collection of proverbs I am currently compiling.

References

Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27: 1–24.

Gries, Stefan Th. 2011. Phonological similarity in multi-word symbolic units. *Cognitive Linguistics* 22: 491–510.

Schlüter, Julia. 2003. Phonological determinants of grammatical variation in English: Chomsky's worst possible case. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 69-118. Berlin, New York: Mouton de Gruyter.

Snider, Neal. 2009. Similarity and structural priming. In Niels A. Taatgen & Hedderik van Rijn (eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, 815–20. Austin, TX: Cognitive Science Society.

A triangulated approach to media representations of the British women's suffrage movement

Kat Gupta
University of Nottingham
alxkg2@nottingham.ac.uk

1 Introduction

The British women's suffrage movement was a complex, diverse campaign that emerged in the mid-nineteenth century. The suffrage movement was not a unified one; it was composed of various groups with differing backgrounds, ideologies and aims. Historians working with suffragist-produced texts have noted different terminology used to describe different factions of the movement. Less attention has been paid to how the suffrage movement was perceived by those outside the movement, and particularly how it was represented in the press.

Focusing on *The Times*, I examine how suffrage campaigners' differing ideologies were conflated in the newspaper, particularly in connection with their support of or opposition to militant direct action. To demonstrate this, I use a range of methodological approaches drawn from corpus linguistics and discourse analysis including collocational analysis, examination of consistent significant collocates, critical discourse analysis and van Leeuwen's taxonomy of social actors. My research offers an innovative insight into contemporary public understanding of the suffrage campaign with implications for researchers examining large, complex protest movements.

2 The corpora

Two corpora were constructed out of texts from The Times Digital Archive. The year 1908 was selected as a start point because it was the year before the direct action campaign began; 1914 was selected as a suitable endpoint because it heralded the outbreak of World War I and, with it, corresponding shifts in both the focus of *The Times* and its production.

The corpora were extracted from xml files using the search term *suffrag**, so including texts containing *suffrage, suffrages, suffragist, suffragists, suffragette* and *suffragettes*. The contents of the corpora are summarised in Table 1.

	Average texts per year	Tokens
Suffrage	546	7,089,889
Letters to the Editor (LttE)	85	395,597

Table 1. The corpora created for this study

3 Terminology and direct action

Historians' research indicates that two terms were used to describe suffrage campaigners, each with different profiles. The term 'suffragist' tended to be used to describe constitutionalists who campaigned by lobbying Parliament. The term 'suffragette' was originally a pejorative and was used to describe campaigners who, variously, saw the vote as an end unto itself, were members of a militant organisation and/or were prepared to engage in direct action (Holton 1986). However, as shown in Table 2, the term preferred by *The Times*' for suffrage campaigner was *suffragist*[1].

	suffragist	***suffragists***	***suffragette***	***suffragettes***
1908	165	302	30	57
1909	139	311	16	35
1910	156	197	9	11
1911	81	104	2	10
1912	388	502	18	30
1913	527	615	34	19
1914	263	270	15	21

Table 2. Frequency of *suffrag** terms

This extended to *The Times'* coverage of direct action. Direct action terms[2] were associated with *suffragist* rather than *suffragette*; this runs counter to how suffrage campaigners tended to self-identify and how they are described in the historical research. I examined six direct action terms in detail: *disturbance**, *outrage**, *violence*, *crime**, *incident?* and *disorder*. Through a combination of collocational analysis and detailed understanding of the historical context, I established different semantic profiles for the terms and show how they are used to describe different activities, encounters with different groups of people and at different points in the escalating direct action campaign.

4 Emily Wilding Davison

An examination of the yearly then monthly 'peaks

[1] This is in contrast to the present day preference for *suffragette*
[2] Obtained through calculating Mutual Information with a score of 3 or higher

and troughs' (Gabrielatos et al. 2012) of the frequency of the *suffrag** direct action terms described in the previous section identified June 1913 as a peak of suffrage direct action. This peak in the lexically-driven corpus analysis is complimented by the historiographical focus on June 1913as the month in which Emily Wilding Davison was struck by a horse at the Epsom Derby and subsequently died of her injuries (Stanley and Morley 1988; Crawford 1999). Again, relatively little attention has been paid to how Davison's actions, death and their aftermath was presented in the press.

As Rosen (1974) describes, Davison had a fraught relationship with WSPU leadership; her independence and inclination to engage in unsanctioned direct actions put her into conflict with the control exerted by the Pankhursts – an ambivalence reflected in the news reporting. Using approaches drawn from critical discourse analysis, I explore how Davison was portrayed in *The Times*. Davison's actions and their aftermath were reported over 20 days, forming a news narrative. Davison was portrayed as different kinds of social actor (van Leeuwen 2009) and in terms of different named discourses. This close reading of a limited number of individual news articles, again informed by a detailed understanding of the historical context, serves to illuminate Davison as a figure around whom coalesced anxieties about the role of women, the hegemony of gendered separate spheres and the danger and instability posed by women outside their proper sphere.

5 What can triangulation of methods offer?

The three approaches I discuss – corpus linguistics, critical discourse analysis and historiography – offer a nuanced analysis that is responsive to the historical context.

The combination of corpus linguistics and critical discourse analysis is an established one (c.f. Baker 2006); in this analysis, I use it to move between the broad picture of thousands of texts offered by corpus analysis to the close analysis of a limited number of texts offered by critical discourse analysis.

The third component of the analysis is the detailed understanding of the time period in which the movement was situated. An analysis that does not take into account the historical context – the political, social and cultural world of the suffrage movement – cannot adequately account for the complexities of the suffrage movement. As Table 2 illustrates, even the terminology used by suffrage campaigners to identify themselves had associations that are not immediately obvious to a present day reader.

Through this combination of established historical approaches, discourse analysis and corpus linguistic methodologies, this investigation refines our understanding of the suffrage movement in its socio-historical context and offers an insight into the media representation of complex political campaigning and activist organisations.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Crawford, E. 1999. *The Women's Suffrage Movement: A reference guide 1866-1928*. London: Routledge.

Gabrielatos, C., McEnery, T., Diggle, P. J. and Baker, P. 2012. "The peaks and troughs of corpus-based contextual analysis". *International Journal of Corpus Linguistics* 17 (2): 151-175

Holton, S. S. 1986. *Feminism and Democracy: Woman's suffrage and reform politics in Britain, 1900-1918*. Cambridge: Cambridge University Press.

Rosen, A. 1974. *Rise Up, Women! The Militant Campaign of the Women's Social and Political Union 1903-1914*. London: Routledge.

Stanley, L., & Morley, A. 1988. *The Life and Death of Emily Wilding Davison*. London: The Women's Press.

van Leeuwen, T. 2009. Critical Discourse Analysis. In J. Renkema (Ed.), *Discourse, of Course : An Overview of Research in Discourse Studies* (pp. 285-300). Amsterdam: John Benjamins.

"Obvious trolls will just get you banned": Trolling versus corpus linguistics

Claire Hardaker
Lancaster University
c.hardaker@lancaster.ac.uk

Trolling – being deliberately antagonistic online, usually for amusement's sake – is a term for behaviour that can be traced back at least as far as the 1980s (e.g. Mauney 1982; Doyle 1989; Maddox 1989). Despite this, trolling has only reached mainstream public consciousness relatively recently (see, for example, BBC 2010; Camber & Neville 2011; Morris 2011), and has received scant academic attention, especially from fields like linguistics (see, however, Donath 1999; Herring et al. 2002; Golder & Donath 2004; Shachaf & Hara 2010; Binns 2011). This is despite the fact that trolling has resulted in criminal convictions and custodial sentences, as in the cases of Colm Coss (BBC 2010; Fogg 2010) and Sean Duffy (Camber & Neville 2011; Morris 2011), amongst others. Coss and Duffy were both charged under §127 of the Communications Act (2003), and each received eighteen week custodial sentences.

Throughout my research I have endeavoured to use corpus linguistic methods in conjunction with large corpora of computer-mediated communication to analyse and understand trolling, and to begin with, I created a CMC corpus created from two Usenet groups with a combined wordcount of eighty-six million words (86,412,727). Usenet data is useful for several reasons. There are newsgroups on an extraordinary range of topics. Some have archives reaching back to the 1980s, and perhaps most usefully, newsgroup posts can be downloaded. This makes tailoring corpora for topic, chronology, language, region and so forth relatively easy. Secondly, whilst there is no direct evidence for this, trolling is said to have begun on Usenet (Tepper 1997) and indeed, one can find Usenet examples of *troll* being used to indicate deliberate online trouble-causing up to three decades ago (e.g. Mauney 1982; Maddox 1989; Miller 1990). However, there are also drawbacks to this data. Usenet is no longer as widely known about or used, probably due to the increasing prominence of feature-rich social networks. As such, the Usenet demographic may tend towards older long-term members, and this may mean that the behaviour found on Usenet may be different than that on, say, social networks. That said, this is a difficult hypothesis to test, and overall, the benefits of the data – its scope of topics, longevity, and processing compatibility – were considered greater than the drawbacks for this investigation.

The first corpus (hereon RE) was created from part of the newsgroup, *rec.equestrian*. RE's theme is equestrianism, horse competitions, and breeding, along with related topics such as animal welfare, agricultural legislation, and livestock nutrition. The second corpus (hereon SF) was created from a subset of the newsgroup, uk.sport.football. SF's theme is (English) football, event fixtures, and league tables, along with related topics such as footballer wages, club management, and refereeing decisions.

Once RE and SF were created, *WordSmith* (Scott 2009) was used to retrieve all instances of TROLL*. Searching the corpora with an open-ended wildcard resulted in around ~9% false hits (e.g. ***Troll****ope*), but using this wildcard also retrieved derivations, inflections, compounds, neologisms, and some typographic errors that might otherwise have been excluded. RE returned 2,643 instances, whilst SF returned 1,456 instances. This created an initial sub-corpus of 4,099 examples that reduced to 3,727 once the false hits were excluded. (Though *WordSmith* retrieved an impressive set of results from RE/SF, no search can currently retrieve implicit references to trolling, e.g. *it has a sub-bridge apartment*.)

The issue that I quickly discovered whilst trying to analyse and code the results for trolling strategies, however, was that this behaviour is often heavily deceptive and covert, and this aspect has continually presented two specific, inter-related problems when endeavouring to perform the best analysis possible.

In this paper, I present those two issues, primarily with the intention of seeking advice and/or insights from others into similar experiences that may be able to assist me in moving my research forwards. The two issues in question are as follows:

The first, and perhaps most serious consideration is that due to the phenomenon in question, analysis of trolling is primarily based on examples that H *claims* to *perceive* as trolling[1]. This is because typically, S's do not announce an intention to troll, since this usually backfires and

[1] Whilst not a perfect abbreviation, H should be understood to mean hearer/reader/observer, who, for purely alliterative convenience, is referenced with the pronouns *he, him, his* etc. Likewise, S should be understood as speaker/writer/producer, who, likewise, is referenced with the pronouns *she, her, hers* etc.

causes the attempted trolling to fail. The result is that most examples of Ss discussing trolling involve denials, defences of conduct, and counter-accusations.

The preponderance of H-interpretation-based examples is of concern, principally because we have no guarantee that H is correct in is judgement, nor can we be sure that he isn't lying about his interpretations of S's intentions, e.g. he might accuse someone he dislikes of trolling simply to get her banned. Alternatively, he might wrongly accuse one S of trolling, and fail to recognise that another S has been carrying out highly skilled, covert trolling for years. In short, Hs may unwittingly attack the innocent and defend the guilty.

This is not to suggest, of course, that H interpretation is not important. Indeed, it is used as an implicit measure in a number of UK statutes that are designed to deal with various types of linguistic aggression (e.g. the Malicious Communications Act 1988; the Defamation Act 1996; the Protection From Harassment Act 1997; and the Communications Act 2003). However, especially when dealing with a potentially life-changing conviction, H-interpretation *alone* ought not to constitute the only 'proof' that S is or was trolling, whether in an analysis or in a courtroom.[1]

The second issue, which is partly affected by the first, is the value of quantitatively processing trolling strategies (i.e. the methods by which a troll sets about causing offence, such as digression, (hypo)criticism, antipathy, shock, aggression). As mentioned above, H can profess mistaken or dishonest interpretations. However, beyond that, trolling can incorporate multiple strategies, and when analysing examples, Hs do not always identify the specific grievances they may have had that triggered them to make an accusation of trolling in the first place. This leaves the analyst in the position of ascribing the potential trigger.

In short, I am firmly of the opinion that corpus linguistic methodology can be an extremely useful method of investigating and describing this behaviour, but at the same time, it is also clear that it will take a great deal of care to produce thorough and, more importantly, *meaningful* analyses of pragmatic behaviours such as intention, interpretation, manipulation, and deception using these methods.

[1] Notably, the Communications Act (2003), which deals with CMC in §127, was enacted *prior* to the peak of major social networks. Further, these Acts typically use the concept of a 'reasonable person'—a common law concept of a normative, decontextualised, objective fiction whose beliefs, knowledge, and behaviours represent an idealised standard against which others are measured. Much research already shows, however, that assessments of behaviour are highly contextually dependent. What may be admirable in one context may be highly offensive in another, even if carried out by the same person, in the same place, before the same company.

References

(1988). Malicious Communications Act. http://www.legislation.gov.uk/ukpga/1988/27/contents. United Kingdom.

(1996). Defamation Act. http://www.legislation.gov.uk/ukpga/1996/31/contents. United Kingdom.

(1997). Protection From Harassment Act. http://www.legislation.gov.uk/ukpga/1997/40/contents. United Kingdom.

(2003). Communications Act. http://www.legislation.gov.uk/ukpga/2003/21/contents. United Kingdom.

BBC (2010). Jade Goody website 'troll' from Manchester jailed. *BBC News* October 29th.

Binns, Amy (2011). Don't feed the trolls: Managing troublemakers in magazines' online communities. *Mapping the Magazine* 3.

Camber, Rebecca and Neville, Simon (2011). Sick internet 'troll' who posted vile messages and videos taunting the death of teenagers is jailed for 18 WEEKS. *Daily Mail* September 14th.

Donath, Judith S. (1999). Identity and deception in the virtual community. In Marc A. Smith and Peter Kollock (eds.), *Communities in Cyberspace* 29-59. London: Routledge.

Doyle, Jennifer (1989). Re: <hick!>. *alt.callahans* 14th December, https://groups.google.com/d/msg/alt.callahans/SphfCkUsdtY/FggkP4rvoQEJ.

Fogg, Ally (2010). Do not jail the troll. *Guardian* November 04th, http://www.guardian.co.uk/commentisfree/libertycentral/2010/nov/04/trolls-are-offensive-but-not-criminals.

Golder, Scott A. and Donath, Judith S. (2004). Social roles in electronic communities. *Association of Internet Researchers (AoIR) Conference: Internet Research 5.0* 1-25. Brighton, England: 19-22 September.

Herring, Susan C., Job-Sluder, Kirk, Scheckler, Rebecca and Barab, Sasha (2002). Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society* 18, 371-384.

Maddox, Thomas (1989). Re: Cyberspace Conference. *alt.cyberpunk* 22nd October, https://groups.google.com/d/msg/alt.cyberpunk/976Vj9FPX3Q/3Ytxg-JeCdMJ.

Mauney, Jon (1982). second verse, same as the first. *net.nlang* 05th July,

https://groups.google.com/d/msg/net.nlang/YNX8PlANL_g/Uveya3fB1ZkJ.

Miller, Mark (1990). FOADTAD. *alt.flame* 8th February, https://groups.google.com/forum/?fromgroups=#!msg/alt.flame/RMlMz6ft4r8/SwPcwihXE4AJ.

Morris, Steve (2011). Internet troll jailed after mocking deaths of teenagers. *Guardian* September 13th.

Shachaf, Pnina and Hara, Noriko (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 357-370.

Tepper, Michele (1997). Usenet communities and the cultural politics of information. In David Porter (ed.), *Internet Culture* 39-54. New York: Routledge.

Lexical bundles performed by Chinese EFL learners: From quantity to quality analysis

Dick Kaisheng Huang[1]
University of Hong Kong
huangks@hku.hk

1 Introduction

For more than half a century linguists have been interested in phraseology, the study of the structure, meaning and use of word combinations (Cowie 1998). The literature has noted a phraseological tendency and its processing advantages (Sinclair 1991; Hoey 2005; Wray 2002, 2012). They are widely believed to be of great value in second language acquisition as well as foreign language learning.

Lexical bundles are a corpus accessible feature of phraseology studied by corpus linguists. They are defined as the most frequently occurring sequences of words in a given register (Biber et al. 1999). Though being not as fixed and complete as idioms, these form-meaning composites are considered as extended collocations which can help to achieve certain discourse functions.

Studies show that lexical bundles are frequently used by both first and second language speakers in either academic or other registers (Biber 2005; Biber and Barbieri 2007; Wei 2007; Paquot and Granger 2012). When discussing bundle performance, some researchers have also analyzed the distributional characteristics by classifying bundles into structural and functional types (Altenberg 1998; Biber et al. 2004; Hyland 2008).

Most descriptions of lexical bundles have focused on frequency and classification analysis but the question of how accurately second language writers or speakers use these bundles has not been fully addressed. To be able to store and output large quantity of lexical bundles is one thing, to be capable of using them in grammatically correct and functionally appropriate ways in a certain register is quite another. The description on lexical bundles needs to take accuracy into consideration and this can only be done by studying concordance lines.

The current research aims to describe and analyze lexical bundles written by Chinese EFL majors at different stages during their 4-year tertiary English learning. More specifically, it

[1] The author is a PhD candidate from the Centre for Applied English Studies, the University of Hong Kong. Special thanks are due to Professor Ken Hyland for his supervision.

intends to answer the following questions:

(1) Do senior students use lexical bundles more frequently than junior students in their timed essays?

(2) Do senior students use lexical bundles more accurately than junior students in their timed essays?

2 Methodology and findings

The dataset for the current research is a large collection (about 6000 texts, 1.7 million tokens) of timed argumentative essays taken from 3 learner corpora which represent the written language production by Chinese undergraduates majoring in English at universities across China. Two sub-corpora were re-sampled to represent junior students (Year 1 & 2) and senior students (Year 3 & 4) respectively to make contrastive interlanguage analysis (Granger 1996).

Three parameters were set up for identifying target bundles. After initial attraction for all the 3- to 5-word bundles by WordSmith 5.0 (Scott, 2010), candidates with a standardized frequency lower than 40 per million words and multi-text occurrences under 5% of the total texts have been cut off. Then Collocate 2.0 (Barlow, 2012) was used to calculate and exclude candidate bundles with MI scores lower than 3.0. Finally bundles which overlapped or included content words were removed, leaving 40 and 71 target bundles from the two sub-corpora for descriptive and statistical analysis.

	Corpus 1	Corpus 2
data size	1,000,937	677,746
types of bundles	40	71
tokens of bundles	15,176	13,007
type/token ratio	1/379	1/183

Table1: Target bundles from the 2 corpora

	Corpus1	Corpus2
overall accuracy	0.92	0.93
	p = 0.819	
grammatical accuracy	9.75	9.75
	p = 0.496	
semantical accuracy	9.45	9.55
	p = 0.556	

Table 2: Statistics on the accuracy of bundle use

By looking at the quantity of the target bundles from the 2 sub-corpora, it was found that senior students used more bundles (110 per million) than juniors (40 per million). Again the type/token ratio of all the target bundles used by senior students (1/183) is much higher than that of the juniors (1/379) showing that fewer bundles were used repeatedly by senior students and they used a wider variety of bundles in their essays.

To investigate further, I selected 40 bundles from each of the 2 lists using stratified sampling then examined 10 examples from the concordance lines of each bundle. With the help of dictionaries and a reference corpus (Wordbanks online), each bundle out of the 400 instances was judged according to its grammatical and semantical correctness in context. All the error types were then analyzed and categorized. The statistical results showed the mean grammatical accuracy of bundles used by the 2 groups of students remained unchanged (97.5%), but the semantic accuracy (95.5%) and the overall accuracy (93%) of bundles used by senior students were considerably higher than that of the juniors (94.5% and 92% respectively). These differences, however, were not statistically significant according to Mann-Whitney Test. Detailed description on bundle misuses shows that the most common grammatical/semantic errors made by both junior and senior students are agreement error and collocation error respectively.

3 Conclusion

It can be therefore concluded that, senior students tend to use lexical bundles more frequently and in a wider variety, but do not use them more accurately than juniors. In other words, Chinese EFL majors have not achieved significant progress in bundle performance during their 4 years of English learning.

The significance of the current research lies in two perspectives. Methodologically, while most contrastive interlanguage analysis are made between L2 or EFL learners corpus and native reference corpus, comparisons on bundle performance in this research has been done with different learner groups in the same population of Chinese tertiary students. Pedagogically, as noted by Meunier (2012), the impact of formulaic language to teaching English for general purposes seems to lag behind since it is more perceptible in EAP and ESP; this research therefore suggests that, phraseological competence should be further strengthened in EFL learning and teaching in China and other EFL contexts. Future research might focus on ways of teaching lexical bundles more effectively.

References

Altenberg, B. 1998. "On the phraseology of spoken English: The evidence of recurrent word-combinations". In A. P. Cowie (ed.) *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Barlow, M. 2012. *Collocate Unicode (Version 2.0)*. Houston: Athelstan.

Biber, D. 2005. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. and Barbieri, F. 2007. "Lexical bundles in university spoken and written registers". *English for Specific Purposes* 26: 263-286.

Biber, D., Conrad, S. and Cortes, V. 2004. "If you look at...: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25(3): 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Cowie, A. P. 1998. *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hyland, K. 2008. "As can be seen: Lexical bundles and disciplinary variation". *English for Specific Purposes* 27(1): 4-21.

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies*. Lund: Lund University Press.

Meunier, F. 2012. "Formulaic language and language teaching". *Annual Review of Applied Linguistics* 32: 111-129.

Paquot, M. and Granger, S. 2012. "Formulaic language in learner corpora". *Annual Review of Applied Linguistics* 32: 130-149.

Scott, M. 2010. *WordSmith Tools (Version 5.0)*. Liverpool: Lexical Analysis Software.

Sinclair, J. M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Wei, N. X. 2007. "Phraseological characteristics of Chinese learners' spoken English: Evidence of lexical chunks from COLSEC". *Modern Foreign Languages* 30(3): 280-291.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. 2012. "What do we (think we) know about formulaic language? An evaluation of the current state of play". *Annual Review of Applied Linguistics* 32: 231-254.

A complementary approach to corpus study: a text-based exploration of the factors in the (non-) use of discourse markers

Lan-fen Huang
Shih Chien University, Taiwan
Lanfen.huang@gmail.com

1 Introduction

Using corpus approaches to investigate discourse markers reveals that the type of activity and the degree of interactivity are key factors in using discourse markers (Huang, 2011). For instance, such discourse markers as *well*, *you know* and *I mean* occur more often in the dialogic genres, while *now* occurs more in the monologic genres. However, corpus methodologies are unable to give adequate explanations for the under- or over-representation of discourse markers in a particular text. This paper proposes that a qualitative, text-based analysis be used to test some hypotheses which cannot be completed with corpus techniques. It is generally believed that discourse markers are inherently interactive and the use of them contributes to the interaction between speakers. Nevertheless, is it true that the fewer discourse markers, the lower the interaction is? If speakers do not use discourse markers, they might employ other devices to signal listeners' engagement, thereby increasing the degree of interactivity. This paper also seeks to explain how and why discourse markers occur more in one text than another, in the light of contextual information, such as the roles of speakers, the relationship between them, and settings.

2 Texts for analysis

The two texts were selected from two broad categories – a highly monologic discourse mode, and a highly interactive discourse mode – respectively in the MICASE corpus (Simpson, Briggs, Ovens, & Swales, 2002). Text 1 was chosen because there are relatively few occurrences of the discourse markers under investigation. This text (MICASE: LEL485JU097[1]) was a lecture on Physics, in which the primary speaker was a senior faculty

[1] The full-length transcript is available at http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;cc=micase;view=transcript;id=LEL485JU097, retrieved 12 May 2013.

member. Text 2 (MICASE: SGR175SU123[1]) was discussion among a peer group leader and students (senior undergraduates) from a biochemistry study group. It is assumed that there are more instances of discourse markers in Text 2, the highly interactive discourse mode, than in Text 1, the highly monologic discourse mode.

3 Text of highly monologic discourse mode: lecture

The main speaker in Text 1 is constructing his institutional identity: senior faculty member. As he holds authority over the listeners in the classroom, the relationship between the speaker and listeners is primarily asymmetrical. The speaker is lecturing and maintaining the floor all the time.

Discourse markers can contribute to interaction in discourse. The lecture in Text 1 is among the activities in which interaction is feasible. However, if the speaker employs relatively few discourse markers, he might use other devices to engage listeners. Five devices in written discourse are discussed by Hyland (2009); they are used to analyse the engagement strategies in Text 1.

The first type of device, interrogatives, is an explicit engagement feature, given that questions invite listeners to respond orally or cognitively. This device is not used by the lecturer, but the second type of device, pronouns, are heavily used, the inclusive pronoun, *we* (*our*), in particular. The third type of device, directives, can be seen as constructing power differences in the classroom. This instance of directives seems to demonstrate the authority of the speaker as a lecturer, over the listeners, as students. The fourth device, references to shared knowledge, seems to address the listeners as if distinctions of power from knowledge and academic status do not exist, thus making it easier to engage the listeners through (an assumption of) shared knowledge or experience. The last device, asides and interruptions to the prevailing argument, is used to extend the information point, and offer a personal comment, focusing on the interaction between speaker and listener(s) rather than the development of the proposition.

The analysis of Text 1 indicates that discourse markers are not the primary means of increasing interaction in lectures. Possible alternatives include the use of lexical items as boundary markers and devices signalling listeners' engagement. In this analysis of engagement strategies, how speakers engage their listeners and construct their identities has been exposed.

From the choices of engagement strategies, it may be inferred that the lecturer in this text avoids conveying his academic authority, and attempts to express solidarity with the students. The lecturer rarely uses interrogatives and directives to engage listeners. While explicit questions and directives tend to add an interactional dimension, these devices imply the speaker's authority is based on knowledge, and suggest that the speaker is in control of both the exposition as well as the audience. Moreover, the use of the inclusive pronoun, *we*, to replace the first person singular pronoun *I* and the appeals to shared knowledge reduce the authority/distance relationship.

4 Text of highly interactive discourse mode: group discussion

In reference to the contexts, four interpretations are submitted to explain why the speakers in Text 2 use more discourse markers than the speaker in Text 1. First, the study group discussion in Text 2 is less likely to have been pre-prepared. This may be the reason why the speakers use discourse markers to search for lexis or content information, and to indicate a restart and repair.

Another factor in the frequent use of discourse markers can be the number of speakers. Text 2 involves more than one speaker and a mixture of students. The involvement of several speakers and constant shifts of participant turn-taking lead to the frequent use of *oh*, *well* and *you know*; therefore, this increases the frequency and total of discourse markers.

Third, it appears from the contrast between Text 1 and Text 2, that discourse markers are used in symmetrical conversations rather than asymmetrical ones. The primary identity of the interlocutors in Text 1 is either teachers or students. The teacher-student relationship between speakers is not always stable. The more symmetrical relationship developed through interaction is influential. In a sense, the instructor with his institutional identity probably uses *you know* to demonstrate solidarity, to reduce the difference of status and to downplay his authority. If the student also uses *you know*, a less asymmetrical interaction will be created.

Fourth, using discourse markers can be a way of expressing solidarity and establishing and maintaining rapport. In Text 2, both of the students (one is a peer group leader) use *like* as a discourse marker. It is not possible to identify the functions of all the uses. It seems that the use of *like* has little connection with the proposition in

[1] The full-length transcript is available at http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;cc=micase;view=transcript;id=SGR175SU123, retrieved 12 May 2013.

the discourse, but is used to construct the speakers' personae. The two speakers use *like* to show solidarity with their peers. *Like*, on the whole, is used in this discourse as an in-group marker. At the same time, the speakers present the identity of a competent student in biochemistry. The two speakers make some changes in what Sacks (1992, pp. 327-328) calls "operative identities."[1] For example, Speaker 1 makes some changes of identity as the conversation proceeds. The use of academic terms suggests that she is presenting the identity of a competent student or group leader. The use of common discourse marker *like* displays her role of a member. The identities of Speaker 1 are dynamic rather than fixed.

5 Conclusions

The text-based analysis shows that it is possible to explain why discourse markers occur more in one text than another. More interpretations of the use of discourse markers are made with reference to the type of activity and relationship between speakers. The relationship between speakers can be built on the "discourse identities" [2] and "situated identities"[3] (Zimmerman, 1998, p. 90) which speakers employ in speech. The relationship can also be formed by the changes in "operative identities" (Sacks, 1992, pp. 327-328) as speech proceeds.

In the analysis of the two texts, it is reasonable to assume that the distinctions in using discourse markers correlate with the type of activity and the speakers' identities. As the literature has reported, the use of discourse markers is sensitive to type of activity (e.g. (Fuller, 2003)) and the relationship between speakers (e.g. (Jucker & Smith, 1998)). In this paper, it is interesting to further identify that the use of discourse markers is relevant to the speakers' construction of dynamic identities in discourse. These affect the speakers' decisions, based on the type of interaction, to give priority to fluency, the engagement of the listeners and the creation of solidarity.

Regarding the connection between discourse markers and contexts, it is difficult to be precise, but it is interpreted either that there is a connection or occasionally that there is no connection and that discourse markers are being used for constructing speakers' personae.

In general, the corpus methodologies reveal general patterns of the use of discourse markers. The text-based analysis helps to better understand the contexts in which discourse markers are used and the relationship between discourse markers and contexts. This paper has explained how and why discourse markers occur more in one text than another. The corpus study and text-based analysis can be seen as complementary approaches which inform and enrich each other, thereby leading to a better understanding of the use of discourse markers.

[1] Sacks (1991: 327-8) argues that speakers have *operative identities*, which are the identities they have in the world and are not employed at the beginning of the discourse.

[2] Zimmerman (1998: 90) defines *discourse identities* thus: "Discourse identities are integral to the moment-by-moment organization of the interaction. Participants assume discourse identities as they engage in the various sequentially organized activities: current speaker, listener, story teller, story recipient, questioner, answerers, repair initiator, and so on."

[3] Zimmerman (1998: 90) defines *situated identities* as follows: "Situated identities come into play within the precincts of particular types of situation. Indeed, such situations are effectively brought into being and sustained by participants engaging in activities and respecting agendas that display an orientation to, and an alignment of, particular identity sets, for example, in the case of emergency telephone calls, citizen-complainant and call-taker."

Acknowledgement

This work was supported in part by the National Science Council, Taiwan, under the grant number NSC101-2410-H-158-012.

References

Fuller, J. M. 2003. The influence of speaker roles on discourse marker use. *Journal of Pragmatics* 35(1): 23-45.

Huang, L.-f. 2011. *Discourse markers in spoken English: a corpus study of native speakers and Chinese non-native speakers.* Unpublished PhD thesis, University of Birmingham, UK. Available online at http://etheses.bham.ac.uk/2969/

Hyland, K. 2009. "Corpus informed discourse analysis: the case of academic engagement". In M. Charles, D. Pecorari & S. Hunston (eds.) *Academic writing*. London: Continuum.

Jucker, A. H., & Smith, S. W. 1998. "And people just you know like 'wow': Discourse markers as negotiating strategies". In A. H. Jucker & Y. Ziv (eds.) *Discourse markers: description and theory*. Amsterdam: John Benjamins.

Sacks, H. 1992. *Lectures on conversation 2*. Oxford: Blackwell.

Simpson, R., Briggs, S. L., Ovens, J., & Swales, J. M. 2002. *The Michigan corpus of academic spoken English*. Retrieved February 18, 2009, from Ann Arbor, MI: The Regents of the University of Michigan. Available online at http://quod.lib.umich.edu/m/micase/

Zimmerman, D. H. 1998. "Identity, context and interaction". In C. Antaki & S. Widdicombe (eds.) *Identities in talk*. London: Sage.

Lexical bundles in private dialogues and public dialogues: A comparative study of English varieties

Dora Zeping Huang
Chinese University of Hong Kong
huangzeping@gmail.com

1 Introduction

Lexical bundles, brought into light by corpus linguistics a decade ago, are described as a type of word combinations on the basis of frequency data. Over the past ten years, most studies on this field focused on two major varieties of English (British English and American English) (e.g., Biber 2009), or learner English in the academic contexts (e.g., Nesi and Basturkmen 2009). However, the concept of English today has undergone fundamental changes with the increasingly rapid pace of globalization. English has been labeled as a lingua franca, an international language, or a global language that has been widely used for the purpose of both intranational and international communication. The research interest in the English language has no longer been restricted to the English in the core speaking countries, but shifted to the localized forms of English worldwide.

This study attempts to complement relevant research by focusing on a new perspective the core and the periphery of lexical bundles in two types of Modern English dialogues (face-to-face private dialogues and public dialogues) among four varieties of English from the International Corpus of English (ICE Corpora). These four varieties are the British English (hereinafter refers to ICE-GB), the Canadian English (ICE-CA), the Hong Kong English (ICE-HK), and the Singapore English (ICE-SIN).

2 Literature review

In the past ten years, there has never been a dearth of studies investigating lexical bundles in the English language (e.g., Biber, Johansson, Leech, Conrad and Finegan 1999; Biber, Conrad and Cortes 2004; Hyland 2008; McCarthy and Handford 2004; McCarthy and Carter 2006; Simpson 2004). As a pioneering study, Biber et al. (1999) examined three- to four-word combinations based on frequency in two registers (conversation and academic prose) from *the Longman Grammar of Spoken and Written English Corpus* (LSWE). Their findings showed striking differences in the use of lexical bundles between speech and writing.

Different from Biber et al. (1999), some research studied lexical bundles by narrowing down in one particular register rather than across registers. McCarthy and Carter (2006) investigated multi-word strings automatically retrieved from a five-million-word corpus of conversational English from Britain and Ireland. McCarthy and Handford (2004), based on *the Cambridge and Nottingham Corpus of Business English* (CANBEC), studied the two-, three-, four-, five-, and six-word clusters generated from the one million words of spoken business data recorded in a variety of business settings, including meetings, sales presentations and telephone conversations, etc. However, the English varieties under consideration in these studies are restricted to merely British English, American English or Irish English.

There have been plenty of studies reported the use of lexical bundles in the academic context (e.g., Cortes 2004; Biber 2006, 2009; Liu 2012). The research scope of these projects is limited to contrasting academic written discourse with academic spoken discourse, or native academic writing with learner academic writing by exploring texts ranging from lectures, classroom discussion, textbooks to academic writing. Research on lexical bundles beyond academic contexts is still rare.

3 Research questions

In view of the research gap, the present study attempts to explore the use of lexical bundles in two types of dialogues, namely, private dialogues and public dialogues across four varieties of English. Three research questions are addressed in this study.

What is the size of the core and periphery of lexical bundles in private dialogues and public dialogues?
Is there any linguistic variation in the use of lexical bundles between private dialogues and public dialogues?
Are there any similarities and differences in the use of lexical bundles across the four varieties?

4 Methodology

There are 360 texts of private dialogues with approximately 720,000 running words, and 280 texts of public dialogues with about 560,000 words. To sum up, there are 1,280,000 tokens in the language dataset for the current study. In order to have a thorough and comprehensive investigation, the present study sets out to examine three-word bundles by setting a

minimum frequency of five times for each genre category.

Following Nelson's (2006) framework, the core refers to items overlapping in all four varieties while the periphery refers to items only existing in one variety. To calculate the core and periphery bundles in the four corpora, all four lists of three-word bundles are combined into a single list. The new list is indexed by WordSmith. If the frequency of a bundle is 4, it means all the four corpora had this bundle. If it has a frequency of 3, it means it is a bundle overlapping in 3 corpora.

Grammatical features of core bundles are investigated on the basis of Biber et al.'s (1999) taxonomy. Lexical bundles are placed into four categories: clausal bundles, phrasal bundles, repetitive bundles, and indeterminate. Log-likelihood scores (Rayson and Garside 2000) are employed as a statistical index to identify bundles that are typical and distinct to either private dialogues or public dialogues.

5 Results

Overall, the size of the core appears to be relatively small in terms of types in both private dialogues and public dialogues. However, in terms of frequency, the core bundles occur far more frequently than the periphery bundles. There are 6,037 types of bundles in total found in private dialogues. Among these bundles, 518 types of bundles (8.6 per cent) are found in all four bundle-lists, so the "absolute core" represents just 8.6 per cent of types. At the other extreme, the absolute periphery, i.e., items occurring in one corpus only, consists of 3,779 types, or 62.6 per cent of all types. Therefore, in terms of types, the absolute core is quite small, and the absolute periphery is very large, at about over half of all types.

However, if we compare the total frequency, it is found that the total number of tokens represented by the absolute core (total frequency 48,172) outnumbers the absolute periphery (total frequency 26,161). In other words, 6.3 per cent of the whole dataset of private dialogues is composed of the absolute core bundles, that is, about two per cent bigger than that of the absolute periphery. What is remarkable is that, on average, each core bundle occurs 93 times, whereas items in the absolute periphery occur only seven times.

Similar to private dialogues, only a small number of bundles (8.6 per cent) are found in all four lists of public dialogues. At the other end, the absolute periphery, i.e., items occurring in one corpus only, consists of 2,742 types, or 65.3 per cent of all types. However, on average, the core bundle in public dialogues occurs much more frequently (66 times) than the periphery.

In terms of grammatical features, over half of the core bundles in both private dialogues (72%) and public dialogues (63%) are clausal bundles. Although phrasal bundles rank the second most types of bundles in both modes of dialogues, there are more phrasal bundles occurring in public dialogues (30%) than in private dialogues (16%). Bundles distinctive in either private dialogues or public dialogues are observed and compared based on the log-likelihood scores.

A comparison of bundles overlapping between individual varieties also yields an interesting result that there is more overlap between the data from ICE-GB and ICE-CA (ENL) than between the data from ICE-HK and ICE-SIN (ESL) in terms of lexical bundles. The proportions of bundles overlapping between the two ENL varieties are found to have more consistency and similarity. In contrast, those between the two Southeast Asian counterparts show more disparity; in particular, lexical bundles from the Hong Kong corpus appear to be most distinctive, displaying deviations from the other three datasets. Furthermore, distinctive bundles in each English variety are also identified and analyzed according to the log-likelihood scores.

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Biber, D., Conrad, S. and Cortes, V. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25* (3), 371-405.

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers.* Amsterdam/Philadelphia: John Benjamins.

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311.

McCarthy, M. and Handford, M. 2004. "Invisible to us": A preliminary corpus-based study of business spoken English. In U. Connor and T. Upton (eds.), *Discourse in the professions: perspectives from corpus linguistics* (pp167-202). Amsterdam/ Philadelphia: John Benjamins.

McCarthy, M. and Carter, R. 2006. *This that and the other:* Multi-word clusters in spoken English as visible patterns of interaction. In M. McCarthy (ed.), *Exploration in corpus linguistics* (pp7-26). Cambridge: Cambridge University Press.

Nelson, G. 2006. The core and periphery of world Englishes: A corpus-based exploration. *World Englishes, 25*(1), 115-129.

Nesi, H. and Basturkmen, H. 2009. Lexical bundles

and discourse signaling in academic lectures. In J. Flowerdew and M. Malhberg (eds.), *Lexical cohesion and corpus linguistics* (pp23-44). Amsterdam/Philadelphia: John Benjamins.

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling. Proceedings of the workshop on comparing corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). Hong Kong.

Simpson, R. 2004. Stylistic features of academic speech: The role of formulaic expressions. In U. Connor and T. Upton (eds.), *Discourse in the professions: perspectives from corpus linguistics* (pp37-64). Amsterdam/Philadelphia: John Benjamins.

SAE11: a new member of the family

Sally Hunt
Rhodes University
s.hunt@ru.ac.za

Richard Bowker
Rhodes University
r.bowker@ru.ac.za

The Brown-LOB family of corpora are well known as ground-breaking and influential one-million word corpora of English, built using variations of the same sampling frame and providing snapshots of usage of American and British English, respectively, at various points in time beginning in the 1960s. The relative uniformity of method, especially in terms of balance between genres, allows for diachronic and cross-variety comparison. Similar corpora have been built for other varieties of English, such as the Wellington Corpus of Written New Zealand English (Bauer 1993) and the Kolhapur corpus of Indian English (Shastri 1986). More recently, BE06 (Baker 2009) utilised the internet as a resource to collect the Sampling Frame Units (SFUs) in an already digital form, vastly improving ease of access, rather than scanning hard copies or typing texts in order to add them to the corpus. To date, no comparable corpus has been compiled for South African English (SAE). As part of an ongoing project to build a large monitor corpus of SAE, and following Baker (2009), in 2012 we embarked on collecting SAE11, a one-million word corpus of South African English from online sources, with the majority of texts originally published in South Africa in 2011. The process of building this corpus and the challenges we encountered are the subject of this paper. We believe that the problems we faced and the decisions we made are relevant to future corpus builders, but also make an important contribution to the ongoing discussion in the corpus community concerning best practice and the overall aims and assumptions of corpus linguistics. Broadly, the issues to be discussed in this paper include the nature of SAE as well as the availability of internet material and its status as representative of general usage. We also consider the applicability of a now rather elderly sampling frame, especially to an online medium unimaginable when the frame was devised, with genres of its own which need to be included in order to be representative, and the adjustments we made as a consequence of what we view as mismatches between the frame and usage.

The nature of South African English itself is at the root of a number of the challenges we faced while building the corpus. In fact, whether a

unified variety exists which can legitimately be called South African English is debatable. South Africa has eleven official languages, of which English is one, with nearly 10% of the population as L1 speakers. The majority are classified as Bantu languages while Afrikaans is Germanic, like English. The extent of bi- and multilingualism, especially amongst mother-tongue speakers of African languages, means that there are multiple influences on the kind of English spoken in the country, and it could be argued that there is no such thing as South African English, but rather several South African Englishes, including a mother tongue variety and second or additional language varieties (Mesthrie 2002). This makes the collection and analysis of data problematic but also potentially richer and more exciting, if distinct patterns of usage are evident from speakers of specific languages or language groups. Influences from all the major languages are evident in the corpus, including in the English of speakers of other languages. For instance, *braai*, from Afrikaans, is the standard word for *barbeque* in South Africa, and is used by speakers from all language groups. The possibility exists therefore for claiming some sort of pan-South African variety, especially with regard to lexis.

In terms of compiling the corpus, we were frequently faced with decisions about what constitutes South African English. Is a novel written by someone residing in Kenya for the last three decades and published in the UK to be regarded as South African English if the author grew up in South Africa? What of the author who grew up in neighbouring Botswana but has lived in South Africa all her adult life and publishes novels set in Cape Town, using convincing South Africanisms in her dialogue? And how were we to deal with the tendency for texts to lose 'local flavour' as their formality increases, to the extent that they appear indistinguishable from British English? Although our criteria for inclusion were strictly external, the relative scarcity of texts in some genres meant that the centrality or otherwise of some texts as representative of the variety was a real constraining factor in our collection efforts. For instance, the list of novels for adults published in English in South Africa in 2011, compiled by the National English Literary Museum which monitors and archives South African literary publications, comprises only 96 works: not only are extracts of relatively few of these available online (15 – 25%), those we could find were often subject to questions such as those above in terms of their status as SAE. This is in contrast to collection efforts for corpora based in Britain or America where not only is there a larger pool of texts from which to choose, but questions of eligibility are less pressing. In fact, it had been our initial intention to build SAE06, as a direct comparison with Baker's BE06, but the dearth of online material in even fairly prominent genres meant that we had to opt instead for 2011, a year for which much more published material is available online. Therefore, unlike Baker (2009) who reports taking 83 hours to collect the requisite one million words, our experience was that after the same time period had elapsed, we were no more than a quarter of the way, due largely to the fact that there is a much smaller quantity of published South African English on the internet than there is in British English, and thus searching for sufficient text in a specific sub-genre from a particular year to fill a SFU can take a considerable period of time.

A second dominant group of issues arose from the sampling frame. The original balance of genres in the Brown corpus was designed to reflect the proportions of published English in the USA in the chosen year. It must be asked whether the same proportions were published in 2011, especially given technological advances and the development of new genres, and in South Africa, especially given the relatively low literacy levels of the population (see Pretorius and Mokhwesana 2009). While the larger categories, such as news and fiction, are without question still published in great quantities, sub-categories such as *belles lettres* are not a particularly prominent genre currently, while blogs are not catered for in the frame. In similar vein, in South Africa, "Westerns" are not published as local literature but there is fiction which deals with similarly rural tales of adventure and which could legitimately be substituted, preserving the distinction between genres and also more adequately reflecting the English published in South Africa in 2011.

The adjustments outlined above, amongst others, will be discussed in more detail in the paper. If the ultimate goal of building this type of corpus is to create a microcosm of language use in a particular place at a particular time, then adjustments can and must be made with this in mind (cf. Baker 2009). If, however, the goal is to mirror the shape of the existing corpora to allow for comparison between corpora, then competing principles may create a tension if the original sampling frame is out of kilter with current usage. This is an ongoing debate which will require "on-the-ground" decisions, as well as abstract principled consideration.

References

Baker, P. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14:3. 312–337.

Bauer, L. 1993. Manual of Information to accompany the Wellington Corpus of Written New Zealand English. Downloaded from http://icame.uib.no/wellman/well.htm

Mesthrie, R. 2002. Language in South Africa. Cape Town: CUP

Pretorius, E.J. and Mokhwesana, M.M. 2009. Putting reading in Northern Sotho on track in the early years: Changing resources, expectations and practices in a high poverty school. In South African Journal of African Languages, 29(1), pp. 54-73.

Shastri, S.V., C.T. Patilkulkarni and G.S. Shastri. 1986. Manual of Information to accompany the Kolhapur corpus of Indian English. Downloaded from http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM#sourc

Bridging genres in scientific dissemination: popularizing the 'God particle'

Ersilia Incelli
University of Rome Sapienza
ersilia.incelli@uniroma1.it

Advances in modern communication are changing 'expert – expert' and 'expert – lay' information exchange, above all in scientific communication, which relies on effective knowledge dissemination to raise public awareness of vital scientific research being carried out, as well as to attract institutional and governmental funding.

This contribution aims to investigate the discursive properties at the interface between meaning and knowledge in a contrastive analysis of texts from 'specialist' scientific journals and popular science discourse in the media. The work draws on a collection of texts relating the recent discovery of Higgs Boson, commonly called the 'God particle'. The corpus consists of three different genre types; one consisting of articles from the scientific research journals Advanced Physics and Contemporary Physics, the second consisting of articles from scienceblog.com, and the third consisting of articles from online newspapers, namely The Guardian, The Independent. Each sub-corpus totals about 100,000 words. A fourth dimension will be explored and contrasted in two short 3 minute You Tube videos which explain the Higgs boson to the general public. Each genre presents different ways of speaking for different rhetorical purposes, therefore scientists and journalists use a range of registers moving between several repertoires (Myers 2003).

In particular, the study focuses on the lexico-grammatical choices and linguistic strategies, used to illustrate scientific data in the different genre, highlighting differences in textual form and rhetorical structure, involving different mental and process verb choices. For example, in scientific research journals there is a predominance of causal conditionals and purposive ways to argue and assert proof, for example: 'if x is y, z will happen', e.g.

> a) if a particle can move through this field with little or no interaction, there will be no drag, and that particle will have little or no mass;
>
> b) if a particle interacts significantly with the Higgs field, it will have a higher mass.

Whereas, in popularizing discourse, scientific

content is simplified for the lay public, through denominations, reformulation markers, and the recontextualization and redefinition of ideas, such as the use of metaphors to conceptualize complex scientific processes (Calsamiglia and van Dijk 2004). For example,

> a) Particles wading through the field gain heft the way a bill going through Congress attracts riders and amendments, becoming ever more ponderous.
>
> b) Without the Higgs field, as it is known, or something like it, all elementary forms of matter would zoom around at the speed of light, flowing through our hands like moonlight.

In this case study the 'God' metaphor helps simplify the universe in the attempt to bring understanding to its mass and order: 'the teams [...] still need to determine whether Higgs boson behaves as the God particle is thought to behave – and therefore what its role in the creation and maintenance of the universe is'. The phenomenon has really nothing to do with 'God', and for this reason it may even be misleading to the public. In a way the metaphor is more a form of sensationalism, like a sensational hyperbole, used to attract reader attention, to get the public interested in a major scientific event, or to increase viewership and readership numbers. In the video clips, the way scientists explain abstract meanings to the lay public using analogies and even 'allegory', is particularly worth exploring. The most popular kind of analogy in this genre, usually involves a round object like a pearl moving through some kind of syrupy substance. The You tube examples include a canteen tray full of sugar, e.g. You tube video Higgs boson Science explained using sugar and ping-pong balls; or You tube The Guardian video What is the Higgs boson?

The analysis is primarily qualitative and discourse-analytical in approach, relying on quantitative data retrieved through corpus linguistic techniques, such as keywords, collocation and cluster analysis, identifying recurrent phraseological strings and clusters pertaining to the genre. e.g. the forces that cause; particles interact with; gaining mass. In this way the methodological framework also follows a corpus-based approach which combines quantitative and qualitative techniques (Baker 2006) which contributes to a better understanding of patterns within a genre. The theoretical framework also draws on a social semiotic approach, hence multimodal discourse analysis to reveal what and how aspects of knowledge about scientific processes are communicated in the discourse of a video, referring to Van Leeuwen's (2009) model which views discourses as 'socially constructed ways of knowing some aspects of reality which can be drawn upon when that aspect of reality has to be represented or [..] context specific frameworks for making sense of things', (Van Leeuwen 2009:144).

The work also highlights the changes taking place in how the public, (still a selected community), now 'consumes' information.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum

Calsamiglia H., van Dijk T.A. 2004. "Popularization Discourse and Knowledge about the Genome", *Discourse & Society*, 15(4), 2004, Special issue Genetic and genomic discourses at the dawn of the 21st century, guest-edited by Brigitte Nerlich, Robert Dingwall, Paul Martin: 369-389.

Greaves, Chris. 2005. *Concgram* software tool.

Myers, G. 2003. Discourse studies of scientific popularisation: questioning the boundaries. *Discourse Studies*, 5 (2). pp. 265-279.

Rayson P. 2003. *WMatrix/USAS – Semantic Annotation System*. University of Lancaster

Scott. M. 2005. *Wordsmith tools*. Version 5.0. Oxford University Press.

Van Leeuwen, T. 2009. Discourse as the recontextualization of social practice: A guide. In: Wodak R. and Meyer M. (eds) *Methods of Critical Discourse Analysis*. London: Sage, 144-161

Youtube video Higgs boson Science explained using sugar and ping-pong balls. Available online at: http://www.youtube.com/watch?v=ZzQpeqE_wLg

Youtube The Guardian video. Available online at: http://www.guardian.co.uk/science/video/2012/jul/03/what-is-a-higgs-boson-video—

The TenTen Corpus Family

Miloš Jakubíček[▲▼], Adam Kilgarriff[▲], Vojtěch Kovář[▲▼], Pavel Rychlý[▲▼], Vít Suchomel [▲▼]
[▲] Lexical Computing Ltd.
[▼] Masaryk University
<name>.<surname>@sketchengine.co.uk

1 Introduction

Everyone working on general language would like their corpus to be bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata. In this paper we describe out programme to build ever better corpora along these lines for all of the world's major languages (plus some others).

Baroni and Kilgarriff (2006), Sharoff (2006), Baroni et al (2009), and Kilgarriff et al (2010) present the case for web corpora and programmes in which a number of them have been developed. TenTens are a development from them.

2 Names

Two of the programmes above used the WaC suffix for corpus-naming. To forestall confusion with a name like FrWaC being ambiguous between two different corpora (though both French and web-crawled) a new name was needed. The new batch of corpora are in the order 10^{10} (10 billion) words, so this is the TenTen family. [1] The corpus name is then formed by prefixing with the two-letter ISO-639-1 code for the language, and, optionally, suffixing with two-digits for the year of collection, to give e.g. enTenTen12 for English collected in 2012, zhTenTen for Chinese.

3 Major world languages

We treat the following as major world languages (based on number of speakers and sizes of associated economies): Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish. We have created, and will maintain and develop, TenTen corpora for each of these eleven languages. We have also developed them for several other languages we have particular interests in, currently Czech, Hungarian, Polish and Slovak.

All these corpora are available within the Sketch Engine (Kilgarriff et al 2004).[2]

4 Spiderling, jusText, Onion

The processing chain for creating the corpus is:

- Crawl the web with spiderling[3] (Pomikalek and Suchomel 2012), a crawler designed specifically for preparing linguistic corpora
- Remove non-textual material and boilerplate with jusText (Pomikalek 2011). JusText uses the working definition that we want only 'text in sentences' (and not, e.g. headers and footers). The algorithm is linguistically informed, rejecting material that does not have a high proportion of tokens that are the grammar words of the language, so, in the course of data-cleaning, most material which is not in the desired language is removed.
- De-duplicate with onion (Pomikalek 2011). We de-duplicate at the paragraph level, as, for many linguistic purposes, a sentence is too small a unit, but a whole web page (which may contains large chunks of quoted material) is too large.

These tools are designed for speed and we use them installed in a cluster of servers. For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day. The 12-billion-word enTenTen12 was collected, in 2012, in twelve days.

Then, we want to tokenize the corpus into words, lemmatise, and part-of-speech tag. For these processes we examine the available tools for the language and apply the best we can find (after considering, firstly, accuracy, but also speed, quality of engineering, and licence terms). We have made extensive use of TreeTagger and FreeLing for European languages; Stanford tools for Chinese, meCab (with UniDic lexicon) for Japanese, Han Nanum for Korean, and MADA (in collaboration with Columbia University) for Arabic.

5 Static corpora and monitor corpora

A static corpus is a fixed dataset. A monitor corpus moves on, adding more material over time,

[1] We continue to use the WaC suffix in the 'Corpus Factory' programme, which uses slightly different methods (see Kilgarriff et al. 2010), mainly for languages with fewer speakers and less of a web presence.

[2] http://www.sketchengine.co.uk

[3] http://nlp.fi.muni.cz/trac/spiderling

so it can monitor change in the language (Clear 1986). The advantage of the static corpus is that it is a fixed point that can be referred to in years to come and always means the same thing. The advantage of the monitor corpus is that it stays up to date.

We do not see these two goals as conflicting. Our plan is to re-crawl each language every year or two, and then, after filtering out any paragraphs in the new material that we already had in the old, adding the new to the old, with metadata that allows us to search in, and gather statistics over, 'only the new' or 'only the old'. This also allows us to contrast the new with the old, using Sketch Engine functions such as keywords and sketch-diffs.

6 Virtual corpora

A corpus is a collection of texts. If you add one collection to another, you get a bigger collection. 1+1=1. There are often benefits to treating two corpora of the same language as two parts of a larger whole. We have recently developed technology that implements the intuition, allowing two or more existing corpora, indexed in the Sketch Engine, to be seen as a single corpus from the user's point of view.

Virtual corpora, or super-corpora, have several benefits. They make maintenance of these very large objects easier, as different component corpora can be stored and indexed separately. Also when we add new material, to a very large corpus, we will not need to re-index the whole. They encourage the super-corpus designer to be disciplined in their use of metadata fields, as queries will only make sense if there is a unified system covering the metadata of all component corpora.

7 Fixed corpora: pros and cons

As already noted, many people would like their corpus to be fixed, so that queries and experiments run over it give exactly the same results now and in ten years time. Some argue that such replicability is central to the scientific integrity of the field.

This presents us with a substantial difficulty. We often find problems with our corpora, for example, sets of pages from a spam website. We would like to remove that spam, and the corpus will then be more useful for most users, but those who want replicable results object.

A similar issue arises with NLP tools. If there are better tools, or even just debugged or otherwise improved versions of those we are already using, should we upgrade? For most of our users, we would like to, but those who want replicability will object.

To some extent these problems can be solved by keeping numerous versions. But corpora are large, and management and maintenance is in any case a large task, and there are limits to our willingness to keep multiple versions.

As a policy, our priority is good, up-to-date data and mark-up, and we give higher priority to data quality than to 100% replicability. We think a metaphor from the natural sciences is more apt here than one from computer science. Where biologists replicate an experiment with a new sample of tissue, they do not expect 100% replicability. Replicability will be within margins according to the variability of the material under scrutiny.

8 Metadata

One of the limitations of web-crawled corpora is that they come with very little metadata.

Date of production is one problem: none of the dates on a web page reliably state when it was written – unless it is one of a few types of text such as newspaper, blog, or press release. We are supplementing general crawls (where we have the date of crawling, which is of some use, but little else) with targeted crawls for these text types (see Minocha et al 2013).

Another concern is region. For Spanish, Portuguese and Arabic, we have metadata fields according to the top level domain of the website that the text came from. For English we have trained a classifier to distinguish British and American English, and applied it to all of enTenTen, so we have data-derived metadata.

We have also classified all documents in enTenTen for readability, based on Kilgarriff et al (2008) and plan to do the same for formality, using a method based on Heylighen and Dewaele (1999).

We are exploring domain corpora using both bottom-up methods and targeted crawling (Avinesh et al 2012) so in due course, large parts of the TenTen corpora will have a value for the 'domain' attribute.

9 Conclusion

We have presented a new family of corpora, the TenTens, of the order of 10 billion words. We have described how we are building them, what we have built so far, and how we shall continue maintaining them and keeping them up to date in the years ahead. While, as yet, they have very little metadata, we are working out how to gather and add metadata attribute by attribute. The

corpora are all available for research at http://www.sketchengine.co.uk.

References

Avinesh PVS, D. McCarthy, D. Glennon and J. Pomikálek (2012) Domain Specific Corpora from the Web Proc *EURALEX*. Oslo, Norway.

Baroni, M., and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. *Conference Companion of EACL* 2006.

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. J. *Language Resources and Evaluation* 43 (3): 209-226.

Clear, J. 1986. Trawling the language: Monitor corpora. *Proceedings of Euralex*.

Heylighen, F. and J-M Dewaele. 1999. *Formality of Language: definition, measurement and behavioural determinants*. Internal Report, Free Univ Brussels.

Kilgarriff, A., P. Rychly, P. Smrz, D. Tugwell. 2004. The Sketch Engine. *Proc Euralex*, Lorient, France.

Kilgarriff, A. 2009. Simple Maths for Keywords. *Proc Int Conf on Corpus Linguistics*.

Kilgarriff, A., M Husak K McAdam M Rundell P. Rychly. 2008. GDEX: Automatically Finding Good Dictionary Examples. *Proc. EURALEX*, Barcelona.

Kilgarriff, A., S. Reddy, J. Pomikalek, Avinesh PVS. 2010. A corpus factory for many languages. *LREC*, Malta.

Minocha, A.,. S. Reddy and A. Kilgarriff 2013. Feed Corpus: an ever-growing up-to-date corpus. *8th Web-as-Corpus workshop*, Lancaster, UK.

Pomikalek, J. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masaryk University, Brno, 2011.

Pomikalek, J., and V. Suchomel 2012. Efficient Web Crawling for Large Text Corpora *Proc. 7th Web-as-Corpus workshop*, Lyon, France.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, (eds), *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Imagining the Other: corpus-based explorations into the constructions of otherness in the discourse of tourism

Sylvia Jaworska
Queen Mary, University of London
s.jaworska@qmul.ac.uk

1 Introduction

The figure of the Other has recently attracted considerable attention in linguistic research. The representations of the Jew, the Migrant and the Homosexual, to name just a few, have been examined in detail including historical analyses as well as contemporary perspectives (e.g. Baker 2004; Baker and Gabrielatos 2008; Jaworska 2011; Rash 2012; Wodak 1997). One of the dominant themes revealed in this research is that an encounter with the Other triggers a dichotomy based on a negative other-representation and a positive self-representation. The Other is frequently portrayed as a threat, which is demonised or criminalised. However, most of the studies on the topic to date have been concerned with the representation of the Other in political contexts and there is very little research outside this particular setting. However, as recently suggested by Coupland (2012: 241), representing a group or groups as the Other may not be to inherently marginalise or demonise them. He further adds that it is a challenge for future research to analyse the strategies of Othering in a variety of contexts as each context may bring about different social effects (cf. Koller 2012).

The aim of this study is to respond to this challenge by looking at the ways in which the Other is represented in the discourse of tourism. In our modern times, the tourist experience is a space when regular encounters with the Other take place. This is enhanced by the desire for authenticity, which is equated with the exposure to local people and customs, especially alternative and 'primitive' ways of life (MacCannell 1976). Although the notion of authenticity has been contested in tourism research (e.g. Cohen 2007), a quick look at tourist promotional materials, normally saturated with descriptions of 'real people and life', provides enough evidence for the saliency and endorsement for the notion of authenticity and the figure of the native Other as its central component (White 2007). Recently, in tourism research, there has been some interest in the representation of local people, revealing the prevalence of stereotypic and essentialised images

that depict them as exotic, erotic, primitive and timeless (White 2007; Aitchison 2001). For example, White (2007) highlights that local people in tropical destinations are frequently fetishised by making references to their friendliness and 'inherent' sociability. In her view, the imagery of warmth and hospitability of the 'natives' resonates, on the one hand, with the tourist desire for authenticity and intimacy. On the other hand, such a portrait revives colonialist images of "an expected servility and docility of natives" (ibid.: 35). Hence, it has been argued that tourism is an industry of difference per se perpetuating colonial and gendered discursive practices.

This claim has been contested by researchers who have recently examined the representations of people and places in materials produced by local tourist industries. For example, Bandyopadhyay & Morais (2006) and Amoamo & Thompson (2010) have shown that the self-representations focus more strongly on cultural diversity, hybridity and modernity, and in doing so, challenge the Western other-representations. Contrary to the previous claim, these scholars see tourism as a site of resistance.

While this research provides some evidence for the existence of hegemonic and divergent discourses, its empirical basis is very small. These studies are based on rather selective samples of texts and are not grounded in a systematic linguistic analysis despite the frequent references to linguistic concepts such as language, discourse or metaphor. Equally, with a few exceptions (Baker 2006, Jaworski & Thurlow 2010, Jaworska 2013), linguistics has, to date, paid little attention to tourism despite the sociological recognition of the centrality of language in constructing and ordering tourist experiences (Dann 1996).

2 Aims of the present study

The present study intends to contribute to the under-researched area of tourism discourse in Linguistics by looking at the portrait of local people in a large corpus of promotional tourist materials, including those produced by Western (British) and local tourist industries. The main questions this research addresses are:

- How is the Other represented in a variety of tourist contexts?
- What are the dominant discourses surrounding the representation of the Other?
- Do stereotypical images prevail?
- Are there any differences between the self- and the other-representations?

This study aims primarily at identifying the ways of how local people are referred to and in what thematic contexts they are mentioned. It will do so by using the tools and methods of Corpus Linguistics. The benefit of CL lies in the fact that it can reveal repetitively occurring lexico-grammatical patterns, which, in turn, point to salient representations and the majority ways of viewing the phenomena under investigation (Baker 2006: 14). As Stubbs (2001: 215) highlights: "Repeated patterns show that evaluative meanings are not merely personal or idiosyncratic, but widely shared in a discourse community. A word, phrase or construction may trigger a cultural stereotype." Since cultural stereotypes are precisely the concern of this study of representations, CL proves to be a useful methodological approach for the current investigation.

3 Methodology and corpus data

In order to address the research questions outlined above, two parallel corpora were created. The first consists of texts describing the most popular tourist destinations produced by tourist companies operating in Britain such as Thomson Holidays, Virgin Holidays, Thomas Cook and Kuoni Travel UK. The second includes descriptions of the same destinations but produced by local tourist agencies. Since this research was interested in post-colonial discursive practices, only those destinations that have been previously colonised by the British were included in the corpus. Table 1 summarises the size of each sub-corpus. Both corpora were searched by using the software Sketch Engine™.

Corpus	Tokens
Corpus 1: Other_Represent	83,959
Corpus 2: Self_Represent	85,250
Total	169,209

Table 1: The Data

In the first instance, frequency lists of both corpora were created and carefully examined to identify the most frequent nouns (and pronouns) used to refer to local people. These were then categorised into semantic groups including location, occupation, historical reference, gender, kinship, social organisation. Subsequently, the most frequent descriptors of local people were selected and their collocational profiles investigated.

4 Preliminary results

This study is a part of a larger project which

investigates the constructions of otherness as manifested in the descriptions of places and people in tourism (see also Jaworska 2013). A preliminary analysis revealed considerable differences in the way local people are referred to in both corpora. In the texts produced by Western (British) tourist organisations, the local people are frequently positioned in occupations that are associated with lower status, serving and primitive or traditional ways of live ('butler', 'fisherman', 'hawker'). There was an absence of references pointing to higher status professions or activities associated more commonly with modern life style. Moreover, as the collocational analysis demonstrated, local people were mostly attributed with activities such as 'selling', 'serving' and 'smiling'. In this data set, there were practically no references pointing to cultural, national or religious diversity. There were also no instances of gendered representations. Different patterns emerged in the data consisting of tourist material produced by local tourist organisations. There, local people were also frequently described by making references to occupations. However, they were of higher often intellectual status ('artist', 'poet', 'writer', 'scientist'). Interestingly, in this corpus there were also frequent references to gender, mainly to 'woman/women' who were attributed with 'dance', 'performance' or piece of clothing and a traditional way of live. The findings corroborate, but also challenge some of the results obtained in previous, qualitative research. Firstly, by positioning local people in lower status occupations and attributing them with 'smiling' and 'serving', the other representations re-create colonialist imagery of the local servility and docility (White 2007). Previous research 'accused' the Western tourist industry of reproducing gendered stereotypes when describing local people. This could not be confirmed in the present study. Interestingly, a stronger presence of gendered representations was detected in the corpus of 'self-representations'. Some possible explanations for the absence of gendered imagery in the first data set and its stronger presence in the second will be offered.

References

Aitchison, C. 2001. "Theorizing Other discourses of tourism, gender and culture. Can the subaltern speak (in tourism)?" *Tourist Studies* 1 (2): 133-147.

Amoamo, M. & Thompson, A. 2010. (re)Imaging Maori tourism: Representation and cultural hybridity in postcolonial New Zealand. *Tourist Studies* 10 (1): 35-55.

Baker, P. 2004. "'Unnatural Acts': Discourses of Homosexuality within the House of Lords Debates on Gay Male Law Reform". *Journal of Sociolinguistics* 8 (1): 88-106.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. and Gabrielatos, C. 2008. "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005". *Journal of English Linguistics* 36 (1): 5-38.

Bandyopadhyay, R. and Morais, D. 2005. Representative Dissonance. India's Self and Western Image. *Annals of Tourism Research* 32 (4): 1006-1021.

Cohen, E. 2007. "Authenticity in Tourism Studies: Apreś la Lutte". *Tourism Recreation Research* 32 (2):75-82.

Dann, G. 1996. *The Language of Tourism: A Sociolinguistic Perspective.* Wallingford: CAB International.

Fairclough, N. 1989. *Language and power*. London: Longman.

Jaworska, S. 2011. "Anti-slavic imagery in German radical nationalist discourse". *Patterns of Prejudice* 45 (5): 435-452.

Jaworska. S. Forthcoming. "The quest for the 'local' and 'authentic': Corpus-based explorations into the discursive constructions of tourist destinations in British and German commercial travel advertising." In D. Höhmann. *Tourismuskommunikation. In Spannungsfeld von Sprach- und Kulturkontakt*. Frankfurt a. M.: Peter Lang.

Jaworski, A. and Thurlow, C. 2010. *Tourism Discourse: The Language of Global Mobility*. Basingstoke: Palgrave Macmillan.

Koller, V. 2012. How to analyse collective identity in discourse: Textual and contextual parameters. *Critical Approaches to Discourse Analysis Across Disciplines*, 5 (2): 19-38.

MacCannell, D. 1976. *The Tourist: A New Theory of the Leisure Class*. New York: Schocken Books Inc.

Rash, F. 2012. *German Images of the Self and the Other in German Nationalist, Colonialist and Anti-Semitic Discourse 1871-1918*. Basingstoke: Palgrave Macmillan.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

White, C. 2007. "More authentic than thou: Authenticity and othering in Fiji tourism discourse". *Tourist Studies* 7 (1): 25-49.

Wodak, R. 1997. "Das Ausland and antisemitic discourse: the discursive construction of the Other." In S. H. Riggins (ed.) *The Language and Politics of Exclusion: Others in Discourse*. Thousand Oaks, CA: Sage.

"Hold on a minute; where does it say that?" – Calculating key section headings and other metadata for words and phrases

Stephen Jeaco
Xi'an Jiaotong-Liverpool University
smjeaco@liv.ac.uk

Corpus users are familiar with the concept and usefulness of keyword analysis. It is a feature of concordance software that calculates which words are key in a corpus when compared with a reference corpus (Scott 1997; Scott and Tribble 2006). Despite recent questions about the wording of the definition and whether log-likelihood contingency tables are appropriate (Gabrielatos and Marchi 2012), it is a widely utilized tool and seems to be providing concordance users with information they find meaningful.

However, it may also be fruitful to develop a procedure to look at data from the opposite direction; that is to consider which texts or text sections are key for a word or phrase. Some work in this direction has been done using equally sized strips of text and comparing relative frequencies of words within one section against the others (Liang 2012). The idea of looking at where words tend to occur is also related closely to the well-established concept of dispersion (Oakes 1998). One way of showing the user how words or phrases are spread throughout texts and the corpus is through dispersion plots (Scott and Tribble 2006). Attempts have also been made to search keyword databases for words with a specific pragmatic function in order to see whether their role in the text can be automatically identified (Scott 2000). While dispersion calculations and keywords on strips do provide some insights, corpora often have tags and metadata which could provide much more detail. It seems that currently these metadata are usually only used to filter searches rather than to examine the distribution of specific words and phrases under investigation.

This paper presents a new approach which uses log-likelihood contingency tables with Bayes Factors to create a list of key metadata and section labels (called KeyTags) which are then displayed using a tag cloud. The log-likelihood contingency table which is used to rank and test the significance of the relationships is given in the table below. It can be seen that this contingency table is formed by comparing the number of instances of a word or phrase within a text or section which is mapped to a metadata tag against the number of times the word or phrase occurs outside this context. The log-likelihood formula also balances this against the overall number of other words within the same context. A similar procedure is used to calculate KeyTags for multi-word units, where the frequencies are multiplied by the length in words of the multi-word unit, since each instance of a two word multi-word unit occurring within a metadata tag would account for two words from the total word count for that tag.

	Sub-Corpus 1	*Sub-Corpus 2*	*Total*
Node Word	Node word inside XML node	Node word outside XML node	Frequency of node word
Other Words	Other words inside XML node	Other words outside XML node	Frequency of other words
Total	Word count inside XML node	Word count outside XML node	Whole Corpus

Table 1. KeyTags Contingency Table.

Wilson (In press) recommends the use of Bayes Factors in keyness calculations to distinguish between very strong evidence and less strong evidence based on the overall size of the corpus, and this approach is used in order to standardize the cut-off point for the KeyTag rankings.

In the present work KeyTag data are generated using the BNC (2007) and a corpus of biomedical academic articles from SpringerOpen (2011). These corpora are pre-processed using an application which allows a corpus manager to designate KeyTag processing rules for XML tags before the texts are passed through CLAWS (Garside and Smith 1997) and imported into a relational database. This provides flexibility for KeyTags to include details of the source (e.g. date and publication details) and the author as well as headings or other labels which are embedded in the XML texts. Each complete corpus is pre-processed and summary data are held in the relational database, meaning that users can get instant access to the KeyTag Clouds, without waiting for a minute. The summary data can also be retrieved using the tag as a search item, producing similar results to traditional keyword analysis, but on a much wider range of tags. The overall aim of this new concordance software feature is to provide additional information about the distribution of words and phrases to unsophisticated users of the system. The clouds are to be displayed alongside concordance lines

and other summary data as a means of enriching the contextual clues available. It could also be considered as a possible way to approach the automatic identification of what Hoey (2005) calls pragmatic association primings.

The results provide some indication of typical environments but there are also some issues regarding presentation which need to be considered. For example, a KeyTag cloud of text metadata for *therefore* in the BNC provides (in descending order of keyness) "ACPROSE", "FICTION", "NEWS", "W fict prose", "W ac:polit law edu", "NONAC", "W commerce" , "W ac:soc science" and "OTHERPUB" followed by some specific publishers. As expected, this suggests strongly an association with written texts. The same search for *thus* shows "ACPROSE", "FICTION", "Written Text", "NEWS", "OTHERSP", "Spoken Text", "NONAC", "OTHERPUB", some of the same written domains, and "S meeting" further down the list but still with an approximate BIC value reaching very strong evidence. Depending on the size of the KeyTag cloud, there could be a danger that this last KeyTag would not be visible, but again as expected, the results show a high level of formality for texts containing *thus.*

Looking at text KeyTags for words like *regime* and *pose* gives some information about the kinds of texts in which they occur, but section heading KeyTags can give insights into the actual topics. The word *regime* in the BNC has key section headings such as "The Neutralisation of Afghanistan", "States as third parties to treaties" and other sub-headings showing topics of the politics of governments, historical figures, and religious groups as well as dieting. The word *pose* shows little of interest at the text level in the BNC, but KeyTags for the section level indicate its importance in choreography. For the biomedical academic corpus, a search for *aim* produces section KeyTags including "Background", "Results", "Abstract", "Methods", "Aim" and "Introduction". It might be thought desirable to filter out KeyTags matching the search term itself, because each time a section heading is added to the corpus as text it guarantees increasing its identification with itself, but this issue needs to be considered more fully since users might find it helpful to see that a word they have searched for is often used as a heading. In the same corpus, section KeyTags for *goal* are "Background", "Results" and "Methods", with "Abstract" rather lower down the list. For this node, "Introduction" is below the strong evidence threshold and "Aim" is absent. Turning to another fairly predictable academic term, the word *significant* in the SpringerOpen corpus has section KeyTags of "Results", "Methods", "Background", "Statistical Analysis", "Authors' contributions", "Statistics", "Implementation" and "Statistical analyses", perhaps hinting at its dual use in academic texts: related to statistics and previous research.

As well as text and section information, KeyTags can also include details about the author or speaker. Some mixed results are produced for KeyTags in the BNC for *gosh* and *sorry*. For *gosh*, "sex:f" is in third position, suggesting females in the corpus use this relatively more frequently. Several other speaker description tags appear but, on close examination, other than a code for British English, they all mean the data were unknown or not available ("soc:UU", "role:unspecified", "dialect:NONE" and "ageGroup:X"). It would be important to consider whether these tags should be excluded or whether information about the way in which the original data samples were gathered would need to be presented. The results for *sorry* show a similar lack of information for the top five tags, but also have "sex:M", "firstLang:EN-GBR", age group codes, "sex:F", an occupation and then "age:50+" and another occupation. The age group codes correspond to "35 to 44 years of age" and "45-59 years of age", demonstrating a slight tendency for this word to be associated with more mature speakers. Showing both genders within the same cloud may be confusing, but is logically possible since only the spoken texts contain these tags.

As well as providing new kinds of data for corpus users, this approach also tries to bridge the gap between the sophisticated mark-up of modern XML corpora and visual presentation of KeyTags which might aid users in interpreting typical contexts for search terms. It is argued that the procedure can make a useful contribution to the range of summary information available, but further consideration needs to be made as to how to help users interpret the significance of the KeyTags, and also how they should interpret "thin" or "empty" clouds.

References

BNC (2007). *The British National Corpus*, Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Gabrielatos, C. and A. Marchi (2012). "Keyness: Appropriate metrics and practical issues.". CADS International Conference 2012. University of Bologna, Italy.

Garside, R. and N. Smith (1997). "A hybrid grammatical tagger: CLAWS4". In R. Garside, G.

Leech and A. McEnery Corpus annotation: Linguistic information from computer text corpora. London: Longman: 102-121.

Hoey, M. (2005). Lexical priming : a new theory of words and language. London : Routledge.

Liang, M. (2012). "Patterned Distribution of Phraseologies within Text: the Case of Academic English". Corpus Technologies and Applied Linguistics. Xi'an Jiaotong-Liverpool University, Suzhou, China.

Oakes, M. P. (1998). Statistics for corpus linguistics. Edinburgh : Edinburgh University Press.

Scott, M. (1997). "PC analysis of key words -- And key key words." System 25(2): 233-245.

Scott, M. (2000). "Mapping Key Words to Problem and Solution". In M. Scott and G. Thompson Patterns of Text in Honour of Michael Hoey. Amsterdam, Netherlands: Benjamins: 109-127.

Scott, M. and C. Tribble (2006). Textual patterns : key words and corpus analysis in language education. Amsterdam: J. Benjamins.

SpringerOpen. (2011). "SpringerOpen's open access full-text corpus for text mining research." Retrieved 6 July, 2011, from http://www.springeropen.com/about/datamining/.

Wilson, A. (In press). "Embracing Bayes Factors for Key Item Analysis in Corpus Linguistics". In New approaches to the study of linguistic variability. Frankfurt/Main: Peter Lang.

Rape, madness, and quoted speech in specialized 18th and 19th century Old Bailey trial corpora

Alison Johnson
University of Leeds
a.j.johnson@leeds.ac.uk

1 Introduction

At 120 million words the *Old Bailey Proceedings Online* (Hitchcock et al. 2012) (hereafter *OBP*), which contains 197,000 trials, is a seemingly homogenous large corpus. But for discourse analysts, as Hunston (2011) points out, "a corpus, unlike a text, cannot be analysed". This paper suggests a solution to this problem, with both pedagogical and research relevance: the extraction of smaller corpora from the larger whole. The paper uses two sub-corpora of the *OBP* to investigate professional trial discourse in history. It outlines the methods, design, and rationale of the extraction process required to obtain what Cameron and Deignan (2003) call "small corpora" and Flowerdew (2004) "specialized corpora". The first is a corpus of 18th century rape trials and the second a corpus of 19th century trials where an insanity defence is offered.

The particular research focus is the use of quoting (direct, indirect and simulated verbatim speech), particularly of the defendant, looking at who quotes and who is quoted, where it is in the trial, and when it occurred in relation to the offence and its alleged commission.

2 The corpora – scope and design

Given the scope of the *OBP* – it contains trials relating to many different offences from murder and treason to pickpocketing, and these range over two and a half centuries – it would be unreasonable not to consider contextual, sociolinguistic, and field-specific questions in the compiling of specialized corpora from within it. The corpus consists of "spoken interaction (re)cast as writing" (Culpeper and Kytö 2000), but the methods, skills, and scribes vary across the 17th, 18th and 19th centuries and across individual trial records. By the 19th century, for example, Pitman shorthand had been invented and the adversarial trial resembled modern trial discourse more closely. Defence lawyers were present in the Crown Court (Archer 2010) and advocacy, rather than individuals prosecuting their own cases, was the norm. Individual trials contain structural parts – prosecution and defence cases – and both

professional and lay voices: barristers, medical and lay witnesses. And there are considerable spelling variations to negotiate. Corpus design, based on these and other variables is outlined in the first part of the paper, giving examples of decisions made in relation to the two corpora.

The corpus of 18th century rape trials contains 111 trials of between 300 and 14,000 words each, with an average of around 2,000 words (and a total of around 340,000 words). These are all the trials of substance in the period where such trials were not censored. The corpus of 19th century trials, the Monomania Corpus, is a deliberately biased corpus: it is created by using the uniquely Victorian word 'monomania' to denote madness, and a search for that word finds 15 trials of between 300 and 164,000 words (a total corpus of around 690,000 words).

3 Quotation

Bell (1991: 207-209) outlines the functions of direct quotation in news discourse: producing "incontrovertible facts", "distancing and disowning", and producing "soundbites". And Clift and Holt (2007: 6) point out the "dramaturgical quality of DRS [direct reported speech]". These are all important to consider, alonside the Labovian insight of the evaluative aspect of quotation in storytelling. Quotation produces what I describe and show as evaluative surplus in the simultaneous "replay" of an interaction with the opportunity for the quoter to "convey his or her attitude towards the reported utterance" (Clift and Holt 2007: 6). This interaction provides an occasion for the jury to internally accept or reject the action, as they weigh the evidence. Quotation can be used differently by the prosecution and defence, to demonstrate the defendant's criminal liability, or his incapacity to commit a criminal offence, respectively. The establishment of a binding legal reality crucially hinges on the differential weights of prosecution and defence evidence. Embedding reported speech is a highly selective and thus powerful resource of institutional meaning-making (legal and medical) and reproduction.

4 Quoting in rape trials and insanity defences

The second part of the paper presents the methodological implications of working with these corpora in relation to the focus on direct quotation, which is found using the searchwords: 'say', 'says', 'said' and quotation marks as a punctuation feature. In the Rape Corpus, using subdivided files (adult cases and child victims) we see different results for child victims and adults, and also see the value of treating summary third person trials and first person 'verbatim' records differently. In the Monomania Corpus, I show the value of news reports alongside the *OBP*, with these containing additional witness evidence, opening speeches, judges' monologues. The additional and overlapping trial discourse in the news reports highlights the relative incompleteness of the *OBP*. Each of the 15 Monomania Corpus trials is a text with component parts (prosecution case; defence case; witness appearances) and only partial, as a comparison with trial proceedings reported in 19th century newspapers shows. In further illustrations from the Monomania trials of Cooper (1842) and McNaughten (1843), I show how the method of dividing the trials up into their structural parts (prosecution and defence cases) in additional files, allows us, through corpus techniques using *Wordsmith* (Scott 2011) and *CFL LFM* (Woolls 2012), to see very different pictures of the defendant. The picture that is painted through quoting his words in the prosecution case, compared with quoting the defendant's words in the defence case, produces different worlds and different identities: one criminal and culpable and the other insane and unable to distinguish between right and wrong.

Since legal discourse is conservative (rather than varying widely in lexis and structure over time) and a clear example of Reisigl and Wodak's (2009) discourse-historical approach, which suggests that discursive practices are embedded in history, we can also test to what extent current defence advocacy and attitudes towards the police, expert witnesses and defendants are evident in 19th century trials. In the case of Cooper, the prosecution quotation is damning and incriminating and the witnesses seem well 'rehearsed' in supporting the prosecution case, whereas the defence picture, drawn out through family members, presents the working-class misery of the time and a young man seen as easy prey by a constabulary looking for results. In McNaughten's case, we see the two institutions of the law and medicine fighting over culpability and insanity, as the defence quotes from medical interviews with the accused.

5 Conclusions

The paper offers a number of conclusions relating to methodological and analytical concerns. A large and complex corpus of specialized professional discourse, such as the *OBP*, requires informed and discipline-specific thinking in order to extract and develop sub-corpora of use to

discourse analysts. There are almost unlimited possibilities for corpus extraction from this corpus and this paper details just two research possibilities, as illustrated with the two corpora extracted by the author and discussed here. In addition to the methodological detail reported here and the *modus operandi* for doing socio-legal discourse studies within the *OBP* that is illustrated, there are both pedagogic and research benefits revealed. Through extraction of small and specialized corpora from the larger whole, the linguistic researcher is able to produce new socio-legal and socio-historical linguistic knowledge in the learning and teaching domain and in terms of academic research into particular phenomena. The quotation phenomenon can be seen in the context of practices of text-production (Fairclough 2001: 20), and, in doing so, we can see that the record selects this vivid representation mode as a particular act. Making the record include verbatim speech gives us unique access to the words and worlds of speakers in trials across time. We see that: what, who, and where words are quoted are matters for record and the result is the production of evidentially powerful testimony for the prosecution and defence and for the jury's decision-making.

References

Archer, D. 2010. "The historical courtroom. A diachronic investigation of English courtroom practice". In Coulthard, M. and Johnson, A. (eds) *The Routledge Handbook of Forensic Linguistics.* Abingdon: Routledge, 185-198.

Bell, A. 1991. *The Language of News Media.* Oxford: Blackwell.

Cameron, L. and Deignan, A. 2003. "Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse". *Metaphor and Symbol* 18 (3): 149-160.

Clift, R. and Holt, E. 2007. "Introduction". In Holt, E. and Clift, R. (eds) *Reported Talk. Reported Speech in Interaction.* Cambridge: Cambridge University Press, 1-15.

Culpeper, J. and Kytö, M. 2000. "Data in historical pragmatics: spoken interaction (re)cast as writing". *Journal of Historical Pragmatics* 1 (2): 175-199.

Flowerdew, L. 2004. "The argument for using English specialized corpora to understand academic and professional language". In Connor, U. and Upton, T. A. (eds) *Discourse in the Professions.* Amsterdam: John Benjamins, 11-33.

Fairclough, N. 2001. *Language and Power* (2nd Edn.). Harlow: Pearson Education Limited.

Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S. and McLaughlin, J. 2012. *The Old Bailey Proceedings Online, 1674-191.* www.oldbaileyonline.org, version 7.0, 24 March 2012 [accessed 14 January 2013].

Hunston, S. 2011. "Doing analysis in discourse and corpus: the case of evaluative language". Plenary paper presented at Corpus Linguistics 2011. Birmingham, 20-22 July 2011.

Reisigl, M. and Wodak, R. 2009. "The discourse-historical approach". In Wodak, R. and Meyer, M. (eds) *Methods of Critical Discourse Analysis* (2nd edn) London: Sage, 87-121.

Scott, M. 2011. *Wordsmith Tools.*

Woolls, D. 2012. *CFL Lexical Feature Marker.*

Family in the UK – risks, threats and dangers: a modern diachronic corpus-assisted study across two genres

Jane Helen Johnson
University of Bologna, Italy
janehelen.johnson@unibo.it

The institution of 'the family' in Britain is undergoing considerable change, involving "the formation and dissolution of families and households, and the evolving expectations within individuals' personal relationships (Giddens 2001: 178). Changes may of course be positive but the family may be affected both externally and internally by factors which may be described as 'risk', perhaps determined by fundamental changes both in and to society. Particularly in the last few decades indeed the subject of risk has become increasingly significant and of topical interest. (following Beck 1992).

Though much research has focussed on risk within a sociological framework, not much cross-disciplinary work has been done which brings together risk sociology and linguistics (Zinn 2010). One notable exception is Hamilton et al (2007), who use a corpus of spoken discourse (CANCODE) to investigate the semantics, prosody and phraseology of risk in spoken British English. The special issue of Critical Approaches to Discourse Analysis across Disciplines (2010: 4:2) also features research which makes use of corpus linguistics to examine risk (cf. Marko 2010; Grundmann and Krishnamurthy 2010; and Sandor 2010). Even more recent work includes Hardy and Colombini (2011) who use the Corpus of Contemporary American English (CoCA) to investigate the semantic prosody of risk across different genres in US English.

The focus of this paper is instead to investigate risk in two specific genres and with reference to the particular reality of the family, bearing in mind that "risk is not inherent in any situation [....] but is created through discourses" (Pickard 2009: 69), since "discourse and society shape each other" (Berger and Luckmann 1966; Fairclough 1992: 9). Connections will thus be investigated between perception and expression of risk to the family, both in news discourse and in academic discourse focussing specifically on sociological issues. Reference will be made to sociological issues such as how the family is affected by 'disembedding' from traditional commitments and support relationships (cf. Beck 1992), as well as the 'disembedding' of social relations (cf. Giddens 2001) (Charles et al 2008).

In order to do this, the study uses an MD-CADS approach (see Partington 2010) to investigate the semantic field of risk represented linguistically in relation to family through the lemmas RISK, THREAT and DANGER across two different genres. More specifically, we consider:

- the genre of **news** (in a purpose-built corpus of British newspaper texts from the Guardian and the Daily Mail, 'about' the family and selected on the basis of the presence of the word *family/families*;
- the **academic** genre of **sociology** articles, comprising: (i) all the articles from the online version of Sociology, the journal of the British Sociological Association, between 2008 and 2012, consisting of 2 million tokens; (ii) a subset of the aforesaid articles 'about' the family, selected on the basis of the presence of the word *family/families* in the article abstract.

A diachronic element is provided by the collection of news articles from two separate time periods more than a decade apart, and more specifically:

- the 1993 corpus: featuring 500 articles from the 1993 editions of the Guardian and 500 articles from the 1993 editions of the Daily Mail (total tokens 654,000);
- the 2005+ corpus: featuring 500 articles from the Guardian published between 2005-2011 and 500 articles from the Daily Mail published between 2005-2011 (total tokens 532,000).

A quantitative investigation of linguistic patterns around the lemmas RISK, THREAT and DANGER is made, focussing on concordances, collocates, frequency lists and keyword lists based on both single words and word clusters, followed by a qualitative analysis of particular stretches of discourse to build a comparative picture of how risk, threat and danger to the family is described both across genres and diachronically. Though previous research on a smaller scale (Johnson, 2011) suggested that political orientation would certainly condition what newspapers perceive as threat to the family (in the case of the Daily Mail between 2005 and 2011, for example, gay marriage, abortion, IVF and lesbians wanting to bear children not surprisingly all figured among threats to the traditional family), this study extends findings diachronically, across genres, and in greater linguistic depth.

Specific research issues concern: the

phraseology of wordforms of RISK and other near synonyms such as DANGER and THREAT in a family context; focus on the agents involved in causing the risk; and evaluation (Hunston 2004, 2011; Hunston and Thompson 2000) of the risks, danger and threats to the family. Investigation of these issues will help to build up a snapshot of risk to the family across genres and in different time periods.

References

Beck, U. 1992. *Risk Society*. London: Sage.

Berger, P.L. and Luckmann, T. 1966. *The Social Construction of Reality*. Garden City NY: Anchor Books.

Charles, N., Davies, C. and Harris, C. 2008. *Families in Transition: Social Change, Family Formation and Kin Relationships*. Bristol: Policy Press.

Fairclough, N. 1992. Introduction. In N. Fairclough (ed) *Critical Language Awareness*. London and New York: Longman.

Giddens, A. 2001. *Sociology*. 4th ed. Cambridge: Polity Press.

Grundmann, R. and Krishnamurthy, R. 2010. "The discourse of climate change. A corpus-based approach". *Critical Approaches to Discourse Analysis across Disciplines* 4(2): 125-146.

Hamilton, C., Adolphs, S. and Nerlich, B. 2007. "The meanings of 'risk': A view from corpus linguistics". *Discourse and Society* 18 (2), 163–181.

Hardy, D. E. and Colombini, C.B. 2011. "A genre, collocational, and constructional analysis of RISK*". *International Journal of Corpus Linguistics* 16(4): 462–485.

Hunston, S. 2004. "Counting the uncountable: problems of identifying evaluation in a text and in a corpus". In A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse*. Bern: Peter Lang, 157-188.

Hunston, S. 2011. *Corpus approaches to evaluation. Phraseology and Evaluative language*. London: Routledge.

Hunston, S. and Thompson, G. (eds). 2000. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: OUP.

Johnson, J.H. 2011. "Negotiating differences in the linguistic representation of the family in sociology texts: a corpus-assisted study". Talk presented at the 38th International Systemic Functional Linguistics Conference, 25-29th July 2011, Faculty of Letters, University of Lisbon, Portugal.

Marko, G. 2010. "Heart disease and cancer, diet and exercise, vitamins and minerals: The construction of lifestyle risks in popular health discourse". *Critical Approaches to Discourse Analysis across Disciplines* 4(2): 147-170.

Partington, A. 2010. "Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: an overview of the project". In A. Partington (ed). *Modern Diachronic Corpus Assisted Discourse Studies on UK newspapers*. Edinburgh: Edinburgh University Press, 83-108.

Pickard, S. 2009. "Governing Old Age: The 'Case Managed' Older Person". *Sociology* 43(1): 67–84.

Sandor, A. 2010. "Automatic detection of discourse indicating emerging risk". *Critical Approaches to Discourse Analysis across Disciplines* 4(2): 171-179.

Zinn, J.O. 2010. "Risk as discourse: interdisciplinary perspectives". *Critical approaches to discourse analysis across disciplines* 4(2): 106-124.

Reader comments on online news articles: a corpus-based analysis

Andrew Kehoe
Birmingham City University
andrew.kehoe@bcu.ac.uk

Matt Gee
Birmingham City University
matt.gee@bcu.ac.uk

1 Overview

In recent years, there has been a growth in so called 'Web 2.0' technologies. Whereas the first generation of websites were information sources designed to be read passively, second generation websites allow users to interact with online content and participate in its creation. Web 2.0, or the social web, encompasses wiki-based sites which allow users to collaborate on production of texts (Wikipedia being most famous example), and popular social networking sites such as Facebook and Twitter. As we have explored elsewhere, Web 2.0 also includes blogs (Kehoe & Gee 2012) and 'folksonomies': sites which allow users to assign single word keywords ('tags') to videos, photographs or general web content (Kehoe & Gee 2011).

The growth of these more interactive websites has been well documented since the 'first glimmerings of Web 2.0' were noted at the end of the 1990s (Di Nucci, 1999: 32). From a linguistic perspective, the part of the social web that has received the most attention has been the blog, following early work by Herring et al (2005). Another Web 2.0 site that is receiving an increasing amount of attention is Twitter, with a conference dedicated to the language of twitter hosted at Lancaster University in early 2013, for example.[1]

An area that has received less attention is the influence of Web 2.0 on more established online formats, in particular its influence on news sites. In this paper, we examine the introduction of a blog-like commenting feature on the website of the UK newspaper *The Guardian*.[2] This feature allows readers to comment on some (but not all) articles on the *Guardian* website and take part in discussions with other readers. In the paper, we describe how we have extracted comments from the *Guardian* website and added these to a pre-existing corpus of *Guardian* articles. We present a linguistic case study, using a keywords-based approach to determine the topic of reader comments and to examine, from a web-indexing perspective, the relationship between the comments and the article with which they are associated. Through this analysis, we are able to determine which topics *The Guardian* does not permit its reader to comment upon at all and, amongst the remaining topics, which generate the most comments and the most heated debate.

2 Detailed analysis

We took as a starting point our existing corpus of articles from *The Guardian and The Independent*, searchable through the WebCorp Linguist's Search Engine. This corpus contains one billion words covering the period 1984-present, and has been used in a wide range of linguistic studies, including work on neologisms (Renouf forthcoming), diachronic trends (Kehoe & Gee 2009), and collocation/repulsion (Renouf 1996; Renouf & Banerjee 2009). The *Guardian* section of the corpus covers the period 2000-present. We had previously focused on the main body of online news article, ignoring the comment section at the bottom, but we have recently discovered that the commenting option was first enabled on the *Guardian* website in March 2006. Hermida & Thurman (2008: 6) report that five other UK newspaper sites were allowing reader comments on articles by November 2006.

Although we did not previously include reader comments in our newspaper corpus, we have now re-accessed all *Guardian* articles published since March 2006 and downloaded the associated comments, being careful to preserve the date and author of each comment. Of the 528,729 articles published since March 2006, 116,880 (22%) have at least one comment associated with them. We have also looked at the distribution of comments across those articles which do include at least one comment, comparing this with our previous findings on blog comments (Kehoe & Gee 2012). We have found the average number of comments per article in *The Guardian* to be 26 – more than double the average in blogs – and the average comment length to be 93 tokens – more than three times as long as the average blog comment. The latter is even more significant given that stricter moderation procedures on the Guardian website are likely to eliminate the long spam comments responsible for increasing the average comment length in our blog data (Kehoe & Gee 2012).

We, thus, argue that comments on newspaper articles are a rich source of information which cannot be ignored in corpus compilation. We illustrate this with a linguistic case study based around the problem of indexing documents in a

[1] http://www.lancs.ac.uk/fass/events/twitter_and_microblogging/
[2] http://www.guardian.co.uk/

web-scale collection. Our approach extracts supplementary information about the topic of a document from the comments associated with that document. We have applied this technique successfully to blog data in the past and apply it to articles in our newspaper corpus for the first time here. We use the log-likelihood statistic to compare the set of comments associated with an article (our sample) with the complete set of all *Guardian* comments (our reference corpus). The aim is to identify topic-related keywords which appear in the comments associated with an article, particularly those keywords which do not appear in the article itself. We thus illustrate how comments can be used to aid the indexing of data on news websites by refining the model of textual aboutness extractable from the article alone.

In the second section of the paper we look at the distribution of comments by article in more depth. Taking the aboutness profiles produced in the first phase, we explore the relationship between the topic of an article topic and the number of comments associated with it. The FAQ document on the *Guardian* website states that comments are not permitted at all on 'stories about particularly divisive or emotional issues'[1] but does not specify exactly what topics are included in this ban.

Our analysis in this paper reveals that articles relating to politics, race, religion and ongoing legal cases are most likely to face an outright ban on comments. However, our in-depth analysis of the 116,880 articles which do contain at least one comment reveals that there are nevertheless a large number of 'allowed' topics which tend to generate longer and more heated debates whenever articles on those topics are published. We suggest that our newly expanded corpus of news articles and reader reactions to these articles could also be a useful source of data for work on discourse analysis in general and on impoliteness and conflict in particular. We present a preliminary analysis of the lexis of conflict in newspaper comments, drawing upon our previous work on blog data. We also illustrate how the assignment of topic labels to individual entries in the thread of comments associated with an article could be used to refine automatic spam detection and moderation procedures.

As newspaper data have been a staple of many corpora and linguistic studies in the past, the shift in the delivery and reception of newspaper content described in this paper has important implications for future work.

[1] http://www.guardian.co.uk/community-faqs

References

DiNucci, D. (1999) 'Fragmented Future'. Print 53(4): 32. http://darcyd.com/fragmented_future.pdf

Hermida, A. & N. Thurman (2008) 'A clash of cultures: the integration of user-generated content within professional journalistic frameworks at British newspaper websites'. *Journalism Practice* 2(3), 343-356: http://opendepot.org/147/1/hermida_thurman_JP_2_3.pdf

Herring, S., L.A. Scheidt, E. Wright & S. Bonus (2005) 'Weblogs as a bridging genre'. *Information Technology & People*, 18(2), 142-171: http://ella.slis.indiana.edu/~herring/itp.pdf

Kehoe, A. & M. Gee (2012) 'Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus' in S. Oksefjell Ebeling, J. Ebeling & H. Hasselgård (eds.) *Proceedings of ICAME32* (provisional title). e-journal, VARIENG: http://www.helsinki.fi/varieng/journal/volumes/12/kehoe_gee/

Kehoe, A. & M. Gee (2011) 'Social Tagging: A new perspective on textual "aboutness"' P. Rayson, S. Hoffmann & G. Leech (eds.) *Studies in Variation, Contacts and Change in English Volume 6: Methodological and Historical Dimensions of Corpus Linguistics*, e-journal, VARIENG: http://www.helsinki.fi/varieng/journal/volumes/06/kehoe_gee

Kehoe, A. & M. Gee (2009) 'Weaving Web data into a diachronic corpus patchwork' in A. Renouf & A. Kehoe (eds.) *Corpus Linguistics: Refinements & Reassessments*, Amsterdam: Rodopi, pp. 255-279.

Renouf, A. (forthcoming) 'A Finer Definition of Neology in English: the life-cycle of a word', in H. Hasselgård, S. Oksefjell Ebeling & J/ Ebeling (eds.). *Corpus Perspectives on Patterns of Lexis*. Amsterdam/ Philadelphia: John Benjamins.

Renouf, A. & J. Banerjee (2009) 'The Phenomenon of Repulsion in Text' in *Special edition of Proceedings of 25th International Conference on Lexis and grammar*, Palermo, Sicily, Sept. 6-10, 2006', in C. Leclère et al (eds.) *Lingvisticare Investigationes*. Amsterdam: John Benjamins.

Renouf, A. (1996) 'The ACRONYM Project: Discovering the Textual Thesaurus', in I. Lancashire, C. Meyer & C. Percy *(eds.) Papers from English Language Research on Computerized Corpora (ICAME 16)*, Amsterdam: Rodopi, .pp. 171-187.

Collocation analysis and marketized university recruitment discourse

Baramee Kheovichai
University of Birmingham
Kiao_ra@yahoo.com

1 Introduction

Recently, universities in the UK have been subject to marketization policies, resulting in structural and cultural changes (Tomlinson 2005). Marketization results in commodification of knowledge, customerization of students and the deprofessionalization of academics (Naidoo & Jamieson 2005). Critical discourse analysis has investigated marketization of UK higher education institutions through the lens of discourse analysis and one of the major claims is that the order of discourse of UK universities is being transformed along the line of business organizations in the sense that promotion is increasingly more salient in the discursive practice of UK universities (Fairclough, 1993). This change has been claimed to be a colonization of universities' orders of discourse by businesses (ibid.). Mautner (2010) refers to this phenomenon as "discursive alignment", where universities adapt their discursive practice to be more similar to business organizations. She claims that, considering this change from the speech accommodation theory, the powerless have to align their discursive practice with the powerful. By aligning their discursive practice with business organizations, universities reinforce their powerlessness.

This paper furthers the investigations of marketization of higher education through critical discourse analysis and corpus linguistics. This paper reports on a corpus-based comparison of job advertisements produced by UK universities and business organizations. Data consists of 3,000 job advertisements from www.jobs.ac.uk and 2,000 job advertisements from www.efinancialcareer.co.uk. This paper uses collocation analysis to investigate how universities discursively construct their identity along the line of business organizations. The focus of this paper is on the evaluative adjectives which collocate with organizational reference terms in each corpus. There are two rationales for investigating evaluative adjectives. First, according to Hunston (1994), the language of evaluation can cast light on the value system within a social group. Furthermore, as noted by Hunston (2011) adjectives are commonly used for evaluation and thus evaluative adjectives are strong candidates for research into the promotional discourse of job advertisements. The investigation into evaluative adjectives can therefore show whether universities share the same values as business corporations, which has an implication for the identity of academic institutions. Second, Bhatia (2005) notes that positive description of a product is one of the key strategies in promotional discourse. The investigation into evaluative language adjectives can show the extent to which promotional discourse permeates the orders of discourse of universities.This study will show the saliency of promotional discourse and the similarities and differences between the underlying value systems that universities and business organizations use to construct organizational image.

2 Procedures

The academic job corpus and the business job corpus were annotated with a part-of-speech tagger in Wmatrix (Rayson, 2008). The annotated corpora were taken out of Wmatrix and uploaded into Antconc (Anthony, 2004). The second step involves generating search terms that refer to organizations in each corpus. To achieve this, I used Wmatrix and investigated the semantic tags that can potentially lead to terms referring to organizations in each corpus and I read sampled texts from each corpus to identify these terms as well. Once I compiled all the words, I used them as search terms in Antconc and generated collocates. The collocates were exported into excel and the other parts of speech apart from adjectives were filtered out. Only adjectives that express positive or negative attitudes towards organizations were included. The adjectives that classify the organizations such as 'medical school; were excluded. Concordances were generated to ensure that the adjectives are used for evaluative meaning.

Once the list of evaluative adjectives in each corpus was finalized, I classified them according to the semantic groups that they form. Existing frameworks for evaluative language such as Martin & White's (2005) and Bednarek's (2008) are too broad for the study of marketized discourse and thus they are not used for classification. I arrived at the categories by looking at the meaning of these adjectives in a dictionary, thesaurus and reference corpus (BNC). I also read concordances to determine the meanings of these evaluative adjectives in the context of organizational description. In due process, the grouping emerges from adjectives

that share similar meaning.

3 Data analysis

There are seventeen semantic categories or evaluative resources used for constructing organizational image. These categories as well as examples of evaluative adjectives are shown in Table 1.

1. SIZE e.g. large, biggest, sizable
2. GROWTH e.g. thriving, growing, expanding
3. DYNAMISM e.g. dynamic, vibrant, active
4. REPUTATION/ACHIEVEMENTS e.g. successful, recognized, renowned
5. RANK e.g. world-class, top, top-tier
6. COMPETITIVENESS e.g. leading, competitive, world-leading
7. GLOBAL REACH e.g. global, international
8. INNOVATIVENESS e.g. innovative, new, modern
9. ESTABLISHED e.g. established, integrated, traditional
10. WEALTH e.g. well-resourced, profitable, lucrative
11. VISION e.g. research-led, entrepreneurial, ambitious
12. EMOTIONAL APPEAL e.g. exciting, impressive, attractive
13. CARING e.g. collegiate, supportive, helpful
14. UNIQUENESS e.g. unique, distinctive, specialized
15. INCLUSIVITY e.g. inclusive, affirmative
16. GOODNESS e.g. excellent, good, best
17 OTHERS e.g. purpose-built, well-developed

Table 1. Semantic categories of the evaluative adjectives

A statistical comparison of each category across the two corpora was made. Fisher exact test was computed to determine the differences and effect size was used to identify which categories are more strongly associated with either corpus. The categories that are strongly associated with academic job corpus and significantly different from business job corpus are: 1) CARING, 2) DYNAMISM, 3) OTHERS, 4) VISION, 5) UNIQUENESS, 6) EMOTIONAL APPEAL, 7) GROWTH and 8) INNOVATIVENESS. The categories that are more strongly associated with the business job corpus and significantly different from the academic job corpus are: 1) GLOBAL REACH, 2) RANK, 3) COMPETITIVENESS and 4) GOODNESS. The categories that are not statistically significantly different are: 1) WEALTH, 2) ESTABLISHED, 3) SIZE and 4) REPUTATION/ACHIEVEMENTS. The first two are proportionately more frequent in
the academic job corpus while the latter two are proportionately more frequent in the business job corpus.

Furthermore, I ranked each category within each corpus according to the frequency and this is shown in Table 2.

After that, Spearman's Rank Correlation was computed to determine the correlation between the ordering of the semantic categories. The results indicate a strong positive correlation between the two corpora ($r = 0.787$, $n = 17$) which is also statistically significant ($p < .01$).

Academic	**Business**
GLOBAL REACH	COMPETITIVENESS
DYNAMISM	GLOBAL REACH
COMPETITIVENESS	REPUTATION/ACHIEVEMENT
REPUTATION/ACHIEVEMENT	RANK
SIZE	SIZE
INNOVATIVENESS	GOODNESS
RANK	INNOVATIVENESS
VISION	GROWTH
GROWTH	ESTABLISHED
ESTABLISHED	DYNAMISM
EMOTIONAL APPEAL	VISION
UNIQUENESS	EMOTIONAL APPEAL
CARING	UNIQUENESS
GOODNESS	WEALTH
INCLUSIVITY	CARING
OTHERS	OTHERS
WEALTH	INCLUSIVITY

Table 2. Semantic categories ranked by frequency

This statistical result suggests evidence that academic institutions and business organizations put comparatively similar weight on the values associated with employer organizations.

In addition, there is another observation that can be made from the ranking of the semantic categories in Table 2. In the academic job corpus, it seems that the first ten categories are dominated by the values related to competition and measurement (GLOBAL REACH, COMPETITIVENESS, REPUTATION/ACHIEVEMENT, SIZE, RANK and GROWTH). Moreover, in the first five categories in the academic job corpus, four constitutes the semantic categories more strongly associated with business organizations (GLOBAL REACH, COMPETITIVENESS, REPUTATION/ACHIEVEMENT and SIZE). This is an indication that universities often employ evaluative resources that business organizations use to construct organizational image, thereby showing the similarities between universities and business organizations.

However, within these semantic categories, there are also evaluative resources associated with academic institutions which represent social-oriented values (INCLUSIVITY, CARING), intellectual excitement (DYNAMISM, INNOVATIVENESS) and having a sense of purpose (VISION, UNIQUENESS). These evaluative resources represent a possibility for universities to develop a sense of identity that diverges from business organizations.

References

Anthony, L. (2004). AntConc : A Learner and Classroom Friendly , Multi-Platform Corpus Analysis Toolkit. *IWLeL 2004: An Interactive Workshop on Language e-Learning* (pp. 7–13).

Bednarek, M. (2008). "An increasingly familiar tragedy"
 Evaluative collocation and conflation. *Functions of Language*, *15*(1), 7–34.

Bhatia, V. K. (2005). Generic patterns in promotional discourse. In T. Virtanen & H. Halmari (Eds.), *Persuasion across genres: A linguistic approach* (pp. 213–225). Amsterdam/Philadelphia: John Benjamins Publishing Co.

Fairclough, N. (1993). Critical discourse analysis and the marketization of public discourse: the universities. *Discourse & Society*, *4*(2), 133–168.

Hunston, S. (2011). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. Taylor & Francis Group. Hunston, Susan. (1994). Evaluation and organization in a sample of written academic discourse. *Advances in written text analysis*, 191–218.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation*. Palgrave Macmillan: Great Britain.

Mautner, G. (2010). *Language and the market society: critical reflections on discourse and dominance*. New York: Routledge, Taylor & Francis Group.

Naidoo , R., & Jamieson, I. (2005). Empowering participants or corroding learning? Towards a research agenda on the impact of student consumerism in higher education. *Journal of Education Policy*, *20*(3), 267–281. doi:10.1080/02680930500108585

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549.

Tomlinson, S. (2005). *Education in a post-welfare society*. Buckingham and Philadelphia: Open University Press.

Genre in a frequency dictionary

Adam Kilgarriff
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Carole Tiberius
INL
Leiden, NL
carole.tiberius@inl.nl

1 Routledge Frequency Dictionaries

The Routledge Frequency Dictionary series is now well-established with titles available for eight languages. They give the five thousand commonest words for their language, with indexes for access alphabetically or by frequency, and additional features varying from dictionary to dictionary including English translations, example sentences and their translations, listings by word class, collocations, and tables for the vocabulary of various domains. We are currently preparing a Dutch title to add to the series.

2 Genre in dictionaries

There is some labelling in existing titles in the series for genre, region and dialect, as there is in traditional dictionaries, but the core of the dictionary is based on a single language-wide, genre-blind corpus-based list. While it has always been a premise of dictionary-making that one could work with 'the language in general', and there is just a small minority of words that are marked for genre, it leaves those of us who work with corpora uncomfortable. In any corpus-based dictionary, one has to choose what text types to include in the corpus, and those choices will determine the outlook on the language that the dictionary presents. In a frequency dictionary, the whole issue is staring the compilers in the face, since the dictionary content is directly determined by the corpus.

For our dictionary, we decided to foreground genre in a way that other dictionaries in the series had not, by setting aside the 'one list' approach and presenting a number of lists, some of which would be genre-specific. This paper is a discussion of the issues that arise in that context, and our responses to them.

3 Corpus construction

We chose to use four different kinds of text – hereafter, 'genres'– for our corpus: spoken, fiction, newspaper and web. Of course there are arguments for going for a finer-grained set of genres, like Brown, but we restrained ourselves to genres where we had access to large numbers of texts, and to a modest number of genres so that the complexities of analysis were manageable.

Three of our four genres are the same as used by Biber and colleagues in much of their work. Their fourth is 'academic'; our fourth is 'web', in response to the growing importance of the web in our lives since Biber's research programme began, and also because of the heterogeneity of the web. It is less likely that we shall miss common words that mainly occur in genres other than conversation, fiction and newspaper.

For **Fiction** books from Flemish and Dutch authors, were available from the period 1970 to 2009.

Newspaper was taken from the SoNaR (2010, *Stevin Nederlandstalig Referentiecorpus*) corpus and comprises material from Belgian and Dutch newspapers. It makes up the largest part of our corpus.

For **Spoken** we used material from the Spoken Dutch Corpus (2004, CGN – *Corpus Gesproken Nederlands)*.

Web was again taken from the SoNaR corpus and includes blogs, discussion lists, e-magazines, newsletters, press-releases, and Wikipedia entries.

4 A fixed-sample-size corpus

In the Brown corpus, all samples are 2000 words long. While the strategy of truncating samples has its detractors – notably John Sinclair, who insisted that corpora should comprise complete texts – and we would not dispute that it makes a corpus unsuitable for some research questions, it also has many advantages. If all samples are 2000 words long, then we immediately know that any results will not be distorted by different sample lengths. A frequency of ten for a word in one sample will have the same weight as its frequency of ten in another.

Interpreting statistics where sample sizes vary often becomes complex and subtle, and we suspect one of the reasons for the ongoing success of the Brown model is its fixed sample size.

One of the central problems in preparing frequency lists is the whelks problem: if there is a text about whelks (a variety of mollusc) then the word *whelk* will probably occur many times. We would rather not give all of those occurrences equal weight in our word frequency list. (Gries 2008 presents a review of methods used to address the issue, and the Routledge dictionaries use a range of mathematical devices.) One simple and appealing strategy is not to count the number of occurrences of each word, but the number of samples the word occurs in. Then, however many times whelks are mentioned in a sample, it will just count as one sample. If samples are different

sizes – particularly if, as often happens in corpus-building, some are hundreds or thousands of times as long as others – this is problematic and figures are hard to interpret. But if all samples are the same length, there are no such complications and it is a straightforward response to the whelks problem.

For our dictionary, frequencies will always be numbers of samples the word occurs in. In homage to Brown, the fixed length for our samples is 2000 words. We first truncated any very long texts at 40,000 words, so we did not have too many samples from any single text, and then simply concatenated all the text of each genre and cut it into 2000-word slices. We considered more sophisticated strategies which paid heed to the beginnings and ends of texts, perhaps only taking one sample from each text (but then, for fiction, we would not have many samples) or not using short fragments (but most of the spoken material was in short fragments). We doubt that our crude strategy will have diminished the integrity of the resulting lists, though of course this is an empirical question.

5 The frequency lists

Instead of a single list, we propose a multi-part list. This is more complex than a single list. The question, "what is frequent enough to include?" is no longer straightforward. We used the following criterion. For each word, in each genre of the corpus, the frequency was identified. Where the average of these figures was over 1.125 the word would be included in the dictionary. This was the threshold that gave the 'top 5000' which are distributed over six lists as follows.

The Core List: The core vocabulary of a language comprises those words that are used across all kinds of uses of the language. This was implemented as those words that occurred with a frequency of at least *x* in each of the four genres. Table 1 gives the number of words that this method delivers, for various values of *x.*

x	90	50	30	10	5	4.5	4	3
#	36	112	190	477	856	944	1039	1345

Table 3: # words with a frequency of *x* in each of the four genres.

We used the 4.5 mark to identify core vocabulary, giving us 944 words. These words were then set aside and do not feature in any other lists.

The 'Genre' Lists (fiction, newspaper, spoken, web): The 'genre' lists include the high-frequency words for the different genres. For each genre we adopted the following base method:

- list the words according to frequency
- include the top items

The complication with this method is that some words occurred in two, three or four of the lists generated in this way, and for such cases we had to decide whether they go in:

- just one list
- more than one list
- the general list.

Our strategy has been to say there should be some cases of each, as follows:

- if highest frequency is at least double the next highest, list in that genre only
- if two are high and two are low, that is, the first- and second-highest, and both more than double the other two, list in both the top two
- else, list in the general list.

Applying this algorithm gave the following counts:

	This genre only	**This genre and one other**	**Total**
Spoken	65	93	158
Newspaper	561	573	1134
Fiction	830	270	1100
Web	109	419	528

Table 2: words to go in each of the genre-lists.

We note the familiar finding that written texts use more different words than spoken (so a larger proportion of tokens in spoken material will comprise core words) and that web is a mixture, sharing some characteristics of conversation but also sharing vocabulary with newspaper and fiction.

The general list: The general list consists of words which have a high frequency across at least three of the genres. Our method left 2007 words for the general list.

6 Conclusion

The question, "what genres should I include in my dictionary" is always a delicate one, and the more corpus-based we are, the more directly we must address it. For a new Dutch frequency dictionary, we are addressing it by basing the headword selection not on one corpus list, as is normal, but on four, for four main genres. This raises a number of questions such as "what is the threshold for a word being 'core' and what is the threshold for it being specialist", and "under what circumstances (if any) should a word feature in

more than one genre list?" We have given our tentative answers.

References

Corpus Gesproken Nederlands (CGN), (2004), Nederlandse Taalunie, Den Haag.

Gries, Stefan. 2008. Dispersions and adjusted frequencies in corpora. Int Jnl Corpus Linguistics 13 (4), 403-437.

SoNaR (2010), Nederlandse Taalunie, Den Haag.

A macroanalytic view of Swedish literature using topic modeling

Dimitrios Kokkinakis
Univ. of Gothenburg

Mats Malm
Univ. of Gothenburg

{first.last}@gu.se

1 Introduction and background

New research opportunities are plentiful for digital and literature scholars who are currently faced with increasingly large portions of large digitized archives produced during the last decades. Conventional methods of analysis involving a so called *close reading* view are not enough. *Distant reading* or *macroanalysis* is proposed instead, as a better, viable and more pragmatic alternative t*o the traditional methods of analyzing e.g., literature*. According to this view, understanding literature is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data. Therefore, applying macroanalytic methods and technologies is a priority among many research groups in the humanities worldwide. In this paper we explore *topic modeling*, an increasingly popular statistical method used for uncovering themes, topics and patterns in large amounts of text. We use available topic modeling software and, as empirical data, the content of the Swedish literature bank, a constantly growing body of Swedish fiction corpus from the 18th and 19th century. We present preliminary results on a sample of this corpus and discuss how humanistic research can be conducted through this type of computation, as a means to identify potential issues of interest e.g., for historians.

Close reading is the careful, sustained interpretation of a brief passage of text where great emphasis is placed on the particular over the general, paying close attention to individual words, syntax, and the order in which sentences and ideas unfold as they are read. In contrast, *distant reading* encapsulates *quantitative methods*, in which the reality of the text undergoes a process of deliberate reduction and abstraction; (Moretti, 2005; Nelson *et al.*, 2012; Jockers, 2013). Within the latter view, understanding literature is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data. This enables experimentation and exploration of new corpora uses and development that otherwise would be impossible to conduct. For such purposes, several

available techniques can be applied, one popular being *topic modeling* (Wallach, 2006; Brett, 2012; Graham *et al*., 2012). In topic modeling, a 'topic' is a probability distribution over words, or in other words a way of identifying recurring patterns of co-occurring words or clusters of words in a corpus. The models decompose the document collection into groups of words representing the main topics. In the best case, topic modeling can reveal groups of words that are semantically related and interpretable. The clusters of words are grouped together by a process of similarity, usually by variants or extensions of the vector space model. Topic Modeling has been applied successfully for various problems in literature studies; *cf.* Hall *et al.,* 2008; Gohr *et al.,* 2009; Yang *et al.,* 2011; Brown 2012 and Mimno 2012.

2 Corpora

Prose fiction is just a type of textual material that has been brought into the electronic "life" using large scale digitized efforts, an essential source within many disciplines of humanities and social studies. Prose fiction is complex and difficult to use not only because of interpretational complexity but also because of its limited availability. The Swedish Literature Bank, and its sister project the "19th Century Sweden in the Mirror of Prose Fiction[1]", aims to change this by developing a large representative corpus which mirrors society at given points in time, chronologically selected in such a way that historical comparisons can be made. A substantial part of the material is all fiction, written in the original and published separately for the first time, that appeared in Swedish starting from the year 1800 and collected during consecutive twenty year intervals. The material provides a whole century of evolution and social, aesthetic, scientific, technical, cultural, religious and philosophical change. For the experiments described in this paper, we selected the literary production, 13 novels, of a single author, namely Hjalmar Bergman (1883-1931). The selected novels are: *Savonarola* (1909); *Amourer* (1910); *Hans nåds testamente* (1910); *Vi Bookar, Krokar och Rothar* (1912); *Loewenhistorier* (1913); *Falska papper* (1916); *Herr von Hancken* (1920); *Farmor och Var Herre* (1921); *Eros' begravning* (1922); *Chefen fru Ingeborg* (1924); *Flickan i frack* (1925); *Kerrmans i paradiset* (1927); *Clownen Jac* (1930).

[1] The Swedish Literature Bank:
http://litteraturbanken.se/#!om/inenglish
the 19th Century Sweden in the Mirror of Prose Fiction:
http://spf1800-1900.se/#!om/inenglish

3 Experimental setting

Before applying topic modeling on our data, some pre-processing steps were applied, namely i) text conversion to lower case; ii) named entity recognition, that labels several classes of entities such as person, location and organization (Kokkinakis, 2004). This way, entities with multiple words would stay together in the topic modeling phase; and iii) words that are not tagged as named entities are lemmatized. Several topic model software, as well as variations of the techniques implemented are available. In this study we use MALLET (McCallum, 2002; Graham *et al*., 2012), a popular topic model package that implements Latent Dirichlet Allocation or LDA (Blei *et al.*, 2003), which is a special case of topic modeling. LDA is an unsupervised machine learning technique that does not require manually annotated training corpora, which are often expensive and thus unavailable for literary or historical domains. The fact that it accepts unannotated text as input makes it an ideal tool for exploring the large text collections being digitized *cf.* Brown (2012).

4 Case study: consumption patterns and life style

Borin *et al.* (2011) describe a method using various types of lexical resources representing different historical stages of Swedish for the development of semantically oriented text search applications for historical research, e.g., for investigating the emerging consumer society in 19th century Sweden. For comparison, we also decided to conduct experiments in this direction, but this time by *automatically* discovering topics related to the consumption society and life style that emerges in the 19th century Swedish literature corpus. Therefore, we took another approach using topic modeling in order to find the topics more strongly associated with some of the artifacts examined by Borin *et al.*, such as *soffa* 'sofa', *säng* 'bed', *byrå* 'dresser', *porslin* 'porcelain, china', *spegel* 'mirror', *möbel* 'piece of furniture', *klocka* 'clock, watch', etc. Thus, to find topics more strongly associated with such words we performed similar methodological steps as described by Yang *et al.* (2011). For instance, the "*soff*" examples are found by extracting each line in the sample that contains an instance of the string "*soff*" along with a window of five lines on either side. MALLET is then run on these examples, e.g., "*soff*, to find the top general topic groups over 1000 iterations with stopword removal.

5 Results and discussion

In order to evaluate the output of the topic modeling procedure we sampled a small set of novels from the Literature Bank since our hypothesis is that important topics about consumption would emerge from the data. In the near future we will perform expert evaluation in order to objectively assess the results, in a same manner as in Yang *et al.* (2011), i.e. Accuracy (topics) = # of relevant topics / total # of topics.

Topic "*soff*"
soffa bord förmakssoffa tycka midt hörnsoffa liggande släpa schaggsoffa revolver knä lögn rycka ljuga tagen placera soffhörn förgäves uppsöka pinnsoffa plagg fruntimmer rynka publik smyga konst sitta kula höger
Topic "*säng*"
säng sängkammare ligga sängkant täcke dö sängkammardörr kläder golv sitta madam rent resonera husfader hvit käring lof treva kapprak verklig kniv barnsäng nattlampa hviska ganska visst glida helst tak
Topic "*klock*"
klocka ringa klockare park klockslag gård klockeberg person börja rogershus väckarklocka ring kring munk förvaltare vetenskap björkenäs björkenä sällskap sol anlända ana sten lycka allra klockeberga guldklocka sälja högst

Table 1. Top-30 lemmatized words for three topics.

Topic modeling addresses important issues that are related to e.g., keyword searches in databases, since it can be difficult to find related words in topics that scholars may look for. Nelson (2012) discusses that topic modeling as distant reading approach is "comprehensive, allowing us to analyze models that are drawn not from a selection but from the entirety of massive corpora", moreover it "identifies word distributions (i.e. 'topics'), topic modeling encourages – even forces – us to examine larger groups of related words, and it surfaces resonant terms that we might not have expected". Finally, the topics identified are "all the more revealing because they are based on statistical relationships rather than *a priori* assumptions and preoccupations of a researcher".

We believe that topic modeling provides exciting opportunities for research in digital humanities, where new questions can be posed and new knowledge can be produced. Our results so far seem interesting in their own right, presented here primarily as examples of the value of tailoring topic modeling approaches to the available contextual data for a specific domain (Swedish fiction) and to specific threads of scholarly investigation. Apart from the expert evaluation we are also interested to enhance the models with other parameters, such as e.g., author gender, in order to measure thematic innovation and other forms of historical investigations across the 18th and 19th century Swedish literature landscape.

References

Blei D., Ng A. and Jordan M. 2003. Latent Dirichlet allocation. *J of Machine Learning Research* 3 (4–5): pp. 993–1022.

Borin L., Forsberg M. and Ahlberger C. 2011. *Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature*. NODALIDA. Pp. 58–65. Riga, Latvia.

Brown T. 2012. *Telling New Stories about Our Texts: Next Steps for Topic Modeling in the Humanities*. Proceedings of the Digital Humanities. Hamburg, Germany.

Brett M. 2012. *Topic Modeling: A Basic Introduction*. http://www.fredgibbs.net/clio3workspace/blog/topic-modeling/

Gohr A., Hinneburg A., Schult R., Spiliopoulou M. 2009. *Topic Evolution in a Stream of Documents*. Proceedings of the Ninth SIAM International Conference on Data Mining. Nevada, USA.

Graham S., Weingart S., and Milligan I. 2012. Getting Started with Topic Modeling and MALLET. The Programming Historian 2. http://programming historian.org/lessons/topic-modeling-and-mallet

Hall D., Jurafsky D. and Manning C.D. 2008. *Studying the History of Ideas Using Topic Models.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Pp 363-371. Hawaii, USA.

Jockers M.L.. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Kokkinakis D. (2004). *Reducing the Effect of Name Explosion.* Proceedings of the 4th LREC Workshop: Beyond Named Entity Recognition, Semantic Labelling for NLP tasks. Pp. 1-6. Lissabon, Portugal.

McCallum A.K. 2002. *MALLET: A Machine Learning for Language Toolkit*. (UMass Amherst's MAchine Learning for LanguagE Toolkit). http://mallet.cs.umass.edu

Mimno D. 2012. Computational Historiography: Data Mining in a Century of Classics Journals. *J. on Computing and Cultural Heritage (JOCCH).* Vol. 5:1. http://www.perseus.tufts.edu/publications/02-jocch-mimno.pdf

Moretti F. 2005. *Graphs, maps, trees: abstract models for a literary history*. R. R. Donnelley & Sons.

Nelson R.K., Mimno D. and Brown T. 2012. *Topic Modeling the Past*. Proceedings of the DH2012, Hamburg, Germany. http://www.dh2012.uni-hamburg.de/

Wallach H. 2006. *Topic modeling: beyond bag-of-words. Proceedings of the 23rd International Conference on Machine Learning (ICML).* Pp. 977–984. Pittsburgh, U.S. http://people.ee.duke.edu/~lcarin/icml2006.pdf

Yang T-I., Torget A.J. and Mihalcea R. 2011. *Topic Modeling on Historical Newspapers. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* Portland, OR. http://www.aclweb.org/anthology/W11-1513

Czech nouns derived from verbs with an objective genitive: Their contribution to the theory of valency[1]

Veronika Kolářová
Charles University in Prague
kolarova@ufal.mff.cuni.cz

1 Valency of nouns in the framework of the Functional Generative Description

Our approach to issues of valency of deverbal nouns is based on the theory of valency (especially valency of verbs) as developed in the framework of the Functional Generative Description (Panevová 1974, 1975; Sgall et al. 1986).

The following complementations (i.e. the individual dependency relations) are able to fill individual slots of the valency frames of verbs:

a. inner participants or arguments (they can be obligatory or optional): Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), Origin (ORIG);
b. obligatory free modifications or adjuncts, especially those with the meaning of location, e.g. DIR, LOC, and manner (MANN).

While treating the valency frames assigned to deverbal nouns denoting an action, the same set of complementations as with verbs is used (Piťha 1984; Panevová 2002). Such nouns are expected to inherit all participants that are present in the valency frame of their source verbs. Forms of the participants reflect typical shifts in surface forms of participants, e.g. ACC → GEN, cf. (1) (Karlík and Nübler 1998; Panevová 2002; Kolářová 2006).

(1) *Petr vrátil knihu kamarádovi*
'Peter-NOM.SG returned book-ACC.SG friend-DAT.SG'

→ *vrácení knihy kamarádovi Petrem*
'returning-NOM.SG book-GEN.SG friend-DAT.SG Peter-INS.SG'

There are two basic types of Czech deverbal nouns that in one of their meanings [2] can sometimes denote an action, i.e. nouns derived

[1] The research reported in the paper was supported by the Czech Science Foundation under the projects P406/2010/0875 and P406/12/P190.

[2] The differences in meanings are in literature mostly described as the difference between event and result nominals (Grimshaw 1991; Apresjan 1995; Alexiadou 2001; Melloni 2011).

from verbs by productive means, i.e. by suffixes *–(e)ní/tí*, e.g. *honění* 'hunting' or *hubnutí* 'losing weight', and nouns derived from verbs by non-productive means or by the zero suffix, e.g. *honba* 'hunt', *hon* 'hunt' (the differences between the two types of nouns are discussed e.g. in Dokulil 1982; Veselovská 2001; Karlík 2002; Procházková 2006; Kolářová 2006).

2 Czech nouns derived from verbs with an objective genitive

For the present study of valency properties of nouns derived from verbs with an object expressed by a prepositionless genitive (i.e. an adverbal genitive, GEN_{Adverb}), a representative corpus of contemporary written Czech was exploited, i.e. the Czech National Corpus (CNC), namely all morphologically annotated subcorpora of CNC, i.e. SYN2000, SYN2005, SYN2006PUB, SYN2009PUB and SYN2010.

Since Czech verbs with an object expressed by GEN_{Adverb} represent rather a small group of verbs (e.g. in comparison with verbs with an object expressed by a prepositonless accusative), Czech nouns derived from these verbs have not been in the focus of attention yet. Thus, we focus on four topics of the theory of noun valency and show that studying of the nouns can considerably contribute to the theoretical description of valency of Czech deverbal nouns. The topics are as follows: Correspondence GEN_{Adverb} – GEN_{Adnom}? (Section 3), Double post-nominal genitives (Section 4), Possessives corresponding to an adverbal objective genitive (Section 5), and Agents expressed by a prepositionless intrumental modifying nouns derived from intransitive verbs (Section 6).

3 Correspondence GEN_{Adverb} – GEN_{Adnom}?

An adverbal prepositionless genitive is a non-structural case and according to typical shifts in surface forms of participants it should be preserved with the derived noun. Thus the first task was to study adnominal counterparts of adverbal objects expressed by GEN_{Adverb}, e.g. PAT, cf. *obávat se čeho* 'be-INF_afraid REFL sth-GEN.SG', or ADDR, cf. *dotázat se koho* 'ask-INF REFL sb-GEN.SG').

First of all, the list of verbs with GEN_{Adverb} was created on the basis of several Czech valency dictionaries (esp. so-called VALLEX and PDT-VALLEX). The list contains approx. 150 lemmas of verbs. On the basis of the list of verbs, two lists of nouns were created: (i) the list of productively derived nouns, (ii) the list of non-productively derived nouns. After that, all the nouns were searched for, using the following query: ([lemma="..."] [!(tag="[Z|R|V|J].*")]{0,2} [tag="N...2.*"]); all found occurrences were manually checked. The intention was to find out whether Czech nouns derived from verbs with GEN_{Adverb} can be modified by PAT or ADDR expressed by GEN as well.

Concerning non-productively derived nouns, neither other papers nor valency dictionaries mention the genitive form of the respective participants. However, it has turned out that several nouns (namely *dotek* 'touch', *naděje* 'hope', *odvaha* 'courage', *památka* 'memory') modified by PAT in GEN_{Adnom} can rarely be found in the corpora, cf. (2) (Kolářová 2012). Other forms of PAT, i.e. prepositional groups (Jirsová 1966; Novotný 1980) and sometimes also an infinitive or an embedded objective clause, are more frequent. Factors that influence possibility or impossibility to be modified by PAT or ADDR in GEN_{Adnom} are typically connected with the type of the semantic class the noun belongs to; it concerns e.g. the difference between nouns denoting "positive" vs. "negative" mental state or dispositions, cf. (2) vs. (3), or the tendency to avoid syntactic homonymy of ACT and ADDR expressed by GEN_{Adnom}, which is typical of nouns of saying, cf. (4).

(2) *odvaha spolupráce*.PAT
'courage cooperation-GEN.SG'
(3) **obava následků*.PAT
'fear consequence-GEN.PL'
(4) *dotaz kamaráda*.ACT/*ADDR
'question friend-GEN.SG'

Concerning productively derived nouns, approx. 50 lemmas of the nouns were found with PAT or ADDR expressed by prepositionless GEN_{Adnom} corresponding to GEN_{Adverb}, cf. (5). In contrast to non-productively derived nouns, the adnominal genitive form corresponding to GEN_{Adverb} is a typical form, while prepositional groups are rare, cf. (6). The nouns were also divided into the subgroups according to their valency frames and semantic classes.

(5) *dotýkání se předmětů*.PAT
'touching REFL exhibit-GEN.PL'
(6) *zděšení z čeho*.PAT
'horror from sth-GEN.SG'

4 Double post-nominal genitives

In addition to the adnominal genitive form, also combinations with other participants were observed. Special attention was paid to constructions in which two participants (actants,

A_1 and A_2) are expressed by an adnominal prepositionless GEN (double post-nominal genitives; Alexiadou 2001). Up to now, the only Czech nominalized structure (NS) with double post-nominal genitives has been considered to be grammatical, i.e. NS_1, see (7), while the NS_2A, cf. (8), is ungrammatical (Karlík 2002).

(7) NS_1 in which A_1 (GEN_{Adnom} ← Ak) and A_2 (GEN_{Adnom} ← GEN_{Adverb})
*zbavení ženy.*ADDR *starostí.*PAT
relieving woman-GEN.SG worry-GEN.PL
'relieving the woman of worries'

(8) NS_2A in which A_1 (GEN_{Adnom} ← Nom) and A_2 (GEN_{Adnom} ← Ak)
**zkoušení Petra.*PAT *Evy.*ACT
examining Peter-GEN.SG Eve-GEN.SG
'examining Peter by Eve'

On the basis of material obtained from the CNC, we describe several other types of constructions that we consider to be grammatical, e.g. NS_3 and NS_2B, cf. (9) to (11). We claim that nominalizations of support verb constructions or other multi-word predicates (NSs marked by "B", i.e. NS_2B and NS_3B) play important role for the possibility to use a NS with double post-nominal genitives in Czech. We also specified typical and specific word order of A_1 and A_2 within the studied constructions.

(9) NS_3A, in which A_1 (GEN_{Adnom} ← Nom) and A_2 (GEN_{Adnom} ← GEN_{Adverb})
dožití dítěte konce pojistné doby
living_to child-GEN.SG end-GEN.SG of insurance period
'living of the child to the end of insurance period'

(10) NS_3B, in which A_1 (GEN_{Adnom} ← Nom) and A_2 (GEN_{Adnom} ← GEN_{Adverb})
zanechání činnosti řady klíčových hráčů
quitting activity-GEN.SG array-GEN.SG key player-GEN.PL
'quitting of activity by the array of key players'

(11) NS_2B, in which A_1 (GEN_{Adnom} ← Nom) and A_2 (GEN_{Adnom} ← Ak)
sbírání zkušeností nejmladších závodníků
gaining experience-GEN.PL youngest-GEN.PL competitor-GEN.PL
'gaining of experiences by the youngest competitors'

5 Possessives (prenominal genitives) corresponding to an adverbal objective genitive

We also concentrate on a new topic that had arisen, i.e. the expression of the adnominal participant corresponding to GEN_{Adverb} by a possessive adjective or a possessive pronoun (POSS). Under certain conditions, POSS can correspond to adverbal accusative (Anderson 1977; Alexiadou 2001; Karlík 2002). However, usage of POSS (← GEN_{Adverb}) has not been studied yet. Thus, usage of POSS (← GEN_{Adverb}) was searched for; we used the lists of nouns mentioned in Section 3 and applied the following query: ([(tag="PS.*") | (tag="AU.*")] []{0,1} [lemma="…"]). All found occurrences were manually checked. The form of POSS (← GEN_{Adverb}) was found with eight productively derived nouns, cf. (12), and with two non-productively derived nouns, cf. (13). Thus in addition to shifts POSS (← NOM) and POSS (← ACC) also the shift POSS (← GEN_{Adverb}) is possible in Czech (Kolářová, to appear).

(12a) *zanechání studia*.PAT
'quitting one's studies'

(12b) *jeho*.PAT *zanechání*
'its quitting'

(13a) *dotyk volantu*.PAT
'touch of wheel'

(13b) *jeho*.PAT *dotyk*
'its touch'

6 Agents expressed by a prepositionless intrumental modifying nouns derived from intransitive verbs

On the basis of the language material obtained from the CNC, it has also turned out that some nouns derived from verbs with GEN_{Adverb} can be modified by Agent (Actor) expressed by prepositionless instrumental, even though their source verbs do not allow passive constructions. It concerns esp. nouns derived from reflexive intransitive verbs, cf. (14).

(14) *ujímání se zvířátek*.PAT *hodnými lidmi*.ACT
taking_charge REFL animal-GEN.PL good-INS.PL people-INS
'taking charge of small animals by good people'

7 Conclusion

Czech nouns derived from verbs with an objective genitive contribute to the theory of noun valency in several aspects and it is useful to study their valency properties in detail. Corpus-based material obtained from the CNC provides a sufficient base for exemplifying even some marginal phenomena in Czech.

References

Alexiadou, A. 2001. *Functional Structure in Nominals. Nominalization and ergativity.* Amsterdam / Philadelphia: John Benjamins.

Anderson, M. 1977. 'NP Pre-posing in Noun Phrases'. *Proceedings of NELS 8*: 12-21.

Apresjan, J. D. 1995. *Leksičeskaja semantika. Sinonimičeskije sredstva jazyka.* Moskva: Vostočnaja literatura, RAN.

Dokulil, M. 1982. K otázce slovnědruhových převodů a přechodů, zvl. Transpozice (On transfers and transitions among parts of speech: the case of transposition). *Slovo a slovesnost* 43: 257-271.

Grimshaw, J. 1991. *Argument Structure.* Cambridge, Mass: The MIT Press.

Jirsová, A. 1966. Vazby u dějových podstatných jmen označujících duševní projevy (Valency of non-productively derived nouns denoting mental states or dispositions). *Naše řeč* 49: 73-81.

Karlík, P. 2002. Ještě jednou k českým deverbálním substantivům (Once more on Czech deverbal nouns). In Z. Hladká and P. Karlík (eds.) *Čeština – univerzália a specifika 4*: 13-23. Praha: Nakladatelství Lidové noviny.

Karlík, P. and Nübler, N. 1998. Poznámky k nominalizaci v češtině (Notes on nominalization in Czech). *Slovo a slovesnost* 59: 105-112.

Kolářová, V. 2006. Valency of deverbal nouns in Czech. *The Prague Bulletin of Mathematical Linguistics* 86: 5-19.

Kolářová, V. 2012. Valence dějových substantiv odvozených od sloves s předmětovým genitivem (Valency of nouns derived from verbs with an object expressed by prepositionless genitive). In S. Čmejrková, J. Hoffmannová and J. Klímová (eds.) *Čeština v pohledu synchronním a diachronním. Stoleté kořeny Ústavu pro jazyk český*: 609-614. Praha: Karolinum.

Kolářová, V. (to appear). Adverbální předmětový genitiv a jeho protějšky v nominálních konstrukcích: Případ posesiva (Adverbal objective genitive and its counterparts in nominal constructions: The case of possessives). In *Proceedings of the conference Slovo a tvar v štruktúre a v komunikácii.* Bratislava.

Melloni, Ch. 2011. *Event and Result Nominals. A Morpho-semantic Approach.* Bern: Peter Lang.

Novotný, J. 1980. *Valence dějových substantiv v češtině* (Valency of non-productively derived nouns in Czech). Sb. pedagogické fakulty v Ústí nad Labem. Praha: SPN.

Panevová, J. 1974. On Verbal Frames in Functional Generative Description. Part I. *The Prague Bulletin of Mathematical Linguistics* 22: 3-40.

Panevová, J. 1975. On Verbal Frames in Functional Generative Description. Part II. *The Prague Bulletin of Mathematical Linguistics* 23: 17-52.

Panevová, J. 2002. K valenci substantiv (s ohledem na jejich derivaci) (On valency of nouns (with respect to their derivation)). *Zbornik Matice srpske za slavistiku* 61: 29-36.

Piťha, P. 1984. Case frames of nouns. In P. Sgall (ed.) *Contributions to functional syntax, semantics, and language comprehension*: 225-238. Amsterdam / Philadelphia: John Benjamins.

Procházková, V. 2006. *Argument structure of Czech event nominals.* Master Thesis, University of Tromsø.

Sgall, P., Hajičová, E. and Panevová, J. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht: Reidel and Prague: Academia.

Veselovská, L. 2001. K analýze českých deverbálních substantive (A note on the analysis of Czech derived nominals). In Z. Hladká and P. Karlík (eds.) *Čeština – univerzália a specifika 3*: 11-27. Brno: MU.

MotionML: Motion Markup Language – a shallow approach for annotating motions in text

Oleksandr Kolomiyets
LIIR, KU Leuven
oleksandr.kolomiyets@cs.kuleuven.be

Marie-Francine Moens
LIIR, KU Leuven
sien.moens@cs.kuleuven.be

7 Introduction

Many NLP applications require information about locations of objects referenced in text, or relations between them in space. For example, the phrase *a book on the desk* contains information about the location of the object *book*, as *trajector*, with respect to another object *desk*, as *landmark*. Besides static (non-motional) relations, which have been widely studied in SpatialML (Mani et al. 2008) and Spatial Role Labelling (SpRL) (Kordjamshidi et al. 2010), *motion* appears to be the most difficult and controversial information to annotate. There has been an enormous effort in formalizing and annotating motions in natural language. While annotating motions is out of scope for SpRL and SpatialML, the most recent work on the Dynamic Interval Temporal Logic (DITL) (Pustejovsky & Moszkowicz 2011) presents a framework for modelling motions as a change of state, which adapts linguistic background considering *path constructions* and *manner-of-motion constructions*. On this basis the Spatiotemporal Markup Language (STML) has been introduced for annotating motions in natural language. In STML, motion is treated as a change of location over time, while differentiating between a number of spatial configurations along the path.

Being well-defined for the formal representations of motion and reasoning, in which representations either take explicit reference to temporal frames or reify a spatial object for a path, all the previous work seem to be difficult to apply in practice when annotating motions in natural language. It can be attributed to possible vague descriptions of path in natural language when neither clear temporal event ordering, nor distinction between the start, end or intermediate path point can be made.

In this paper we present the motion markup language (MotionML) designed for annotating motions in text, and which simplifies the previously introduced notion of path in order to provide practical motion annotations and their use in NLP applications. The proposed annotation language extends the set of markable elements derived from the previous research, and describes the major principles for their annotations. Yet, MotionML utilizes the best practice of well-established XML-based annotation standards for linguistic annotations such as TimeML (Pustejovsky et al. 2003) and SpatialML. A set of examples presents markable elements and how they are used for annotating motions in terms of relations established between them.

8 Rationale and Motion Markables

Let us consider a text:

> *In Brazil coming from the North-East I stepped into the small forest and followed down a dried creek.*

This text describes a motion, and the reader can identify a number of concepts which are peculiar for motions: there is an object whose location is changing, the motion is performed in a specific spatial context, with a specific direction, and with a number of locations related to the motion. Based on these fundamental concepts we can specify markable elements for MotionML, which are inherited from the Spatial Role Labelling task (Kordjamshidi et al. 2010).

Trajector: Trajector is a label assigned to a word or a phrase which denotes an object which moves, starts, interrupts, resumes a motion, or is forcibly involved in a motion. In the example above token *I* is labelled as a trajector. The annotation of trajectors in text is implemented by an XML tag <TRAJECTOR> which is specified by the following BNF[1]:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Trajector_id

Trajector_id ::= T<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Motion Indicator: Motion indicator is a label assigned to a word or a phrase which signals a motion of the trajector along a path. In our example, a number of motion indicators can be

[1] In the following BNFs, id refers to a unique markable identifier, start and end to the start and end character offsets for markables, text and comment are optional attributes for the annotated text and for annotator's comments.

identified: *Coming*, *stepped* and *followed down*. The annotation of motion indicators is implemented by an XML tag <MOTION_INDICATOR> which is specified by the following BNF:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Motionl_indicator_id

Motion_indicator_id ::= M<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Path: Path is a label assigned to a word or phrase that denotes the path of the motion as the trajector is moving along, starting in, arriving in or traversing it. In MotionML, as opposite to STML, the notion of path does not have the temporal dimension, thus whenever the motion is performed along a path, for which either a start, an intermediate, an end path point, or an entire path can be identified in text, they are labelled as path. A number of paths can be identified in the provided example: *from the North-East*, *into the small forest*, *a dried creek*. The annotation of paths is implemented by an XML tag <PATH> which is specified by the following BNF:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Motionl_indicator_id

Motion_indicator_id ::= M<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Landmark: The notion of path should not be confused with landmarks. For spatial annotations landmark has been introduced as a label for a secondary object of the spatial scene. Being of great importance for non-motion spatial relations, in MotionML landmarks are used to capture a spatial context of a motion, such as *Brazil* in the example above. The annotation of landmarks is implemented by an XML tag <LANDMARK> which is specified by the following BNF:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Landmark_id

Landmark_id ::= L<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Spatial Indicator: In addition to landmarks we employ spatial indicators that label words or phrases that trigger relations between trajectors and spatial contexts of motions, such as *In* in the example above. The annotation of landmarks is implemented by an XML tag <SPATIAL_INDICATOR> which is specified by the following BNF:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Spatial_indicator_id

Spatial_indicator_id ::= S<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Distance: In contrast to the previous annotation standards, in which *distances* and *directions* have been uniformly treated as signals, in MotionML if the motion is performed for a certain distance, and such a distance is mentioned in text, the corresponding textual span is labelled as distance. For example: *25 km*, *100 m*, but also *25 min by car*. The annotations of distances are implemented by an XML tag <DISTANCE> which is specified by the following BNFs:

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Distance_id

Distance_id ::= DIS<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Direction: Additionally, if the motion is performed in a certain (absolute or relative) direction, and such a direction is mentioned in text, the corresponding textual span is annotated as direction, e.g. *the North-East* in the example above. The annotations of directions are implemented by an XML tag <DIRECTION> which is specified by the following BNFs

attributes ::= id start end [text] [comment]

id ::= ID

{id ::= Direction_id

Direction_id ::= DIR<integer>}

start ::= integer

end ::= integer

text ::= CDATA

comment ::= CDATA

Using the markables defined above, the following annotations for our example can be produced[1]:

<SPATION_INDICATOR id="S1" text="In"/>

<LANDMARK id="L1" text"Brazil"/>

<MOTION_INDICATOR id="M1" text="coming"/>

<PATH id="P1" text="from the North-East"/>

<DIRECTION id="DIR1">

<TRAJECTOR id="T1" text="I">

<MOTION_INDICATOR id="M2" text="stepped"/>

<PATH id="P2" text="into the small forest"/>

<MOTION_INDICATOR id="M3" text="followed down"/>

<PATH id="P3" text="a dried creek"/>

9 Motions as Relations

By analogy with TimeML, SpatialML and STML, MotionML employs a technique, in which markables in text are related to each other by means of motion links. A motion link is implemented by an XML tag <MLINK> which is specified by the following BNF:

attributes ::= id trajector_id [landmark_id] [spatial_indicator_id] [motion_indicator_id] [path_id] [direction_id] [distance_id] [general_type]

id ::= ID

{id ::= Relation_id

Relation_Id ::= MID<integer>}

trajector_id ::= IDREF

{trajector_id ::= Trajector_id}

landmark_id ::= IDREF

{landmark_id ::= Landmark_id}

spatial_indicator_id ::= IDREF

{spatial_indicator_id ::= Spatial_indicator_id}

motion_indicator_id ::= IDREF

{motion_indicator_id ::= Motion_indicator_id}

path_id ::= IDREF

{path_id ::= Path_id}

direction_id ::= IDREF

{direction_id ::= Direction_id}

distance_id ::= IDREF

{distance_id ::= Distance_id}

comment ::= CDATA

With the provided specification, three links that capture motions can be annotated in our example:

<MLINK id="MID1" trajector_id="T1" landmark_id="L1" spatial_indicator_id="S1" motion_id="M1" direction_id="DIR1"/>

<MLINK id="MID2" trajector_id="T1"" motion_id="M2"/>

<MLINK id="MID3" trajector_id="T1" motion_id="M3"/>

10 Conclusions

In this paper we presented MotionML – the motion annotation language for annotating motions in natural language. It extends the set of markable elements derived from the previous research, and describes the major principles for their annotations. An important simplification for the notion of path was introduced to enable annotations of underspecified and traversing paths, which are very common in natural language. Moreover, the use of landmarks enables a distinction between the motion path and its locative context, which seems to be impossible with the STML approach. The introduced simplification in MotionML, however, does not restrict the expressiveness of the language and its potential scalability to capture fine-grained locative semantics in the STML fashion.

References

Kordjamshidi, P., Moens, M. F., & van Otterlo, M. 2010. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh conference on International Language Resources*

[1] For reading purposes we omit the start and end offset values, and provide textual extents as text values.

and Evaluation (LREC'10), pp. 413-420.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., & Wellner, B. 2008. SpatialML: Annotation scheme, corpora, and tools. In *6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco*.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., & Radev, D. 2003. TimeML: Robust specification of event and temporal expressions in text. New Directions in Question Answering, 3, 28-34.

Pustejovsky, J., & Moszkowicz, J. L. 2011. The qualitative spatial dynamics of motion in language. Spatial Cognition & Computation, 11(1), 15-44.

Use of dedicated multimodal corpora for curriculum implications of EAP/ESP programs in ESL settings

Menikpura DSS Kumara
Rajarata University of Sri Lanka
skmenikpure@gmail.com

1 Introduction

Corpus linguistics is popular for researching patterns of authentic language use as well as for providing insights in developing curriculum for language teaching. The present research is an attempt to study the spoken academic discourse at a Sri Lankan university setting, where the target language, English, is used as a second language. It was carried out in order to improve the curriculum of English Language Teaching (ELT) program of the Faculty of Applied Sciences of Rajarata University of Sri Lanka. It was initiated on the understanding that the teaching/learning materials used in the research locus, based on commercially available English for Specific Purposes (ESP) and English for Academic Purposes (EAP) text books, are inadequate to help students comprehend spoken lectures, and that they are not immediately authentic. In order to address these issues, a dedicated multimodal corpus of spoken academic discourse of the research locus was compiled.

Among the many descriptions of multimodal corpora, a multimodal corpus has been defined as 'a digitized collection of language and communication-related material, drawing on more than one modality' (Allwood 2008). The type of multimodal corpus used in the present research consists of audio-video recordings, and transcriptions of recorded speech, focusing on sensory and production modalities such as vision, hearing, and gestures. Because the corpus is compiled for pedagogic implications, simple orthographic transcriptions are used. Transcriptions are used for concordance analysis of the corpus as well as for teaching material preparation, whereas audio-video recordings are used for aiding transcription, and, more importantly, for authentic teaching material preparation.

The corpus used in the present research is a 'dedicated corpus' or a 'specialized corpus' in the sense that it does not contain standard language as a whole, but is limited to a particular register, i.e. academic language. In terms of the number of words in transcribed text, dedicated corpora are usually rather small in comparison to general

corpora. Small corpus size is not a problem here because the objective of the research is to work out suitable teaching material for a specific teaching situation.

2 Methodology

In the compilation process, two samples each of the most common speech event, lectures, representing the four major academic divisions of the Applied Sciences study program, were audio-video recorded. Of the eight lectures recorded thus, four were transcribed in the initial stage, making the RAjarata Corpus of Academic English (RACAE) of 25,906 words. Audio-video recordings were saved in separate Digital Video Discs (DVDs).In the initial stage of the corpus, the transcripts do not include detailed 'coding of gestural and other non-linguistic features' (Thompson 2005), and they are not aligned to the audio-video files. Such automation will be done in a future stage of the corpus.

Analysis of the corpus for the intended purpose was done using six tools of the online concordancer, Compleat Lexical Tutor (CLT)[1]: VocabProfile, Frequency, Range, Keywords, N-Gram, and Concordance. Audio-Video files, along with transcripts, were also used for teaching material preparation for the target teaching situation.

3 Results and discussion

The concordance analysis of the compiled RACAE corpus reveals that the realization of structures for discourse topics such as 'definition', 'classification', and 'exemplification' in the spoken academic discourse differs from what is prescribed in ESP text books for science-based ESP/EAP programs. Corroborating Flowerdew (1993) observation, the prescribed structure for definitions, 'X is/can be defined as...', is realized in the present corpus as an active or passive structure with the verb 'call'. Similarly, the prescribed structure for classifications, 'X may be classified as....', manifests in the corpus by nouns such as 'type', 'kind', and 'group'. In place of the prescribed formula for exemplification, 'is/are exemplified/illustrated by...', RACAE corpus sees the phrases, 'for example', 'example of', 'such as, and 'like'. Thus, it is quite evident that the teaching situation has failed to include these realizations of spoken discourse in its teaching materials.

Specific, defining vocabulary for the four major academic divisions of Applied Sciences in the research locus; Biology, Physical Science, Health Promotion, and Information & Communication Technology; was extracted using the 'KeyWords' tool of CLT, which extracts the more frequent words in the sub-corpora of RACAE, proportionately to a general corpus, Brown Corpus in the present case. A similar list of specialised vocabulary was extracted using the 'Range' tool. There, the first and the second most frequent thousand words have to be excluded from the extraction, by using 'Stop-lists' function.

'VocabProfile' and 'Frequency' tools yielded the most common words in the RACAE corpus, which are general lexis, with a high frequency of function words. Deviating from the typical native speaker written text vocabulary distribution of 70%:10%:10%:10% among 0-1000: 1001-2000: Academic Word List: OffList categories, the VocabProfile for the RACAE corpus yielded 80.5%:4.5%:5.5%:9.5 result. This suggests the teaching situation to pay equal attention on common general lexis in order to accommodate spoken discourse in teaching materials. Results for 'N-gram' tool also support this suggestion as they manifest in RACAE corpus as more structural bundles, such as 'if you look at' or 'you have to' than being lexical bundles, such as 'rate of vulcanization' or 'in the presence of'.

In addition to providing these valuable insights to the curriculum and teaching material of the target teaching situation through concordance analyses of the transcripts, audio-video files and transcripts are also used for direct teaching material preparation. Audio-video files of this multimodal corpus are especially useful as listening material because the informants are from the same second language setting; their accent is familiar to the students. Listening activities based on the corpus provide the students with an opportunity to have prior exposure to the language of the teachers they are going to meet in their main courses. There, the audio-video files replace the need to have mini lectures for ESP programs.

Audio-video files and transcripts can also be used to help students learn the structures they need in comprehending the academic discourse. For example, in order to highlight the structures and lexis for learning cause and effect relationships, sections of transcripts having discourse markers like *because* and *therefore* omitted are given as a listening clause exercise, and the audio file is played. This provides a better intake than when using decontextualized examples. Multi modal corpora such as the RACAE corpus can also be used for general comprehension within a specialised register. For instance, comprehending the use of humour or

[1] http://www.lextutor.ca/

references to common topics like politics and sports, used by lecturers to exemplify academic matter, helps the students to understand lectures. This is especially meaningful with multimodal corpora because the video files show the student the changing facial expressions of the lecturer.

For pedagogical purposes, multimodal corpora suit better than textual alone corpora because the former cater to a wider range of learning styles and strategies of learners. When used alone with the text, sound files improve the learners' listening skills. Video files also complement comprehension through body language of the informants and the visuals of the pragmatic situation in class. It is not practically possible to incorporate all these aspects of linguistic input in a text alone corpus.

4 Conclusions

The present research project was launched to address the inadequacies of English teaching curriculum in accommodating aspects of spoken discourse in an ESP course at a second language setting. In order to analyze the target register of the language, a dedicated corpus was compiled. It is a multimodal corpus, which is available in several modes. Analysis of the corpus provided the features of the spoken academic discourse in the target setting that should be incorporated in the ELT curriculum of the target ESP course. Additionally, the various modes of the corpus were used directly for the preparation of authentic teaching materials for the target ESP course. Thus, dedicated multimodal corpora are a very useful tool for improving the curriculum of EAP/ESP programs in ESL settings.

References

Allwood, J. 2008. *Multimodal Corpora*. In: Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter: 207-225. Available online at: http://sskkii.gu.se/jens/publications/docs101-150/109.pdf

Flowerdew, J. 1993. *Concordancing as a tool in course design*. System, 21(2): 231-244.

Thompson, P. 2005. "*Spoken Language Corpora*" in *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 59-70.

Early Modern English vocabulary growth

Ian Lancashire
University of Toronto
ian.lancashire@utoronto.ca

Elisa Tersigni
University of Toronto
elisa.tersigni@utoronto.ca

1 The size of Early Modern English vocabulary

A population of fewer than three million people at any time spoke Early Modern English (EME), but its total vocabulary grew faster than at any other period, thanks to printing and translation. However, the surviving texts supplying the data would have been used only by a small, mainly London-based population. Ninety percent of the English people lived in a rural countryside; and eighty percent of Elizabethan men and ninety-five percent of women were illiterate. It is perhaps risky to calculate vocabulary size solely on what only twelve percent of its speakers used. Although little oral speech survives in transcription from the period, expressive writing in journals and letters does.

In 1611, English existed in three states: a core vocabulary (300-400 hundred words, many of them closed-class), an additional nimbus of between 4,000 and 8,000 common words (what children absorb at home from their mother, and what adults acquire from family, friends, and co-workers), and a substantial lode of hard or technical words (a learned vocabulary, comprising the jargon of many registers, obtained by book-study or by membership in a profession or craft-guild). Our hypothesis is that a true measure of vocabulary increase must take into account all three states, the core, the nimbus, and the hard or technical words. If most words invented or fetched from foreign stock to assist in translation (Bell and Barnard 1992; McDermott 2002) -- words now nestling in OED and *Lexicons of Early Modern English* citations -- meant little to most speakers of England, does it make sense to say that English as used by most people greatly expanded? Although the Human Protein Reference Database has increased English vocabulary today by 30,000 words, it affects the vocabulary of very few people today.

We hypothesize that the well-documented high rate of vocabulary expansion was partly illusory owing to printing, which saved an archival lode of words that, in previous centuries, would have

been recorded only in manuscripts, many of which would then have ultimately been lost through fire, war, and the dissolution of the monasteries in the 1530s.

At present, we have a good estimate of the total vocabulary increase. Wermser's (1976) and Lancashire's studies (2010) of the *Chronological English Dictionary* (*CED*; see *Ordered Profusion* 1973), and of OED and *Lexicons of Early Modern English* (*LEME*) data, agree that new words increase over a century from 1490–1509 to 1590–1609 by 75 percent. According to the OED, by 1500, English had just over 36,100 different word-forms, and 64,300 by 1600: an increase of 78 percent. 95.7 percent of *LEME* word-entries are not used by the OED.

2 The core vocabulary

The size of the core vocabulary of English in the Renaissance grew much less than it did in the thirteenth and fourteenth centuries (Márquez and Speck 2008: 69). This fact might well change the importance we attribute to the EME vocabulary explosion. The core vocabulary of even Shakespeare (the terms used in each play he wrote) consisted of fewer than four hundred words (Spevack 1972).

3 Common words

Our preliminary estimate of common EME unlemmatized word-forms, somewhat over 8,000 items, is the average English vocabulary in a group of seven well-known dictionaries, glossaries, and word-lists, 1480–1580. All but one are bilingual dictionaries, which mainly use well-known English equivalents; it does not make sense to explain a foreign-language term by a hard word unknown to most readers. Here are those lexicons (all found in *LEME*).

- Eng.-Lat. Medulla Gram. (Pepys; ca. 1480): 6,951
- Eng.-Lat. Promptorium Parvulorum (1499): 9,300
- Palsgrave's Eng.-Fr. dictionary (1530): 11,200
- Elyot's Lat.-Eng. Dictionary (1538): 12,700
- Salesbury’s Welsh-Eng. dictionary (1547): 4,444
- Withals's Eng.-Lat. dictionary (1556): 3,845
- Mulcaster’s Eng. word-list (1582): 8,143

The last example, by the finest English teacher of his age, William Mulcaster, lists headwords for a projected monolingual dictionary. Both Mulcaster’s word-list and the average of the two late medieval lexicons are about 8,000 words; that is, between 1480-99 and 1582, there appears to be little change in vocabulary size from these two (admittedly few) texts.

4 Hard words

The total number of hard words, lemmatized, can be estimated from two sources: Schäfer's *Early Modern English Lexicography* (1989), which compiles terminology from hard-word glossaries until 1640; and those hard-word (medical and legal) dictionaries he excludes and that *LEME* includes. By subtracting this total from the total vocabulary up to 1640, we can obtain one estimate for total number of common English words. This analysis is ongoing.

5 Authorial Corpus of Letters, Journals, and Diaries

We measure the common vocabulary of representative individual writers from 1450 to 1660, men and women, in personal letters and journals (cf. Cusack 1998). These writings are less likely to be swelled by hard words transliterated from foreign languages during translation, or by terms of art from professional registers. Court depositions and dialogues are possible sources, but the former give constrained speech, and the latter are artificial concoctions.

Our text-selection principles resemble those used in the Helsinki Corpus of English Texts (Kyto 1996), although our corpus is very modest in size. We pay special attention to gender, date, social class, and type of journal or diary (normally determined by the writer’s occupation), and we begin ca. 1450, a quarter century before the usual beginning date for Early Modern English, to establish a baseline vocabulary for comparison.

6 Lemmatizing the Corpus

Our way of counting vocabulary items in a text is to lemmatize them, that is, to convert all words into (OED) headword forms, and to tag each word-occurrence by its part of speech. A word may be spelled unpredictably in many different ways, and both verb and noun inflections must be reduced to a single form (respectively, the infinitive and the nominative singular). Because the same lemmatized spelling can serve homonyms and because zero derivation (the conversion of a noun to a verb, a verb to an adjective, etc.) is a widely-used vocabulary enhancement technique in this period, the vocabulary unit of our count is a lemma and its (current) part of speech. To ensure that our counts are comparable, we base them on lemma-types in

identically-sized excerpts.

We do lemmatization manually, not having found an EME old-spelling-to-lemma conversion table or computational procedure. Headword old-spellings and encoded lemma equivalents in *LEME* texts have supplied the beginnings of our own manual conversion table. It grows the more texts we lemmatize. At present it has 20,000 entries.

7 Preliminary Results

Word-form counts derived from analysis of modernized and lemmatized word lists support the hypothesis that a 75 or 78 percent increase in English vocabulary in the early modern period, as measured and reported in previous studies, is not reflected in the writing of authors. Margaret Paston, John Paston I, and William Cely (writing in the fifteenth century) are baseline authors.

Baseline authors	Types	Tokens	TTR
Margaret Paston	833	5,000	16.7
John Paston I	946	5,000	18.9
William Cely	711	5,000	14.2
Average	830	5,000	16.6

Table 1: Baseline Authors

The highest increase in a single author's vocabulary is 53.6 percent over the baseline (Chamberlain; see Table 2), two-thirds the increase suggested by studies of dictionaries in the period. The vocabulary of all tested authors in our still small (admittedly unrepresentative) group has an average increase of just 18.1 percent.

Modernization and lemmatization, which remove spelling and inflectional variation in the period, reduce the vocabulary counts. For example, after lemmatizing, Margaret Paston's 1,087 tokens drop to 833 and her TTR (type-token ratio) falls from 21.7 to 16.7. Lemmatization will reduce vocabulary counts by all authors.

Factors affecting vocabulary size include gender, class, and occupation, but the significance of these factors has not yet been quantified. Women in the period, for example, tend to have a significantly lower token count and TTR than men do. However, John Husse, an employee of the Lisles, has one of the smallest calculated vocabularies, suggesting that class may be a more significant factor than gender in determining one's vocabulary. We will test and quantify the significance of these factors over the early modern period. Although our statistical analysis is in the preliminary stage, the increase in TTR over time does not appear to be statistically significant.

Author	Types	Tokens	% increase over baseline	TTR
Lady Lisle	887	5,000	6.6%	17.7
John Husse	843	5,000	1.8%	16.9
J.Chamberlain	1,259	5,000	53.6%	25.2
Lady Hoby	591	5,000	-28.9%	11.8
Lady Mildmay	1,100	5,000	32.5%	22.0
Dorothy Sidney	886	5,000	6.6%	17.7
Dorothy Moore	1,162	5,000	39.8%	23.2
Gilbert Burnet	1,094	5,000	31.9%	21.9
Average	978	5,000	18.1%	19.6

Table 2: Subsequent Authors

In addition to measuring and reporting actual vocabulary increase, we are comparing the results of methods that include and exclude lemmatization. Manual lemmatization, necessary for accurate vocabulary counts, is time-intensive. We hypothesize that vocabulary counts can be estimated from unmodernized, unlemmatized text by establishing known ratios of old-spelling word-forms to their lemmatized terms, decade by decade. These ratios will presumably change as spelling standardization grew. Our hypothesis, that the proportions of types and tokens between authors will be proportionate before and after lemmatization, will be tested, and the margins of error calculated.

Our lemmatization software was developed in Visual Basic and uses a dictionary of early modern spelling forms to identify and modify old-forms to include modernized-form and part of speech. Each word is subsequently checked manually for accurate modernization and part of speech designation. Our methodology will be presented along with the results.

References

Bell, M., and J. Barnard. 1992. "Provisional count of *STC* titles 1475-1640." *Publishing History* 31: 48-64.

Burnet, Gilbert. 1907. *Some Unpublished letters*. Camden Soc., 3rd series. London: Royal Hist. Soc.

The Cely letters. 1975. Ed. A. Hanham. London: Oxford University Press for EETS.

Chamberlain, J. 1861. *Letters written by John Chamberlain*. Ed. S. Williams. Camden Society 79. Westminster: J. B. Nichols. Internet Archive.

The Correspondence ... of Dorothy Percy Sidney. 2010. Ed. M.G. Brennan, N.J. Kinnamon, and M.P. Hannay. Farnham, Surrey: Ashgate.

Cusack, B., ed. 1998. *Everyday English 1500-1700: a*

reader. Edinburgh: Edinburgh University Press.

EEBO-TCP. 2000-. *Early English books online / text creation partnership*. Ann Arbor: University of Michigan. URL http://www.lib.umich.edu/tcp/eebo

Elyot, Sir T. 1538. *The Dictionary of syr Thomas Eliot*. London: T. Bertheleti. LEME.

Finkenstaedt, T., and D. Wolff, with contributions by H. J. Neuhaus and W. Herget. 1973. *Ordered profusion: studies in dictionaries and the English lexicon*. Heidelberg: Carl Winter.

Finkenstaedt, T., E. Leisi, and D. Wolff. 1970. *A chronological English dictionary listing 80 000 words in order of their earliest known occurrence*. Heidelburg: Carl Winter.

Hoby, Lady Margaret. 1998. *The private life of an Elizabethan lady: the diary of Lady Margaret Hoby*. Stroud, Glos.: Sutton.

Kytö, Merja, ed. 1996. *Manual to the diachronic part of the Helsinki corpus of English texts: coding conventions and source texts*. 3rd edn. Helsinki: Department of English, University of Helsinki. Oxford Text Archive.

Lancashire, I. 2010. "Why did Tudor England have no monolingual English dictionary?" *Webs of Words: New Studies in Historical Lexicography*. Ed. J. Considine. Newcastle upon Tyne: Cambridge Scholars Publishing. 8-23.

The letters of Dorothy Moore 1641-1649. 2003. Ed. L. Hunter. Oxford Text Archive. Oxford Text Archive.

Lexicons of early modern English. 2006-. Ed. I. Lancashire. Toronto: University of Toronto Library and University of Toronto Press. URL: leme.library.utoronto.ca

Lisle letters. 1981. Ed. M. St. Clare Byrne. Chicago: University of Chicago Press.

Márquez, M. F., and B. P. Speck. 2008. "The spoken core of British English: a diachronic analysis based on the BNC." *Miscelánea: A Journal of English and American Studies* 37: 53-74.

McDermott, A. 2002. "Early dictionaries of English and historical corpora: in search of hard words." *A changing world of words: studies in English historical lexicography, lexicology and semantics*. Ed. Javier E. Díaz Vera. Amsterdam, Netherlands: Rodopi. 197-226.

Martin, R. 1994. "The Autobiography of Grace, Lady Mildmay." *Renaissance and Reformation* 18.1: 33-81.

Paston letters and papers of the fifteenth century. 3 parts. Ed. N. Davis. Oxford, Clarendon Press, 1971-2005.

Schäfer, J. 1989. *Early modern English lexicography*. 2 vols. Oxford: Clarendon Press.

Spevack, M. 1972. "Shakespeare's English: the core vocabulary." *Review of National Literatures* 3.2: 106-122.

Wermser, R. 1976. *Statistische studien zur entwicklung des englischen wortschatzes*. Bern: Francke Verlag.

Withals, J. 1553. *A shorte dictionarie for yonge begynners*. London: T. Berthelet.

Detecting cohesion: semi-automatic annotation procedures

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
e.lapshinova@mx.uni-saarland.de

Kerstin Kunz
Heidelberg Universität
kerstin.kunz@iued.uni-heidelberg.de

1 Introduction

In the present paper, we describe procedures to semi-automatically annotate a corpus on the level of *cohesion*. Our annotations include fine-grained classifications of various cohesive types which permit a multifaceted investigation of cohesion across languages and registers. In this way, the annotated corpus is one of the few existing resources supporting contrastive studies of cohesion as, to our knowledge, existing resources provide annotations of individual cohesive phenomena only, e.g. *pronoun coreference* in the BBN Pronoun Coreference and Entity Type Corpus, cf. Weischedel and Brunstein (2005), *verbal phrase ellipsis* in (Bos and Spenader 2011) or *conjunctive relations* in PDTB, cf. Prasad et al. (2008) for English, or annotation of *anaphora* in (Dipper and Zinsmeister 2009) and (Dipper et al. 2012). Moreover, most of these studies apply manual annotation procedures only that takes much time and effort. This work is an attempt to apply corpus-based semi-automatic procedures to identify candidates expressing cohesion in English and in German.

2 Motivation and background

Our main aim is to provide a resource which yields empirical data about contrasts between English and German on the textual or discourse level (cohesion). This information will enable studies which can enhance systemic and monolingual insights on cohesion, cf. (Halliday and Hasan 1976; Beaugrande and Dressler 1981; Brinker 2005) and complement English-German contrastive studies on the level of lexicogrammar, cf. (Hawkins 1986; Rohdenburg 1990; Steiner and Teich 2004; König and Gast 2007). This requires an investigation of various types of cohesive devices and the linguistic expressions to which they connect. Contrasts in the realization of these phenomena between English and German are explored in terms of frequency of realisation but also with respect to syntactic and semantic and conceptual functions, which may vary in translations and originals, spoken and written texts or be a consequence of register constraints.

To facilitate this kind of analysis we semi-automatically enrich our corpus with annotations of five main categories of cohesive devices defined by Halliday and Hasan (1976): reference, substitution, ellipsis, conjunctive relations and lexical cohesion.

3 Corpus resources

Our multilingual corpus offers a continuum of different registers from written to spoken discourse. More precisely, it includes English and German texts of 10 registers: eight of these are in written and four – in spoken mode. The written part was imported from an existing corpus described by Neumann (2005). The spoken data has been collected from different sources (cf. Lapshinova et al. forthcoming). GECCo consists of six corpora: GO (German originals), EO (English originals), GTRANS (German Translation), ETRANS (English Translation), EO-SPOKEN (English originals) and GO- SPOKEN (German originals). GO, GTRANS, EO and GTRANS contain eight further subcorpora (one for each written register) [1] and EO- and GO-SPOKEN consist of two subcorpora (for two spoken registers)[2].

GECCo is tagged on the following levels: token, lemma, morpho-syntactic information, part-of-speech, chunk and sentence boundary. The annotation of the written part was partly imported from CroCo (Neumann 2005). For the spoken part, we use Stanford POS Tagger (Toutanova et al. 2003) and the Stanford Parser (Klein and Manning 2003). The corpus is encoded in CWB and can be queried with Corpus Query Processor (CQP) (Evert 2005).

4 Annotation procedures

Categories to annotate: As already mentioned above, we use the categories of cohesive devices proposed for English by Halliday and Hasan (1976). For the comparison of these categories in English and German, we define their types and functions illustrated in table 1.

For the time being, we have annotated the written part of our corpus with the following categories: reference, conjunction and substitution.

[1] Fiction, political essays and political speeches, letters of shareholders, manuals, popular-scientific texts, tourism leaflets an websites.
[2] Interviews and academic speeches.

Device	Type	Function
reference	personal	personal, possessive, *it*-endophoric, *it*-exophoric
	demonstrative	head, modifier, local, temporal, pronominal adverb
	comparative	particular, general
conjunction	connects, subjuncts, adverbials	additive, adversative, causal, temporal, modal
substitution	nominal, verbal, clausal	NA
ellipsis	nominal	subject, object
	verbal	operator, lexical, full
	clausal	full, part

Table1. Categories of cohesion, their types and functions

Automatic annotation: To automatically tag our corpus with the information on cohesive devices, we use the annotation procedure derived from the methods used for the YAC chunker (Kermes 2003). The CWB tools have an option to incrementally enhance corpus annotations, as query results deliver not only concordances of the searched structures but also information on their corpus positions. So we can import the information on queried data back into the corpus using CWB Perl modules. Therefore, our annotation rules are defined in form of CQP queries that allow regular expressions based on string, parts-of-speech, chunk and further constraints.

	CQP query	**example of tagged structures in XML**
(1)	<chunk> ([_.chunk_gf="adv_temp"]+| [word="then|now"&pos="rt"])	<reference type=”dem” func=”temporal”> now </reference>
(2)	[lemma="this|these|those|tha t"&pos="dd.*"] [pos="j.*|n.*|mc|vvg|md"]	<reference type=”dem” func=”modifier”> this </reference>

Table 2. Queries and tagged structures in XML

As a result, we obtain lists of textual instances of a particular cohesive device along with its corpus position that we use to tag this cohesive device with the respective structure, as demonstrated in table 2. For example, query (1) is designed to extract textual instances of local demonstrative reference, whereas query (2) delivers occurrences of demonstrative reference with the grammatical function of a modifier. These instances are tagged as *reference* of type *demonstrative* and function *temporal* or *modifier* respectively. The tags are then imported back into the corpus and saved as CQP structural attributes.

Manual correction and evaluation: As our aim is to produce a corpus with highly precise information on cohesive devices in English and German, we integrate a step of manual correction into our procedures. To facilitate the manual correction, the annotated corpus (with the structures in XML format as shown in table 2 above) is imported into MMAX2, a tool for manual annotation (cf. Müller and Strube 2006).

Correction by human annotators allows us, on the one hand, to evaluate automatic procedures and to improve them, on the other hand (the rules for extraction are improved on the basis of human annotators' observations).

Our preliminary results show that in the automatic identification of reference, we are able to achieve the precision of 94% for English originals and of 66% for the German ones. The evaluation of the identification of substitution provided precision of ca. 84% for English and 71% for German subcorpora. We were also able to calculate recall for the procedures to annotate reference, that respectively estimates 89% and 62% for English and German.

5 Conclusion and future work

In the present paper, we have described semi-automatic corpus-based procedures to annotate cohesion that allow both automatic extraction of cohesive devices from our corpus, and their automatic annotation. Moreover, the integrated procedure of manual correction enables evaluation and improvement of the automatic procedures. This procedures facilitate our corpus-based analysis of German-English contrasts in cohesion, as it saves labour and time. Furthermore, it provides a possibility of consistent annotation on the basis of the pre-defined rules, which cannot be ensured if the entire annotation is of manual character.

Future work will also include annotation of cohesive ellipsis, lexical cohesion and cohesive chains. Besides that, we aim at annotating all categories of cohesion in the spoken part of our corpus, as our first observations lead us to suggest considerable differences in frequency and function between spoken and written registers. These differences result from differing conditions of speech such as strong relation to the communication situation, direct interaction of speech participants and constraints on cognitive processing.

References

Bos, J. and Spenader, J. 2011. *An annotated corpus for the analysis of VP ellipsis.* Language Resources and Evaluation. Springer Netherlands.

Beaugrande, R. A. de and Dressler, W.U. 1981. *Introduction to Text Linguistics*. London, New York: Longman . (German version also in 1981 published by Niemeyer).

Brinker, K. 2005. *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. 6 edition. Berlin: Erich Schmidt.

Dipper, S. and Zinsmeister, H. 2009. Annotation discourse anaphora. In *Proceedings of the Workshop "Third Linguistic Annotation Workshop", LAW III, ACL-IJCNLP-2009*. Suntec, Singapore, pp. 166-169.

Dipper, S., Seiss, M., and Zinsmeister, H. 2012. The Use of Parallel and Comparable Data for Analysis of Abstract Anaphora in German and English. In *Proceedings of LREC-2012*. Istanbul, Turkey.

Evert, S. 2005. *The CQP Query Language Tutorial.* IMS Stuttgart. CWB version 2.2.b90.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. London, New York: Longman.

Hawkins, J. A. 1986. *A Comparative Typology of English and German: Unifying the Contrasts*. London: Croom Helm.

Kermes, H. 2003. Off-line (and On-line) *Text Analysis for Computational Lexicography.* Ph.D. thesis, IMS, Universität Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Klein, D. and Manning, C.D. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

König E. and Gast, V. 2007. *Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik.* Berlin: Schmidt (revised 2 edition: 2009).

Müller, C. and Strube, M. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In S. Braun, K. Kohn, J. Mukherjee (eds.). *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, (English Corpus Linguistics, Vol.3), pp. 197-214.

Neumann, S. 2005. *Corpus Design.* Deliverable No. 1 of the CroCo Project. Available online at http://fr46.uni- saarland.de/croco/corpus_design.pdf.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, N. 2008. Penn Discourse Treebank Version 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakesh, Morocco.

Rohdenburg, G. 1990. Aspekte einer vergleichenden Typologie des Englischen und Deutschen. Kritische Anmerkungen zu einem Buch von John A. Hawkins. In Gnutzmann, C. (ed.). *Kontrastive Linguistik.* Forum Angewandte Linguistik. Band 19. Frankfurt/M.: Peter Lang Verlag. pp.133-152.

Steiner, E. and Teich, E. 2004. Metafunctional profile of the grammar of German. In Caffarel, A., J.R. Martin and C.M.I.M. Matthiessen (eds). 2004. *Language Typology. A Functional Perspective.* Amsterdam: Benjamins.

Toutanova, K., Klein, D., Manning, C.D. and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL-2003*, pp. 252-259.

Weischedel, R. and Brunstein, A. 2005. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia.

Procedures for automatic corpus enrichment with abstract linguistic categories

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
e.lapshinova@mx.uni-saarland.de

Stefania Degaetano-Ortlieb
Universität des Saarlandes
s.degaetano@mx.uni-saarland.de

Hannah Kermes
Universität des Saarlandes
h.kermes@mx.uni-saarland.de

Elke Teich
Universität des Saarlandes
e.teich@mx.uni-saarland.de

1 Introduction: aims and motivation

In this paper, we describe procedures to semi-automatically enrich corpora with abstract linguistic categories based on extracted information profiting from the results of corpus-linguistic methods (extraction, analysis and interpretation).

Using lexico-grammatical patterns to extract evidence from text corpora is a common procedure in corpus-based linguistic analysis, e.g., a particular sequence of part-of-speech tags. These patterns are realisations of more abstract linguistic categories, such as voice, modality, conjunctive relations, evaluative expressions, etc. It is this level of abstraction that grammarians or lexicologists are typically interested in.

Existing studies on automatic extraction of abstract categories either focus on the extraction of one particular category, e.g., discourse connectives by Hutchinson (2005) or Bestgen et al. (2006) or serve the task of information extraction, e.g. extraction of semantic relations in Thanopoulos et al. (2000) or subjectivity in Riloff and Wiebe (2003), and operate with categories not aimed for linguistic analysis as such. Even though some approaches use automated methods to annotate abstract linguistic categories, e.g. in Peldszus et al. (2006) or Stede and Heintze (2004), most existing corpora which contain annotation of grammatical categories can be searched for in terms of one particular category only, e.g. discourse connectives in Penn Discourse Tree Bank (PDTB; Prasad et al. 2008), coreference and information structure in Potsdam Commentry Corpus (PCC; Stede 2004), which restrict analyses to a limited number of research questions. Moreover, to our knowledge, none of the existing approaches combine annotation procedures with extractions.

The procedures are basically theory independent and can be applied in various theoretical contexts. In this paper, we provide a basic method of corpus annotation in terms of abstract linguistic categories, illustrating it with an example to support analysis of register variation.

2 Corpus and linguistic research agenda

To exemplify our method, we describe examples of annotations based on our own research. Here we look at different linguistic features that are potentially "register-forming", i.e. they cluster in particular ways giving rise to different registers, cf. Halliday and Hasan (1998). We obtain candidate features from register theory, see for instance, Quirk et al. (1995), Halliday and Hasan (1989) or Biber (1995). These are mostly lexico-grammatical features, ranging from features at sentence, clause or phrase level to lexical and morphological levels, e.g., conjunctive relations, evaluative patterns and passive voice.

For our analysis, we use SciTex, a diachronic corpus of scientific research articles (see Teich and Fank 2010; Degaetano et al. *forthcoming*), including a total of ca. 34M tokens. The corpus is tagged with tokens, lemmas, parts-of-speech and sentence boundaries using the TreeTagger (Schmid 1994). Moreover, the corpus includes annotations of document structure (sections, paragraphs, etc.), as well as text and register boundaries. The corpus is encoded for use with the Corpus Query Processor (CQP) (Evert 2005).

3 Extraction and annotation procedures

The technical framework we use is the IMS Corpus Workbench (CWB) that provides an environment for encoding, storing and querying of text corpora with complex regular expressions (with CQP). The advantage of this framework is that the same formalism can be used for annotation and querying: there is no conceptual difference between extraction queries and annotation rules other than their function.

We start with a set of candidate lexico-grammatical patterns and formulate them as CQP queries. The queries contain regular expressions based on the tags already available in the corpus: tokens, parts-of-speech, sentence boundaries, etc.

We determine relevant features based on statistical methods, e.g. correspondence analysis (Degaetano et al. 2012). Relevant features are

then annotated in the corpus. Additionally, we evaluate the extracted results for coverage and precision.

For the annotation process, we use the CWB Perl-modules to apply the query rules, post-process and annotate the extracted results as shown in Figure 1.

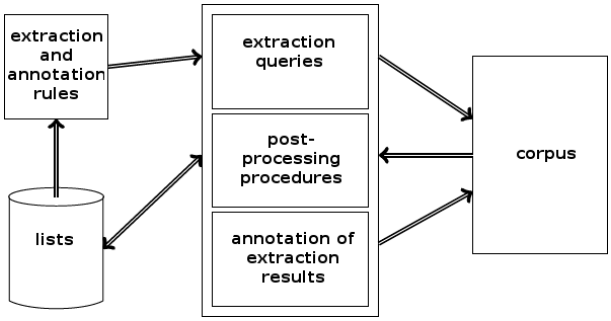


Figure 1. Annotation process

This annotation procedure is derived from the methods used for the YAC chunker (Kermes 2003). Thus, as mentioned above, we use the same rules for extraction and annotation of relevant features, cf. Figure 1.

Post-processing procedures are used to classify the results, determine characteristics (linguistic properties) and/or restrict the results to a subset. The linguistic property is then added to the annotation of the structure. For instance, an evaluative pattern such as *it+be+*adjective+*that/to* can have three different evaluative meanings (*obviousness*, *possibility*, *importance*) depending on the lexical item in the position of the adjective (e.g. *it is clear/obvious/evident that/to:* obviousness; *it is possible/impossible/likely that/to;* possibility; *it is important/necessary/relevant that/to:* importance), cf. Hunston (2004). The lexical items are contained in specific word lists which are applied to classify the results.

We can also apply contextual restrictions to patterns, e.g. sentence position. Consider conjunctions at sentence beginning relevant in terms of cohesive relations. To extract this kind of conjunctions, the sentence tag is needed in the query, as our search is restricted to conjunctions that occur at sentence beginning only. This constraint allows us to achieve highly precise results, as we are only interested in conjunctions which link sentences or clauses and not phrases. In the majority of cases, these are conjunctions occurring at sentence beginning (cf. Lapshinova et al. f*orthcoming)*.

The advantage of this method is that annotation rules can be left unspecified, as the post-processing has control over the results which can be classified, filtered, and changed if necessary. Lexical information can be included easily without multiplying the annotation rules.

4 Examples of annotations in SciTex

The first example is about additive conjunctions (1 in Table 1). According to Halliday and Hasan (1976), there are four major categories of cohesive conjunctions based on logico-semantic relations they express in a text: *additive* (expressed by conjunctions *and, moreover, in addition,* etc.), *adversative (*expressed by *but, however, although,* etc.), *temporal* (expressed by *first, then, finally,* etc.) and *causal* (expressed by c*onsequently, therefore,* etc.). We generate lists for these categories according to works by Halliday & Hasan (1976), Martin and Rose (2003), Quirk et al. (1985) and Biber et al. (1999) among others. The annotation of these categories in a corpus is based on lexical lists ($additive, $adversative, $temporal, $causal) included into a query. These lexical lists are then used in queries rules to annotate the respective category. We also add a contextual element – sentence start (see the element <s> in Table 1). The extracted results do not need further processing and can be directly annotated in the corpus.

	Query blocks	**description**	**examples**
1	<s> [lemma=$additive]	sentence start conjunctions from 'additive' list'	*And, In addition,* etc.
2	[lemma="it"] [lemma="be"] [pos="JJ."] [word="that|to"]	*it* verb *to be* adjective *that/to*	*It* *is* *important* *that*
3	[pos="MD"]?|([pos="V(B|H)(Z|P|D)"][pos="TO"])[pos="RB.*|JJ"]{0,3}[pos="VB"][pos="RB.*|JJ"]{0,3}@[pos="VVN"] [word!="by"]	optional modal auxiliary verb *to* optional adverbs verb *to be* optional adverbs past participle no *by*	*was* *thus also* *raised*

Table1. CQP query for conjunctions, evaluative patterns and passives

The second example (2 in Table 1) shows recurring word sequences that convey evaluative meanings, e.g. the evaluative pattern *it+be+adjective+that/to.* Depending on the adjective (JJ in the Table – Penn Treebank tagset) filler, the pattern expresses different meanings, e.g. importance or obviousness. Thus, if the adjective in Table 1 is filled with *important*, the meaning of the pattern is associated with the category of importance. The classification into meaning types is performed as a post-processing step, in which the adjective fillers are checked

against lexical lists to determine the respective category (adjectives expressing obviousness, possibility or importance).

The third example (3 in Table 1) demonstrates a more complex query that contains a number of morpho-syntactic restrictions (formulated with the help of part-of-speech tags). It is designed to extract cases of *short passives*, which require context information for the identification. As the extracted structure exceeds the target annotation structure, we add a '@'-tag to mark the range of the structure to be annotated, cf. Table 1.

The abstract linguistic categories are annotated in the corpus as structural attributes with features specifying the specific subtype of each category (cf. Table 2).

feature	subtype		example
conj	type	additive	*In addition*
		adversative	*On the other hand*
		causal	*Therefore*
		temporal	*From now on*
evaluation	meaning	importance	*it is necessary to*
		possibility	*it is possible to*
		complexity	*It is easy to*
		obviousness	*It is clear that*
passive	type	finite	*has been proved*
		non-finite	*to be proved*
	length	short	*can be evaluated*
		long	*is dictated by*

Table 2. Annotated features and their subtypes

The annotated linguistic information is easily accessible with the queries /region[conj], /region[evaluation] or /region[passive]. The annotated subtypes are available as "substructures" (conj_type, evaluation_meaning, passive_type, passive_length). Annotating this kind of abstract linguistic categories makes them more easily available for further extraction processes. The distribution of different meanings and types across registers, e.g., can be extracted without having to use complex patterns and lexical lists. Thus, the annotated information can be used for the extraction of more complex linguistic phenomena.

5 Evaluation of procedures

As a full quantitative evaluation of our procedures would require the preparation of a gold standard corpus, we carried out a partial evaluation of selected data samples. For passives, we achieve a precision of 99% for identification and classification. Conjunctive relations can be identified with 100% precision. Their classification is more challenging because of ambiguous cases, e.g., *while* is ambiguous between the adversative or temporal meaning, or *since* can be either causal or temporal. We leave such cases underspecified, annotating their ambiguity (as a feature set) and leave the disambiguation to a further processing step (manual or automatic).

6 Conclusion and discussion

We described a method of semi-automatic annotation of commonly used linguistic categories based on the extraction of lexico-grammatical patterns. We illustrated this with the annotation of categories relevant for register analysis, although it can be used for other categories regardless a theory in the background.

For other corpora, our procedures can be similarly beneficial but less time-consuming as manual annotation of the described categories. In contrast to manual procedures of the described categories, they enable consistent classification of phenomena under analysis into defined categories.

There is a rising interest to use corpus linguistic methods in a wide range of disciplines, for which corpora can provide evidence on the usage of language, on the distribution of certain phenomena shedding light on what is more common in what context or domain. While for some phenomena, simple queries on lemma and part-of-speech are sufficient, other phenomena require more complex queries, e.g. example (3) in Table 1 above. In this case, the annotation can help to simplify the queries, to investigate more complex phenomena as part of the contextual or linguistic information already encoded in the corpus and to make the extraction process more efficient.

References

Bestgen, Y., Degand, L. and Spooren, W. 2006. *Discourse Processes*, Vol. 41, issue 2, pp. 175-193.

Biber, D. 1995. *Dimensions of Register Variation. A Cross Linguistic Comparison.* Cambridge University Press, Cambridge.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*, London, Longman.

Degaetano-Ortlieb, S., Lapshinova-Koltunski, E. and Teich, E. 2012. Feature Discovery for Diachronic Register Analysis: a Semi-Automatic Approach. In *Proceedings of LREC-2012*, Istanbul, Turkey.

Degaetano-Ortlieb, S., Kermes, H., Lapshinova-

Koltunski, E. and Teich, E. forthcoming. Scitex – a diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, (eds.) *New Methods in Historical Corpus Linguistics, 'Corpus Linguistics and Interdisciplinary Perspectives on Language'* – CLIP, Vol. 3. Narr, Tübingen.

Evert, S. 2005. *The CQP Query Language Tutorial.* IMS Stuttgart. CWB version 2.2.b90.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. Longman, London, New York.

Halliday, M.A.K. and Hasan, R. 1989. *Language, context and text: Aspects of language in a social-semiotic perspective.* Oxford University Press, Oxford.

Hunston, S. 2004. Counting the uncountable: problems of identifying evaluation in a text and in a corpus. In *Corpora and Discourse*, Peter Lang, pp. 157-188.

Hutchinson, B. 2005. *The Automatic Acquisition of Knowledge about Discourse Connectives.* Ph.D thesis. Informatics thesis and dissertation collection. University of Edinburgh.

Kermes, H. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography.* Ph.D. thesis, IMS, Universität Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Lapshinova-Koltunski, E., S. Degaetano, E. Teich and H. Kermes. forthcoming. Linguistic evolution of conjunctive relations in emerging scientific registers. In F. Poppi and W. Cheng (eds.). *The three waves of globalization: winds of change in Professional, Institutional and Academic Genres. Proceedings of Clavier-11*. Cambridge Scholars Publishing.

Martin, J. R. and D. Rose. 2003. *Working with Discourse. Meaning beyond the clause.* New York and London, Continuum.

Peldszus, A., Herzog, A., Hofmann, F. and Stede, M. 2008. Zur Annotation von kausalen Verknüpfungen in Texten. In *Proceedings of the KONVENS Conference, companion volume*, Berlin, Germany.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, N. 2008. Penn Discourse Treebank Version 2.0. In *Proceedings of LREC-2008*, Marrakesh, Morocco.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Riloff, E. and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, Sapporo, Japan.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceeding of International Conference on New Methods in Language Processing*, Manchester, UK.

Stede, M. and Heintze, S. 2004. Machine-assisted Rhetorical Structure Annotation. In *Proceedings of the International Conference on Computational Linguistics (COLING-2004)*, Geneva, Switzerland.

Stede, M. 2004. The Potsdam Commentary Corpus. In *Proceedings of the ACL-2004 Workshop on Discourse Annotation*, Barcelona, Spain, pp. 96-102.

Thanopoulos, A., Fakotakis, N. and Kokkinakis, G.K. 2000. Automatic Extraction of Semantic Relations from Specialized Corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany, pp. 836-842.

Teich, E. and Fankhauser, P. 2010. Exploring a corpus of scientific texts using data mining. In S. Gries, S. Wulff, and M. Davies (eds.) *Corpus-linguistic applications: Current studies, new directions*. Rodopi, Amsterdam and New York, pp. 233-247.

The correlation between lexical core index, age-of-acquisition, familiarity and imageability

John Hanhong Li
Open University of Hong Kong
typetoken@gmail.com

1 Introduction

Core vocabulary has been a long-time concern for lexicographers, language teachers and other linguists. Many researchers have realized the existence of core vocabulary in English studies and defined core vocabulary by a variety of criteria. Ogden's (1930) concept of Basic English, the vocabulary control movement in the 1920s and 1930s, and the proposal of culture-free Nuclear English for international use (Quirk 1982; Stein 1979) all boosted the research into core vocabulary.

Past research tends to focus on word frequency in written texts. Modern corpora make it possible for us to explore core vocabulary and their frequency in both of the written and spoken texts (Bullon and Leech 2007; Leech et al. 2001). Word frequency distributed in different sections of a large corpus has also been taken into account (Davies and Gardner 2010).

Moreover, with the development of more corpus resources and more thorough tagging features in modern corpora (e.g. BNC XML), recent studies have integrated other factors such as age, genre and distributed word frequency into the study of core vocabulary (Li 2010, Li and Fang 2011a). Core vocabulary includes not simply those words with high frequency but also those distributed widely in different age groups and genres as displayed in large corpora (Li 2010, Li and Fang 2011a).

Based on this understanding, this research aims to calculate a core index of each word in a large corpus and further explore the relations between a word's core index and its psycholinguistic features: age-of-acquisition, familiarity and imageability. We hope this can provide a deeper perception of the linguistic features of vocabulary and provide language learners with better pedagogy.

2 Dispersion Index

Recently, scholars have realized that a word's raw frequency (total occurrences) in a corpus is not reliable for selecting core vocabulary and dispersion indexes should play a role (Li 2010, Li and Fang 2011a). Gries also points out that "such frequencies in isolation may sometimes be misleading since they do not take into consideration the degree of dispersion of the relevant linguistics variable" (Gries 2008:403). He further reviews the previous dispersion measures and suggests a new dispersion measure so as to raise our awareness of the importance of dispersion.

However, the dispersion-based approach is not sufficient or perfect enough for the selection of core vocabulary. Neither Juilland's *D* nor Carroll's *D* can discriminate the evaluation of the following word frequency dispersion in a five-category corpus: Word A = 1000, 0, 0, 0, 0[1] and Word B = 2, 0, 0, 0, 0. For Word A and Word B, both of their dispersion indexes (*D*-value) are 0 whether in terms of Juilland's *D* or Carroll's *D*. It is the same situation for the frequency distribution of Word C (Word C = 100, 100, 100, 100, 100) and Word D (Word D = 2, 2, 2, 2, 2) where both dispersion indexes are 1. For Word A and Word B, the discrepancy between their total frequencies (1,000 vs. 2) is so huge that we cannot say these two words are equally important or equally core despite the fact that their dispersion indexes are equal (*D* = 0). For Word C and Word D, neither can we conclude that Word C and Word D are of the same importance even if they are both dispersed perfectly evenly (*D* = 1) because the total frequency difference between Word C and Word D is big (500:10). In other words, dispersion index alone cannot solve these problems in selecting core vocabulary.

In our present research, we need to quantify the core degree of each word. We prefer Carroll's *U* (Usage Coefficient) over dispersion indexes to quantify a word's core degree in that Carroll's *U* (Carroll 1970, 1971) not only integrates word frequency but also dispersion information of a word in different sections of a corpus. Carroll's *U* was adopted as a criterion for ranking words in Carroll, Davies and Richman's (1971) *The American Heritage Word Frequency Book* and Francis and Kučera's (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Recently, Carroll's *U* has been adopted as the distributed frequency which blends frequency and dispersion index as a whole for the selection of core vocabulary, and it turns out that Carroll's *U*-based approach can select core vocabulary lists with larger cumulative coverage than raw frequency-based approach (Li 2010b, Li & Fang 2011a).

[1] Word A = 1000, 0, 0, 0, 0 indicates that word A appears in five sections of a corpus with 1000 occurrences in the first section and 0 in the remaining sections.

3 Lexical Core Index

In order to calculate the core index for words in a large corpus with the tagging information of age and genre, we have to build up a corpus of balanced Age-Groups and a corpus of Genres. We mainly base our study on the British National Corpus XML Edition (BNC XML 2007). To ensure the enough data for the balanced Age-Group Corpus, we combine BNC XML with other corpora containing age information such as Lucy Corpus (LUCY), Lancaster Corpus of Children's Project Writing (LCCPW), the Polytechnic of Wales Corpus (EPOW), the Child Language Data Exchange System (CHILDES). [1] All these supplementary corpora have been retagged by CLAWS4 C5 Tagset.[2]

With the setup of the above Age-Group Corpus and Genre Corpus, we adopt Carroll's Usage Coefficient (*U*) to calculate a word's core index in different age groups and text genres for evaluating the core degree of words (Carroll 1970, 1971).

4 MRC Psycholinguistic Database

Age-of-acquisition (AOA), imagery (or imageability) and familiarity are the major psycholinguistic attributes of words. The MRC Psycholinguistic Database 2.0 is a collection of data from a variety of researches in the psychological and linguistic descriptions of word attributes which are helpful for psycholinguistic studies and researches in artificial intelligence and computer science (Wilson 1988) .3 It includes the subjective rating results of age-of-acquisition, imagery, concreteness, familiarity, and meaningfulness (Gilhooly and Logie 1980; Paivio et al. 1968; Stadthagen-Gonzalez and Davis 2006; Toglia and Batting 1978).

5 Age-of-Acquisition

Age-of-Acquisition (AOA) "refers to the age at which a word was learned" (Stadthagen-Gonzalez and Davis 2006:598) and has been studied as a significant factor to language and memory processes (e.g., Carroll and White 1973; Hirsh and Funnell 1995; Juhasz and Rayner 2003; Morrison et al. 1992; Roodenrys et al. 1994). Previous studies show that age-of-acquisition has a unique influence on word recognition performance such as word naming and lexical decision (e.g., Bonin et al. 2004; Brown and Watson 1987; Morrison and Ellis 1995). Some studies propose that the effect of age-of-acquisition reflects the effect of cumulative frequency (e.g., Ghyselinck et al. 2004; Stadthagen-Gonzalez et al. 2004; Stadthagen-Gonzalez and Davis 2006; Zevin and Seidenberg 2002, 2004) since early acquired words will be encountered more often and "also have a much larger cumulative frequency of exposure across the lifetime" (Balota et al. 2006: 316).

A wordlist with the information of age-of-acquisition is produced from the MRC Psycholinguistic Database and we explore the correlation of these words' core indexes and age-of-acquisition.

It turns out that *mummy* is the earliest acquired word, which is also the most frequently used nouns by children and mothers (Li and Fang 2011b). This word ranks 1,089th in terms of core index in the combined Age-Genre Corpus of around 20 million words. The core degree is high for *mummy*, who is the first person a child gets in touch with. Things that are necessary for life and encountered frequently in daily life will be acquired earlier (e.g., *shoe*, *toilet*, *dinner*, *bread*, *milk*, *meat*). What children often do and say will also be acquired earlier (e.g., *kiss*, *bath*, *goodbye*, *mummy*). Things that children often see and play with will be acquired earlier (e.g., *seesaw*, *tree*, *moon*, *colour*, *grass*, *star*, *sheep*, *pony*, *bike*). The parts of body are also acquired earlier (e.g., *finger*, *tongue*). The feelings, sentiments and physical and psychological state children often experience will be acquired earlier (e.g., *hungry*, *funny*, *thirsty*, *sick*).

The result shows that there is a significant negative correlation between the core index and age of acquisition ($p < .01$, $r = -.167$). That implies core words are usually acquired earlier in our lifetime. The more core a word is, the earlier it will be acquired in the language development. Generally speaking, the significant negative correlation between the core index and age of acquisition suggests that core words tend to be acquired earlier in language acquisition. Those words acquired earlier in early childhood are more possibly developed as core words.

6 Familiarity

Familiarity rating has been interpreted as a measure of the frequency of exposure to a word (Stadthagen-Gonzalez and Davis 2006) and familiarity is a better predictor of word

[1] For these supplementary corpus data and the removing of their tags, I need to thank Prof. Geoffrey Sampson, Prof. Tony McEnery, Prof. Eric Atwell, Prof. Clive Souter and Prof. Brian MacWhinney.

[2] *http://ucrel.lancs.ac.uk/claws/trial.html* I need to thank Dr. Paul Rayson's suggestions and help in the process.

[3] *http://www.psych.rl.ac.uk/MRC_Psych_Db.html*
I would like to acknowledge Prof. Michael Wilson's kind help who generously made the full word lists in the MRC Psycholinguistic Database downloadable upon my request.

performance than printed word frequency, particularly for low-frequency words (Gernsbacher 1984). In order to study the correlation between the core index and familiarity of a word, we downloaded a word list with familiarity values from the MRC Psycholinguistic Database. Then we use SPSS to explore the correlation between these words' core indexes and familiarity indexes.

The result shows that familiarity and core index are significantly correlated with each other ($p < 0.01$, $r = .097$). That suggests the familiarity can be a predictor of core vocabulary. If a word is familiar to everyone, it should be used by people from different age groups whether they are young or old. Other research also demonstrates that familiarity norms correlate strongly with age-of-acquisition (Stadthagen-Gonzalez and Davis 2006). Moreover, a familiar word should appear in a variety of text genres. Therefore, a familiar word tends to be highly frequent and widely dispersed across age groups and text genres.

What our research contributes is that not only raw frequency, as found in previous research, but also distributed frequency (Carroll's U) is significantly correlated with familiarity.

7 Imageability

Imageability or imagery rating is about how easily a word can arouse mental pictures or images (cf. Gilhooly and Logie 1980; Paivio et al. 1986; Stadthagen-Gonzalez and Davis 2006).

In order to explore the correlation between the coreness of a word and its imageability, we first searched and downloaded words with imageability values from the MRC Psychological Database. The correlation between their core indexes and imageability values is evaluated by SPSS.

The result shows that the core indexes are significantly correlated with imageability values ($p < 0.01$ $r = -.108$). The significant negative correlation between imageability and core index (U value) implies that words with high imageability tend to have lower distributed frequency. That is to say, a highly imageable word might not be spread widely or used widely in different situations such as age groups or genres.

Take *skunk* as an example. Though it ranks 7th in terms of imageability in the MRC Psycholinguistic Database, which means that it is very imageable and arouses a picture very readily in people's mind. However, its core index or distributed frequency ranks 29,090th in our Age-Genre Corpus of around 20 million tokens. In other words, despite its high imageability, this word is not very core. Moreover, the dispersion index (D) of *skunk* is 0.00 in the Age-Group Corpus and it is 0.22 in the Genre Corpus. That means this word is not widely used.

All in all, highly imageable words are not necessarily the words with high core indexes. The above examples and analysis reveal that highly imageable words do not necessarily scatter through all kinds of text genres or will be frequently used by each age group. Children tend to use many more words of animals which are truly imageable but might not be used frequently by other age groups.

It is probably that core words contain more meaning potentials. In modern lexicography, word meaning is no longer static as defined in traditional dictionaries. With the help of large modern corpora and the development of the theory of Corpus Pattern Analysis (CPA) which treats language as pattern-driven (Hanks 2008), meanings of words are events instead of "static abstract entities" (Hanks 2010). Sinclair (1991:101-102) regards that "frequent words have, in general , a more complex set of senses than infrequent words" and "some words are much more common than others, some senses of one word are much more common than other senses of the same word". He proposes to "discover a statistical relationship between the number of occurrences of a word and the number of different senses it realizes" (1991:101).

If core words have more meanings than non-core words, the imageability of core words might be reduced due to the fact that there are so many meaning potentials for core words that their meanings are not so easy to anchor. Real meanings of words lie in word use between speakers and audience, meaning events and collocated patterns in actual use (Hanks 2010, 2012). In this case, we could explain why the correlation between core index and imageability are significantly negatively correlated. The more core a word is, the more meaning potentials it carries, the less imageable it becomes. Further experiments are needed to give more insights into this expectation.

Generally speaking, the significant negative correlation between core index and imageability demonstrates that highly imageable words tend to show lower distributed frequency or core indexes (U values). Therefore, imageability is not a good indicator of a word's core degree. Moreover, a word's dispersion index could be very low despite its high imageability.

8 Summary

The exploration of the correlations between a word's core index, familiarity, imageability and

age-of-acquisition reveals that familiarity and age-of-acquisition become good indicators of core words, core words tend to be acquired earlier in language acquisition and hence more familiar, but words of high-imageability are not necessarily core words. Moreover, the result makes us infer that core words tend to be lower in imageability because of their rich senses and heavy-loaded meaning potentials. Further studies need to yield more evidence for this.

References

Balota, D. A., M. J. Yap and M. J. Cortese. 2006. Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler and M. A. Gernsbacher (Eds.) *Handbook of Psycholinguistics* (2nd ed.), 285-376. London: Academic Press Elsevier.

Bonin, P., C. Barry, A. Méot and M. Chalard. 2004. The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory & Language* 50: 456–476.

Brown, G. D. A. and F. L. Watson. 1987. First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition* 15: 208–216

Bullon, S. and G. Leech. 2007. Longman Communication 3000 and the Longman Defining Vocabulary. In S. Bullon, G. Leech and J. Harmer (Eds.), *Longman Communication 3000*, 1-7. Harlow: Pearson Education Limited.

Carroll, J. B. 1970. An alternative to Juilland's Usage Coefficient for lexical frequencies and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior* 3: 61–65.

Carroll, J. B. and M. N. White. 1973. 'Word frequency and age-of-acquisition as determiners of picture-naming latency', *Quarterly journal of experimental psychology*, 25: 85-95.

Carroll, J. B., P. Davies and B. Richman. 1971. *The American heritage word frequency book*. Boston: Houghton Mifflin.

Davies, M and D. Gardner. 2010. *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. London and New York: Routledge.

Francis and Kučera. 1982. *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Gernsbacher, M. A. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113: 256-281.

Ghyselinck, M., M. B. Lewis and M. Brysbaert. 2004. Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica* 115: 43-67.

Gilhooly, K. J. and R. H. Logie. 1980. Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behavior Research Methods & Instrumentation* 12(4): 395-427.

Gries, Th. S. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4): 403-437.

Hanks, P. 2008. Lexical patterns: From Hornby to Hunston and beyond. In E. Bernal and J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress*, 89-129. Barcelona: Universitat Pompeu Fabra. Available online at http://www.patrickhanks.com/Hornby_to_Hunston.pdf

Hanks, P. 2010. How people use words to make meanings. In B. Sharp and M. Zock (Eds.), *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* (*NLPCS 2010 Proceedings)*, 3-13. Funchal Madeira, Portugal. Available online at http://www.patrickhanks.com/HowPeopleUseWordsToMakeMeanings.pdf

Hanks, P. 2012 forthcoming. *Lexical Analysis: Norms and Exploitations*. Massachusetts: MIT Press.

Hirsh, K. W. and E. Funnell. 1995. Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal of Neurolinguistics* 9: 23-32.

Juhasz, B. and K. Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 1312-1318.

Leech, G., P. Rayson and A. Wilson. 2001. *Word frequencies in written and spoken English based on the British National Corpus*. London: Longman.

Li, H. 2010. Word frequency distribution for electronic learner's dictionaries. In S. Granger and M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications*, 217-228. Louvain-la-Neuve: Presses universitaires de Louvain.

Li, H. and A. C. Fang. 2011a. Age tagging and word frequency for learner's dictionaries. In John Newman, Sally Rice, and Harald Baayen (eds.), *Corpus-based studies in language documentation, use, and learning*, 157-173. Amsterdam: Rodopi Press.

Li, H. and A. C. Fang. 2011b. Word frequency of the CHILDES corpus: Another perspective of child language features. *ICAME Journal* 35: 95-116.

Morrison, C. M., A. W. Ellis and P. T. Quinlan. 1992. Age of acquisition, not word frequency, affects

object naming, not object recognition. *Memory & Cognition* 20: 705-714.

Morrison, C. M. and A. W. Ellis. 1995. Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory and Cognition* 21: 116–153.

Ogden, C. K. 1933. *Basic English: A general introduction with rules and grammar* (4th ed.). London: Paul Treber & Co., Ltd.

Paivio, A., J. C. Yuille and Madigan S. A. 1968. Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement* 76(1) part 2: 1-25.

Quirk, R. 1982. International communication and the concept of nuclear English. In R. Quirk, *Style and Communication in the English Language*, 37-53. London: Edward Arnold Ltd.

Roodenrys, S., C. Hulme, J. Alban, A. W. Ellis, and G. D. A. Brown. 1994. Effects of word frequency and age of acquisition on short-term memory span. *Memory & Cognition* 22: 695-701.

Sinclair, J. 1991. *Corpus concordance collocation*. Oxford, Oxford Univerisity Press.

Stadthagen-Gonzalez, H. and C. J. Davis. 2006. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* 38(4): 598-605.

Stadthagen-Gonzalez, H., J. S. Bowers and M. F. Damian. 2004. Age-of-acquisition effects in visual word recognition: Evidence from expert vocabularies. *Cognition* 93: B11-B26.

Stein, G. 1979. Nuclear English: Reflections on the structure of its vocabulary. *Peoetica* 10: 27-52.

Toglia, M. P. and W. F. Batting. 1978. *Handbook of semantic word norms*. New York: Erlbaum.

Wilson, M. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Bahvior Research Methods, Instrument & Computers* 20 (1): 6-10.

Zevin, J. D. and M. S. Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language* 47: 1-29.

Zevin, J. D. and M. S. Seidenberg. 2004. Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition* 32: 31-38.

Phraseological discourse actors in English academic texts

Jingjie Li
Donghua University
lijingjie@dhu.edu.cn

Wenjie Hu
Shanghai Jiao Tong University
huwenjie@sjtu.edu.cn

1 Introduction

Austin (1962) raises the concept of speech acts and divides it into three categories, i.e. locution, illocution and perlocution. Searle (1976) further breaks illocution into five parts, i.e. representatives, directives, commisives, expressives and declarations. Both have revealed one important nature of language use: to do things with words. Following this idea, we argue that at a macro-level, members of a certain discourse community take academic texts as a means to exchange ideas, report research results and foster disciplinary development. However, since academic texts have a regular discourse pattern, to comply with this pattern, researchers have to use certain language forms to perform specific acts. These acts are discourse acts. To take substantiation for an example, all academic authors need to specify and justify their views, facts, research findings, etc., then substantiation is a discourse act of academic texts. To perform this discourse act, researchers usually use a sequence as in the following example:

Although only the combined effect significantly slows down processing, the data in Table 4 suggest that both separate effects slow down detection times.

Usually, discourse acts in academic texts differ with disciplines, but some discourse acts are common in all disciplines. Through an observation of the text samples in NEW-JDEST, seven main subclasses are distinguished in this study, depending on their main contributions to discourse acts: presenting views or facts, reporting, text deixis, announcing research features, outlining purposes, presenting results, and focusing. It should be noted, however, that the seven subclasses are by no means exhaustive but the most typical ones in academic texts.

2 Research design

The second generation of JDEST corpus serves as the database of this study, henceforth NEW-JDEST. It contains 6,831,693 word tokens and 208,410 differing types. Altogether 1,202 text

samples are collected in the corpus. The length of each text varies from 181 to 33,341 words.

Specifically, data processing was conducted in the following steps. We first developed a new computing method for extracting contiguous multi-word sequences from a formidably large quantity of data by measuring the internal association of each of the lexical segments. The whole process is completed by pure computer automation, without any human intervention. Then we manually checked the extracted data, removing the disturbing sequences that are both structurally and semantically incomplete and assigning principal discourse functions to each of the finally identified Phraseological Discourse Actors (henceforth PDAs). Third, we generalized the most prominent phraseological patterns and structures to realize each academic discourse act, for which in-depth analysis was carried out within the corpus-driven paradigm.

Altogether this study has finally identified 4,401 varied types of PDAs from NEW-JDEST, with a total of 115,233 occurrences, as shown in Table 1.

N-grams	Types	(%)Type percentage	Tokens	(%)Token percentage
2-grams	1520	34.54	54508	47.30
3-grams	2014	45.76	48274	41.89
4-grams	779	17.70	11667	10.12
5-grams	77	1.75	703	0.61
6-grams	11	0.25	81	0.07
Total	4401	100	115233	100

Table 1. Frequency distribution of PDAs

3 Discussion and summary

This paper focuses on the first five main discourse acts, i.e. presenting views or facts, reporting, and text deixis, as space lacks for a detailed description of all. Table 2 presents the most prominent patterns and lexical realizations for the three discourse acts.

It should be noted, however, that all discourse acts have their own actors. Some of the actors are single words and some are multi-word sequences, with the latter being the focus of this study. It is these phraseological discourse actors that enable us to distinguish academic texts, which have their own phraseological characteristics, from other types of texts (e.g. general English texts).

Besides, this research has also uncovered a great deal of linguistic facts in academic texts which unequivocally point to the underlying regularity. Co-selection is at the very heart of choosing linguistic forms to realize meaning, which inevitably result in the large number of recurrent multi-word sequences in academic texts. To take the PDAs for announcing present research as an example, they have occurred in two main patterns: the collapsed structure (1) "This PAPER Vs" and the standard descriptive form (2) "AC I/we MODAL V". Each pattern has its own set of preferred lexical realizations, such as the high recurrence of *this paper is concerned with* (in Pattern (1)) and *in this paper, we expect to* (Pattern (2)) in NEW-JDEST versus the zero occurrence of *this paper expects to* and *in this paper, we are concerned with*. These co-selections between lexis and structure show the preferred ways of saying things (Granger & Paquot 2008, Stubbs 2009), i.e. the conventionality of language use. The data in this study suggest that language users resort largely to the conventional means for expressing related meanings, not only in spontaneous ongoing communication where speakers are in constant need of easily retrieved expressions to save time, as suggested in Pawley and Syder (1983), Altenberg (1998), Wray (2002), etc., but also in the more deliberated production of academic texts where authors have adequate time to ponder over every expression for the sake of accuracy and effectiveness.

Discourse acts	Patterns
1. Presenting views/facts	(1) DET DATA V (that) *data suggest (that)* *figure # illustrates* (2) It (MODAL) BE V-ed (that) *it follows that* *it is shown that* (3) We (MODAL) V (that) *we can see that* *we assume that*
2. Reporting	(1) AUTHOR (YEAR) V-tense (that) *AUTHOR argue/s that* *AUTHOR found that* (2) It BE-tense V-ed (that) *it has been shown that* *it is said that*
3. Text deixis	(1) as (MODAL) V-ed Ac *as shown in table #* *as described previously* *as demonstrated by AUTHOR*
4. announcing research features	(1) This PAPER Vs *this paper describes* *Chapter # discusses* (2) AC I/we MODAL V *In this paper, we will discuss* *In this section, we will explore*
5. outlining purposes	(1) The (ADJ) PURPOSE (of N) BE to-inf *the purpose of this paper is to* *the purpose of this study was to* *the aim of… is to*

Table 2. Phraseological patterns of PDAs

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In Cowie, A. P. (ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press. 101-122.

Austin, J. L. (1962). *How to Do Things with Words.* Mass: Harvard University Press.

Granger, S. & M. Paquot (2008). Disentangling the phraseological web. In Granger S. & F. Meunier (Ed.), *Phraseology: An interdisciplinary perspective*, 145-160. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Pawley, A. & F. H. Syder. (1983). Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In J. Richards & R. Schmidt (Eds.), *Language and Communication*, 191-225. London: Longman.

Searle, J. (1976). A Classification of Illocutionary Speech Acts. *Language in Society*, 5(1): 1-23.

Stubbs, M. (2009). The Search for Units of Meaning: Sinclair on Empirical Semantics. *Applied Linguistics*, 30/1: 115–137. Oxford University Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

China English Corpus construction on an open corpus platform

Li Wenzhong
Beijing Foreign Studies University
li.wzhong@gmail.com

While it is a fact that China has the largest population of English learners in the world, and that English is the compulsory subject from the 3rd grade in the primary schools in the major developed areas in China, China English as a new variety in the context of English as an International Language (aka English as a Lingua Franca, which may mean differently in some scholars' definitions (cf. Seidlhofer 2005; Jenkins 2012) has long been in controversial discussions in China. People's attitudes vary to a great extent, with a more variable continuum between the extremist YES and NO arbitration. Approached from sociolinguistic and sociocultural perspectives, China English serves as an alternative means in both intranational and international communication, and functions to establish in Chinese users of English their cultural identity and community integrity (Li 1993, 2006). The proponents associate the issue with the national strategic endeavor "Let China go to the world", arguing that English in its localized use could carry the weight of Chinese culture in the international interactions (ibid). To push the argument further, they expound that the term China English is not singular itself, but an umbrella term covering a wide range of variations of English use across the mainland and the related regions such as Hong Kong, Taiwan and Macao, thus the preferred term is 'Chinese Englishes' (Bolton 2006). But from the perspectives of English education, controversies arise as whether China English should be made part of the pedagogical praxis concerning its unsteady status and miscellaneous features. The opponents insist on the point that China English is still on its way to the full formation of a new variety (Xie 1994), and it is even coterminous in some scholars' mind's eye with Chinglish, a mix of Chinese and awkward English which is usually collected and presented in a playful manner both home and abroad. For most of the English teachers and teaching researchers, their worries and cautions are that promoting China English in English learning in China could degrade English teaching at large; the learners' English proficiency will also be affected, because if they are exposed to China English, which is regarded as deviation from

British or American standard English, they are likely to get their errors fossilized. However, most of the research literature on China English is more theoretical and attitudinal than empirical; very little research has been done based on the real data analysis, even less on the corpus data. In the past two decades there has been a flourishing development of various learner corpora such as the CLEC[1] (Gui and Yang 2003), the COLSEC[2] (Yang and Wei 2005), and the SWECCL[3] (Wen, Liang and Yan 2005), yet they are meant to diagnose learners' problems for remedial English instruction. In the research practice, what is prevailing is to determine 'overuse', 'underuse', or 'misuse' by contrasting the learners' usage in the corpora with that of the native speakers, be they British or American. And the rationale behind is that the learners are always wrong and the native speakers are always right, and such ungrounded assumption is not without criticism though (Li, 2007).

The China English Corpus[4] (shortened as CEC hereafter) is constructed as a solution to the scarcity of research data, endeavoring to describe both grammatical and lexical features salient in English texts produced in China. The objectives of constructing the CEC are therefore specified as: 1) to use the corpus as a test bed and on which to place the important hypotheses for verification and falsification; 2) to access the corpus data in a corpus-driven approach with a purpose of comprehensive description of the subject under examination; 3) to locate the basic attributes and properties of China English in a framework of English globalization and nativization and 4) to explore the implications of China English for its cultural expressions and English education. Coming down to the corpus construction, many considerations are taken into account: 1) how to delimit the population and sampling, namely what texts should be included in the corpus and how many? 2) the infrastructure of the corpus: should it be a stand-off or web corpus? 3) Do we need a monitor corpus? 4) How should the corpus be accessed and evaluated? Innovative ideas have been brought forward to meet the growing demands for the availability of corpus data and its capability to be dynamically updated. It is therefore argued that an open corpus, rather than a packed closed one, based on the web service and cloud computing, is of greater potentials for corpus research. Also under close examination are the related issues of corpus development, such as text representativeness, text typology, corpus annotation, automatic redrawing of the interface with XML, user constructing sub-corpus ad hoc, and data retrieval using regular expressions. An Open Corpus Platform is developed using Self-adapted Javascript+XML, as a web-based corpus building environment, for constructing the CEC. The platform is tailored to the users' needs and serves as an interface on which collected texts are made entry into a database with their meta-data automatically formatted as XML files. The interface can be customized by the users so that new items of meta-data can be added or deleted. The texts for the CEC are classified in terms of text type, domain, status (original or translation), and source, the information of which is encoded and recorded as meta-data in each text file. Up to now, the CEC has altogether 19, 091 individual texts with over 42 million running words, collected from 22 sources and of 14 text types in 10 domains and 56 sub-domains. The platform joins both corpus creation and corpus analysis software tools and aims to reducing to a large extent the technological complications in the corpus construction, and concentrating the users on data analysis proper. This paper serves as an introduction to and illustration of the newly constructed CEC, which opens up many possibilities for further research.

[1] Chinese College Learner English Corpus, as a subcorpus of CLEC, was completed at Shanghai Jiao Tong University in 1999.

[2] College Learner Spoken English Corpus was completed at Shanghai Jiao Tong University in 2005 .

[3] The Spoken and Written English Corpus of Chinese Learners was completed at Nanjing University in 2005.

[4] The research project was funded by the National Social Science Fund of China in 2007.

References

Bolton, K. 2006. *Chinese Englishes: A Sociolinguistic History*. Cambridge: Cambridge University Press.

Jenkins, J. 2012. "English as a Lingua Franca from the Classroom to the Classroom". *ELT Journal* 66 (4) (January 10): 486–494.

Li, W. 1993. "China English and Chinglish". *Foreign Language Teaching and Research* (4): 18-24.

Li, W. 2006. "The Humanistic Effects of the Globalization of English and Its Nativization in China". *Journal of Henan Normal University* 33(3): 131-134.

Li, W. 2007. "A Critical Review of CIA". *Computer-assisted Foreign Language Education* (127): 13-17.

Seidlhofer, B. 2005. "English as a lingua franca". *ELT Journal 59*(4): 339–341.

Wen, Q., Liang, M. and Yan, X. 2005. *Spoken and Written English Corpus of Chinese Learners*. Beijing: Foreign Language Teaching and Research Press.

Xie, Z. 1994. "China English: An Interfered Variety in

Intercultural Communication". *Shandong Foreign Languages Journal* 57(4): 63-68.

Gui, S. and Yang, H. 2003. *Chinese Learner English Corpus.* Shanghai: Shanghai Foreign Language Education Press.

Yang, H. and Wei, N. 2005. *Construction and Data Analysis of a Chinese Learner Spoken English Corpus.* Shanghai: Shanghai Foreign Language Education Press.

Sparing a *free hand*: context-based automatic categorisation of concordance lines

Maocheng Liang
Beijing Foreign Studies University
frankliang0086@163.com

1 Background

Early applications of corpora in lexicography (the COBUILD Project) had often depended on the so-called "coloured-pen method", in which "a lexicographer … read the corpus lines, identifying different meanings as they went along, assigning a colour to each meaning and marking each corpus line with the appropriate colour" (Kilgarriff & Koeling 2003). This time-consuming manual categorisation of concordance lines is described in great detail in Sinclair (2003), who was the leader of the COBUILD Project. What is brought to highlight in Sinclair (2003) is how meanings evolve from phraseological patterns and how meanings and patterns are co-selected, as discovered in concordance lines. Illustrating his point with examples such as *free hand*, *the naked eye*, Sinclair recommends some procedures "as a basic strategy for retrieving information from a corpus", in the hope of "motiv[ating] the reader to the pleasure of consulting a corpus at first hand" (Sinclair 2003:viii). However, as typical corpora "contain much more information than a human can handle", Sinclair (*ibid*.) admits "that the investigator can easily get swamped in a large quantity of heterogeneous data".

To tackle the fore-mentioned problem, some innovative work (*e.g*., O'Donnell 2008; Rayson 2002) has been done to make human investigation more cost-effective, either by annotating corpora for meaning, or by facilitating the identification of recurrent patterns.

This study reports work on an interactive tool, which enables the investigator to automatically categorise concordance lines for the co-selection of patterns and meaning.

2 The study

The interactive tool is designed and developed on the basis of the view that "the different senses of words will tend to be distinguished by different patterns" (Hunston & Francis 2000: 83), and that words are typically associated with each other to form distinctive meaningful units. One typical example illustrated in Sinclair (2003) is the phrase *free hand*, which can take on either a literal

meaning (*He moved his free hand around to the front of her*.) or a metaphorical meaning (*He gave Stephanie a free hand in the decoration*). By consulting the corpus, the investigator is presented with some concordance lines with *free hand* as the node. With a few clicks of the mouse, the investigator identifies a few training examples in the concordance as being literal or metaphorical. The programme then analyses the context of the manually categorised examples for phraseological patterns, and then searches all the concordance lines for similar patterns. Drawing on techniques in information retrieval (IR), the programme generates a cosine similarity score for each of the lines. Those lines with the highest scores are then automatically labelled as being literal or metaphorical. In the succeeding sessions of training, lower similarity scores are replaced with higher scores, so that only high scores are retained to make sure that similar patterns are assigned the same label. As the amount of manual labelling increases, the precision of automatic categorisation rises correspondingly. In an experiment with 996 examples taken from COCA (*Corpus of Contemporary American English*) (Davis 2008), 30 concordance lines of *free hand* were manually labelled with literal or metaphorical for the purpose of training the programme. This is rewarded with the result that the programme achieved a high accuracy of over 98% in automatic categorisation.

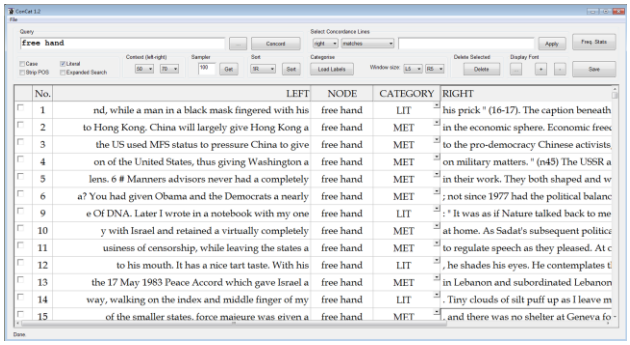


Figure 1: Concordance Categoriser

Results of the study indicate that the context-based method for the automatic categorisation of concordance lines has important potential applications for both linguistic analysis and word sense disambiguation (WSD).

3 The algorithm

As described earlier, in this study, techniques in information retrieval are applied to categorize concordance lines. According to Siddiqui & Tiwary (2006) and other literature, the typical steps in information retrieval to generate a term-document matrix are –

(i) Find individual words and their frequencies in each document

(ii) Use a stop list to remove common words

(iii) Reduce words to their stems

(iv) Assign weights to each term and prepare term-document matrix

In our case, for step (i), we find individual words and their frequencies in the user-specified span of text in each concordance line. Words out of the span are simply ignored. This will greatly reduce the computational expense. Step (ii) above does not apply in our case, as common words, most of which are function words, form important parts of collocation patterns (e.g., in spite of). For step (iii), Porter stemming (Porter 1980) is often used in information retrieval. In our case, we use a lemma list to turn all the possible collocates into their base forms, so that more abstract patterns will be observable. While this process is more computationally expensive, it yields more intelligible results for human readers. For step (iv), we assign more weights to words which are closer to the search word(s), as true collocates are more likely to fall in a span of text separated from the search word by fewer words.

These steps will result in a term-document (term-concordance) matrix, on the basis of which a vector space model (VSM) is created. With this model, similarity scores can be generated for every two concordance lines with the following cosine similarity formula

$$\mathbf{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|},$$

where the numerator represents the dot product of the vector representations for the two documents (concordance lines) while the denominator is the product of their Euclidean lengths (Manning et al. 2008).

When a new query is conducted, the above-mentioned process will be carried out automatically, resulting in a similarity score for each pair of concordance lines generated from the query. If the use assigns a label to a concordance line, those other concordance lines with the highest similarity scores will automatically be assigned with the same label.

References

Davis, M. 2008. *COCA: Corpus of Contemporary American English*. (http://americancorpus.org)

Hunston, S. & G. Francis. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kilgarriff, A. and Rob Koeling. 2003. An evaluation of

a lexicographer's workbench incorporating word sense disambiguation. *Proc. CICLING, 3rd Int Conf on Intelligent Text Processing and Computational Linguistics*, Mexico City. Springer Verlag.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

O'Donnell, M.B. 2008. KWICgrouper: designing a tool for corpus-driven concordance analysis. *International Journal of English Studies*, Vol. 8 (1), 2008, pp. 107-121.

Porter, Martin F. 1980. An Algorithm for Suffix Stripping. *Program*, 14(3): 130–137

Rayson, P. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. Thesis, Lancaster University.

Siddiqui, T. and U. Tiwary. 2006. A Hybrid Model to Improve Relevance in Document Retrieval. *Journal of Digital Information Management*. Vol 4 (1): 73-81.

Sinclair, J. 2003. *Reading Concordances*. London: Longman.

'What is the environment doing in my report?' Analysing the environment-as-stakeholder thesis through corpus linguistics

Alon Lischinsky
Oxford Brookes University
alon@lischinsky.net

At the core of initiatives towards corporate social responsibility (CSR) lies the idea that organisations have an inherent responsibility towards their stakeholders, that is, all groups and individuals affected by the organisation's activities (Freeman 1984). However, the identification of stakeholders – and, *a fortiori*, the design of procedures for reporting and consultation that allow their interests to be taken into account in organisational strategy – is far from a straightforward exercise (Phillips and Reichart 2000). There is considerable variation in the nature of organisation-stakeholder relations, which range from legally-binding duties towards shareholders and regulators, to entirely voluntary contributions to the welfare of local communities and society as a whole. Engagement with primary stakeholders – those essential to the organisation's operations – is normally well-institutionalised, but secondary ones are not necessarily engaged in regular, formal transactions with the organisation, and may often lack the social and economic resources to demand participation in the decision-making process.

Such difficulties become especially acute when considering the environmental impact of the organisation's activities. Contemporary approaches to CSR – such as the "triple bottom line" – are intended to address environmental as well as social issues, and a significant number of CSR reports claim to regard the natural environment as a stakeholder to be satisfied. Nevertheless, such claims must be carefully evaluated. As the environment cannot directly present its demands or participate in negotiation, the possibility remains that organisations may indulge in *greenwash*, simply stating an orientation to environmental concerns without actually engaging with environmental needs (Laufer 2003).

Both environmental activists and critically-oriented scholars have contended that the moral, social and ecological assumptions of "business as usual" are such that cannot be harmonised with a principled conception of environmental sustainability (Dryzek, 2005; Hajer, 1997;

Rutherford, 2006). In consequence, they argue that the adoption of an environmentally-oriented vocabulary by corporations is intended to dilute environmental action by imbuing the field of discourse with ultimately unsustainable short-term, profit-oriented presuppositions (Burchell and Cook 2006:132–3).

This research seeks to contribute to the analysis of this possibility, by exploring the representations of various organisation-stakeholder relationships in a corpus of CSR and environmental reports from major corporations based in Sweden, with a specific focus on the difference between the representation of human stakeholders and the environment. After identifying the range of lexical expressions used to refer to the various organisational stakeholders, their statistically-significant collocational patterns are analysed, as well as the discursive prosody they give rise to. Special attention is paid to formulaic phrasal templates, such as "our impact on the environment", given the highly repetitive and boilerplate-driven nature of these reports. I also explore more flexible constructions, often called *snowclones* (Pullum 2004), such as "what is good for the environment is good for X".

Results show that references to the environment are less frequent and less varied in grammatical and discursive form than those made to other stakeholders. While a difference between the representation of those stakeholders consisting of organised groups (such as suppliers, competitors or shareholders) and less easily-defined ones (such as society as a whole) can be expected, the representation of the environment is uniquely restricted. Most significantly, representation of the environment in an agentive role (or even as a grammatical subject) is exceedingly rare. To a large extent, mentions of the environment appear as formulaic phrases focusing on the company's orientation rather than environmental needs or conditions, and only infrequently do they co-occur with discussions of the company's engagement with conservation-, mitigation- or restoration-oriented NGOs.

References

Burchell, J., & Cook, J. (2006). "Confronting the 'corporate citizen': Shaping the discourse of corporate social responsibility". *International Journal of Sociology and Social Policy*, *26*(3/4), 121–137. doi:10.1108/01443330610657188

Dryzek, J. (2005). *The politics of the earth: environmental discourses* (2nd ed.). Oxford & New York: Oxford University Press.

Freeman, R. E. (1984). *Strategic Management: A Stakeolder Approach*. New York: Pitman Publishing.

Hajer, M. (1997). *The politics of environmental discourse: ecological modernization and the policy process*. Oxford: Clarendon Press.

Laufer, W. S. (2003). "Social Accountability and Corporate Greenwashing". *Journal of Business Ethics*, *43*(3), 253–261.

Phillips, R. A., & Reichart, J. (2000). "The Environment as a Stakeholder? A Fairness-Based Approach". *Journal of Business Ethics*, *23*(2), 185–197.

Pullum, G. K. (2004). "Snowclones: Lexicographical dating to the second." *Language Log*. Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/000350.html

Rutherford, P. (2006). How Have International Business Discourses on the Environment Changed over the Last Decade? *Global Social Policy*, *6*(1), 79–105. doi:10.1177/1468018106061393

Using quantitative measures to investigate the relative roles of languages participating in code-switched utterances

Cathy Lonngren-Sampaio
University of Hertfordshire
C.Lonngren-Sampaio@Herts.ac.uk

1 Introduction

Jake and Myers-Scotton define Code-switching (CS) as 'language use that consists of material from two or more language varieties at any level from the discourse to the clause.'(2009:207). They propose that in 'Classic' CS there is always asymmetry between the two (or more) languages participating in CS clauses. According to the Asymmetry Principle (ibid:209) the abstract morphosyntactic frame of the bilingual clause largely, or entirely, comes from one of the languages, named the Matrix Language (ML) while the other participating language is called the Embedded Language (EL) and typically contributes content morphemes, such as nouns, lexical verbs and adjectives.

In quantitative terms, it is possible to make certain assumptions regarding how the ML/EL asymmetry is realised in code-switched utterances. Firstly, it is reasonable to propose that the ML would contribute more words to a code-switched utterance than the EL. Secondly, the grammatical nature of the words contributed by the ML and their repetitive frequency would mean that there is less diversity in their contribution to code-switched utterances than the lexically-laden content morphemes being inserted by the EL. Thirdly, if one considers that in many languages grammatical morphemes are typically shorter in length (in terms of characters) than content words, one would expect higher mean word lengths for the EL when compared to the ML.

These assumptions form the basis of this study which aims to use quantitative methods to investigate the Asymmetry Principle in a corpus of child bilingual language and answer the following research questions: 1) Do word frequency measures provide evidence for the ML/EL asymmetry?; 2) Can correlations be found between vocabulary diversity scores and the ML/EL?; 3) What can mean word and utterance scores contribute to the investigation of the Asymmetry Principle in bilingual corpora?; and 4) Can these quantitative measures combined provide a useful method for determining the participatory roles of the languages of code-switched discourse?

2 Methodology

Twenty-five hours of recordings of naturalistic interactions between two bilingual Brazilian/English siblings (JAM, 3;6 and MEG, 5;10) and other family members, were transcribed and coded using the CHAT (Codes for the Human Analysis of Transcription) transcription system developed by MacWhinney and colleagues (MacWhinney 2012a). The resulting corpus, named the LOBILL (Lonngren BILingual Language) Corpus, then received more specific coding: language codes to differentiate English and Portuguese material were inserted throughout and a specially developed postcode was used to code bilingual utterances (see example below). Addressee information for each utterance was also included.

*MEG: <<mas a agua>[@pt]>[//] <the water is very very cold>[@en] ? [+ pe]

%add: MOT

Quantitative analyses were carried out using the CLAN (Computerized Language Analysis) software (MacWhinney 2012b), in particular the commands KWAL (which outputs specified utterances), FREQ (which outputs frequency word lists), VOCD (which outputs vocabulary diversity scores) and WDLEN (which outputs mean word and utterance lengths). An example command line is shown below:

kwal @ +t%add +t*JAM +s"PAI" +u +d | vocd +r6 +s"[+ *]"

Consisting of two parts (separated by the upright slash), first KWAL selects all utterances JAM addresses to PAI (his father). Then VOCD performs an analysis on only the CS utterances. By subsequently adding the strings -s"<@pt>" and -s"<@en>", VOCD then performs analyses on only the English material in CS utterances and then on only the Portuguese material in CS utterances.

By systematically substituting the speaker codes (JAM, MEG, MOT PAI), the addressee codes (PAI, MOT, MEG, JAM) and the commands (freq, vocd +r6, wdlen) 36 analyses were performed for each speaker, making a total of 144 analyses.

The results were examined and triangulated with other quantitative and qualitative analyses previously performed on the LOBILL Corpus. Patterns were observed and correlations were made between the four different types of measures (word frequency, vocabulary diversity, mean

word length and mean utterance length) and the roles of the two languages in code-switched utterances.

3 Results

The first set of results, obtained from the frequency analyses (performed by FREQ), provided a measure of the proportion of English and Portuguese words used by the siblings and their parents when code-switching with each other. The following correlations were found: a high proportion of words in one language indicated that that particular language was acting as the Matrix Language; a low proportion of words indicated a role more akin to the Embedded Language.

Secondly, the vocabulary diversity analyses (performed by VOCD) provided separate diversity (D) scores for each language in code-switched utterances. Low D scores were found to correlate with the Matrix Language whereas high D scores were representative of the Embedded Language. What is postulated in both cases (frequency proportions and diversity scores) is that the greater the relative difference between the values for each language, the more asymmetrical their participatory roles appear to be. This in turn means that where the relative difference in values is less disparate we would expect more equal participation of both languages in code-switched utterances.

The third and forth sets of values, which resulted from the WDLEN analyses, measured mean word and mean utterance lengths. A low mean word length was found to correlate with the Matrix Language while a high mean word length correlated with the Embedded Language. In terms of utterance length, the reverse correlation was found: a low mean utterance length provided evidence that the language was acting as the Embedded Language whereas a high mean utterance length was found to be representative of the Matrix Language. As for the first and second sets of results, the evidence suggests that the greater the comparative difference in values between the two languages (this time in terms of means) the more asymmetrical their participatory roles become.

4 Conclusion

This study set out to examine whether quantitative measures could contribute to the investigation of the Asymmetry Principle, a common feature of code-switched discourse. The results of the analyses performed with the CLAN commands KWAL, FREQ, VOCD and WDLEN, led to the formulation of correlations between the four types of quantitative values and the participatory roles (Matrix and Embedded) of the languages involved. These correlations have in turn led to the development and proposal of a novel schema (see below) designed to be used by researchers wishing to interpret such values arising from the analysis of their own code-switched data.

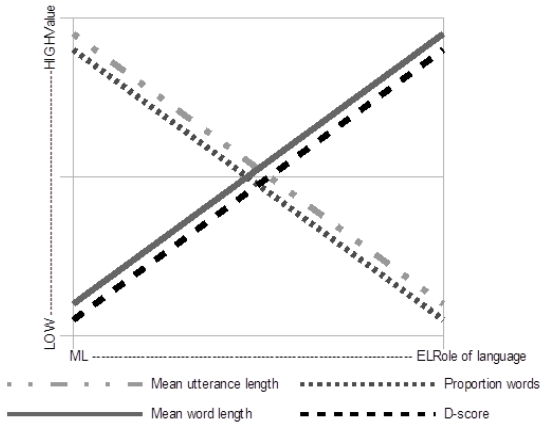


Figure 1. Schema for the interpretation of FREQ, VOCD and WDLEN values when used to investigate the relative roles of the languages participating in code-switched discourse

If the methodology used in this study were replicated on other code-switching data sets, it is believed that the proposed schema (Figure 1) would allow for cross-linguistic comparison: the original results from the LOBILL Corpus could be compared with the results from other bilingual corpora. Such comparisons would have the potential to shed more light on the contribution that the use of quantitative measures could make when investigating the Asymmetry Principle in code-switched discourse.

References

MacWhinney, B. 2012a. *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition.*. Carnegie Mellon University. Available online at: http://childes.psy.cmu.edu/manuals/chat/pdf .

MacWhinney, B. 2012b. *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition.*. Carnegie Mellon University. Available online at: http://childes.psy.cmu.edu/manuals/clan/pdf .

Myers-Scotton, C and Jake, J.L. 2009. "A universal model of code-switching and bilingual language processing and production". In B. Bullock and J. Toribio (eds.) *The Cambridge Handbook of Linguistic Code-switching*. Cambridge: Cambridge University Press.

"The results demonstrate that …". A corpus-based analysis of evaluative *that*-clauses in medical posters

Stefania M. Maci
University of Bergamo, Italy
University of Lancaster, UK
stefania.maci@unibg.it
s.maci@lancs.ac.uk

1 Introduction

Poster sessions at scientific conferences made their first appearance in the US in 1974. Since then posters have rapidly become a major format for scientific and medical communication at conferences. Although there are many scientific guidelines for poster preparation and presentation, little attention has been paid to the study of scientific academic posters from a linguistic perspective. This may depend on the fact that posters are commonly regarded as less prestigious than research papers (Swales 2004; Swales and Feak 2000). This genre has probably been neglected also because of the predominance of visual elements in it. There are, nevertheless, textual traits worth examining which characterise the inductive reasoning typical of scientific discourse, in general, and of medicine, in particular, in that generalization and theoretical abstraction derive from specific observations of certain phenomena.

2 Research Questions

Medicine is an empirical science, which exploits inductive reasoning, in that generalization and theoretical abstraction derive from specific observations of certain phenomena. In posters, due to space/length constraints, generalised and theoretical claims tend to be lexicalized or 'topicalized', rather than discussed.

My purpose here will be to examine the way in which medical discourse is linguistically organized in this genre. My overarching Research Question, therefore, will be:

- How is scientific discourse organised in such a condensed genre?

If posters present 'topicalized' claims, then their relevance needs to be recognized synoptically. As Sala (forthcoming) claims, such recognisability is obtained through adherence to domain-specific epistemological peculiarities in representing effectively how knowledge and the process of knowledge-making take place to persuade the reader of the associated poster's worth.

Given that, my subresearch question here will be:

- How are inductive scientific reasoning and its process of knowledge-making thematized in evaluative meaning forms within the genre of posters?

One of the most powerful ways of expressing evaluative meaning in academic discourse is by means of *that* clauses (Hyland and Tse 2005). It is, therefore, the purpose of this research to investigate how the inductive reasoning of scientific discourse is organized in medical posters, by focussing on the evaluative *that*-clauses, i.e. those complement clauses embedded in a superordinate one (the matrix clause) to complete its construction and project the writer's attitudes about something (Hyland and Tse 2005: 40). In other words, the superordinated or matrix sentence is the clause which contains both the evaluation and the source of evaluation (object or participant) while the *that*-clause projects the entity which is evaluated, as example (1) shows:

(1) We postulate that [MATRIX] this property leads to increased strand transfer by RT [PROJECTED CLAUSE] […] (P012)

To gain reader's attention, poster writer must demonstrate either in visuals or in words that they have something scientifically worthwhile to say. If visuals represent facts, which speak for themselves (Maci 2011), texts express attitudinal meaning and evaluation.

3 Methods

Since the first news item about medical posters was published in 1974 (Maugh, 1974), I started my search from 1975 via MEDLINE [1] *and* PUBMED [2], the most authoritative online databases containing citations and abstracts taken from health and medical journals. Amongst all the medical journals available, I found that the journal which began the publication of poster abstracts was the *American Journal of Epidemiology*. I therefore decided to concentrate on posters specializing in the epidemiological field. I then searched on the Net for all available posters presented at congresses and published online by institutions and medical schools, as well as by online journals with an ISSN code and specialising in poster publication. I thus collected a corpus of 360 medical posters (2,387,328 tokens; 3,652 types; TTR 15.29) presented at international conferences, and, drawing on on

[1] www.medline.cos.com/
[2] www.ncbi.nlm.nih.gov/pubmed/

Hyland and Tse (2005), I tried to classify the type of evaluative *that*-clauses in terms of evaluative stance, evaluative entity, evaluative source and evaluative espression.

4 Results

The results suggest that the various classes of evaluative *that* are differently distributed: from authorial self-reference in the *Introduction* section there is a switch to data reference in the *Results* and *Discussion* sections. In addition, epistemic stance, present almost everywhere, seems to foreground the highest degree of objectivity and reliability. In this context, the authorial voice appears as a mere instrumental tool metadiscursively guiding the reader into the right interpretation of facts and findings. This seems to be supported by the presence of verbs in the matrix clause introducing the *that*-complementizer, which refer to research procedure (*demonstrate*, *show*, *indicate*) and cognitive processes (*suppose*, *assume*, *hypothesize*) featuring the scientific mind at work.

5 Conclusions

The research questions posed at the beginning of my investigation suggest the following answers:

1. Authorial (self-)reference is found in the *Introduction* sections of posters. This is required because of the need to establish credentials supporting scientific credibility and reliability.
2. Evaluative *that*-clauses are absent in the *Methods* sections of posters.
3. In the *Results* and *Discussion* sections of posters, there is a transformation from authorial (self-)reference to data reference to give way to reports of lab work and experimental practice, which, supported by visual data, create the illusion that facts speak for themselves.

Overall, we can say that:

a. Epistemic expressions indicate the highest degree of certainty and objectivity;
b. The authorial voice is mere instrumental tool to meta-discursively guide the reader into the right interpretative path (interpretation of findings);
c. Verbs expressed in the matrix clause contribute to conveying such an illusion as they refer to research activities (*find*, *indicate*, *reveal*, *show*), deriving from experimental procedures to be reported (*note*, *report*, *suggest*) and featuring the scientific mind at work;
d. Argumentation seems absent (no discourse verb) because there is no need for persuading when fact speak for themselves.

Such preliminary analysis certainly has some limitations, and needs to be triangulated.

Nevertheless, given the lack of applied linguistic research on posters, my investigation may offer new insights of this neglected genre realised with a type of discourse aiming at constructing scientific reliability. The results will then hopefully prove to be a valuable pedagogical resource in the future for the different academic communities represented in the corpus, as well as for EAP teachers

References

Hyland, K. and Tse, P. 2005. Hooking the Reader: A Corpus Study of Evaluative *that* in Abstracts. *English for Specific Purposes* 24: 123-139.

Maci, S.M. 2011. Genre variation in medical discourse: the case of medical posters. In S. Sarangi, V. Polese and G. Caliendo (eds). *Genre(s) on the Move. Hybridization and Discourse Change in Specialized Communication.* Napoli, ESI: 169-190.

Sala, M. (forthcoming). Research article abstracts as domain-specific epistemological indicators. A corpus-based study. Paper presented at the 2012 CLAVIER conference, Modena, April 12-13, 2012.

Swales, J.M. 2004. *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

Swales, J. M. and Feak, C. 2000. *English in Today's Research Settings*. Cambridge: Cambridge University Press.

Reading Dickens's characters: investigating the cognitive reality of patterns in texts

Michaela Mahlberg
University of Nottingham
michaela.mahlberg@nottingham.ac.uk

Kathy Conklin
University of Nottingham
kathy.conklin@nottingham.ac.uk

There is more and more research that seems to fall under the umbrella term 'corpus stylistics'. Corpus stylistic research applies corpus methods to the analysis of literary texts giving particular emphasis to the relationship between linguistic description and literary appreciation. Corpus stylistics has much to offer to literary linguistics because it makes it possible to deal with larger amounts of text(s) than any manual 'practical' stylistic analysis might. Linguistic phenomena in literary texts can be quantified and literary texts can be compared to more general language. However, it is also important to ask how patterns identified, and quantified, in literary texts can be explained and interpreted in terms of their stylistic significance in a particular text. It is a crucial aspect of literary linguistics or literary stylistics to aim to account for the potential or perceived effects that texts have on readers. Literary linguists are interested in why readers identify with characters, feel empathy or dislike a particular character. An important step towards finding answers to such questions is to investigate relationships between patterns in texts and observations on actual reading behaviour.

This paper reports on findings of a study piloting an innovative approach within corpus stylistics. We examine whether hypotheses formed on the basis of corpus stylistic research can be demonstrated to have a cognitive reality. Hence, our approach employs psycholinguistic methods to complement corpus stylistic findings. We build on previous corpus stylistic work that has investigated Dickens's techniques of characterisation (Mahlberg 2013, Mahlberg & Smith 2012) and identified patterns of body language and other character cues in the form of five word clusters (eg *his hands in his pockets* or *I don't know how it*). Corpus findings suggest that such clusters can function in the text to provide vivid and authentic descriptions of characters; these patterns have been referred to as literary 'contextualising' patterns (Mahlberg 2013) and contrasted with patterns that 'highlight' or foreground specific character information. In literary criticism it is particularly the 'highlighting' patterns that have received most attention. Corpus stylistic work now draws attention to a range of patterns that require further explanation within established literary scholarship.

Crucially, corpus research in general has shown that corpus findings can bring to light textual phenomena that readers may not be consciously aware of. The degree of awareness of textual patterns may have implications for how readers perceive characters in a novel, as characterisation is an emergent process where textual patterns interact with readers' real world knowledge (Culpeper 2001, Stockwell 2009). To study to what extent readers attend to patterns of body language presentation and other patterns of character information, we conducted an experiment to track readers' eye movements when reading sample passages from Dickens' novels. This experiment is similar to the one described in Siyanova et al. (2011). The sample passages from the novels include patterns discussed in Mahlberg (2013).

One of our key findings is that the phrases and patterns under investigation are associated with reading times that are shorter than the general reading times across the sample passages. Possible explanations for this can be found in the con-textualising functions of the patterns in the given extracts. The eyetracking data will be further related to data that the participants provide on the character information they notice. In this way we investigate relationships between conscious impressions of characters in the reader's mind and the patterns that potentially contribute to creating these impressions. Additionally, the data of the eye-tracking study is complemented with the results of an online question-naire study focusing on the same patterns and textual extracts to gather further data on the actual reading experience. We discuss our results in the context of work in cognitive stylistics and literary criticism and argue that the findings open up new routes for the study of characterisation in Dickens's fiction.

References

Culpeper, J. 2001. Language and characterisation. People in plays and other texts. Harlow: Pearson.

Mahlberg, M. 2013. Corpus Stylistics and Dickens's Fiction. New York & London: Routledge.

Mahlberg, M. and Smith, C. 2012. "Dickens, the suspended quotation and the corpus", Language and Literature, 21(1): 51-65.

Siyanova, A., Conklin, K., & Schmitt, N. 2011. "Adding more fuel to the fire: An eye-tracking study

of idiom processing by native and nonnative speakers". Second Language Research, 27 (2): 251-272.

Stockwell, P. 2009. Texture – A Cognitive Aesthetics of Reading. Edinburgh: Edinburgh University Press.

Experimenting with objectivity in corpus and discourse studies: expectations about LGBT discourse and a game of mutual falsification and reflexivity

Anna Marchi
Lancaster University
a.marchi@lancaster.ac.uk

Charlotte Taylor
University of Portsmouth
charlotte.taylor@port.ac.uk

1 Introduction

In this paper we address the issue of the researcher's individual influence on the research findings and explore the potential of falsification as a counter-corroboration strategy.

It is our intent to follow up on an experiment into researcher triangulation we did a few years ago (Marchi and Taylor 2009), where we tested whether two researchers starting with the same corpus and research question and theoretical /methodological framework would come to the similar conclusions. In that case our main interest was methodological reliability, recently investigated on a larger scale by Baker (2011), here, instead the focus is more specifically on the scientist's personal influence (Myrdal 1970).

In our previous experiment we concluded wishing to 'investigate with greater rigour the extent of the influence of the researcher' (Marchi and Taylor 2009: 20). In this study, as before, both researchers work on the same corpus and the same RQ, but with explicitly different hypotheses. By adding the specific hypothesis element to the experimental design, we aim to test to what extent predictions/expectations impact on the outcome. In the second part of the study we then attempt to falsify each other's hypothesis as a means of testing the potential of falsification.

2 Case study

The topic (i.e. the secondary research question) chosen as testing ground for our primary RQ is the representation of the LGBT[1] community in an Italian (la Repubblica) and British (the Guardian) newspaper of similar political affiliation.

More specifically we are asking: How do the Italian and British liberal newspapers write about

[1] Acronym standing for Lesbian, Gay, Bisexual and Transgender. There are many variations of the acronym (e.g. GLBT, LGBTQIA, etc.) LGBT was adopted here purely on the basis of frequency of occurrence.

LGBT-related issues and are the resulting constructions of LGBT people similar or different?

3 Corpus

Two search term based corpora of the Guardian and la Repubblica were compiled. Equivalent terms for GAY, LESBIAN, *homosex**, *heterosex**, *transsex**, *transgender*, *LGBT*, *GLBT*, *same-sex*, *intersex**, *hermaphrodit** were selected in English and Italian, in order to grant maximum coverage of LGBT related issues. Each corpus consists of approximately 1 million words and covers a period of two years (between October 2007 and October 2009). The choice of the period of time is relatively arbitrary and is based on the criterion of availability of data for the Italian corpus, the contextual timeline of events in the two countries was however taken into careful account as we wanted to look at a relatively 'banal' period, rather than a period of high newsworthiness.

The two corpora have been defined as "comparable" on the basis of the similar political leaning of the two newspapers, although 'true' comparability remains a problem.

4 Method

The methodological and theoretical framework behind this work is that variously defined as corpus-based or corpus-assisted discourse studies. The approach builds, among others, on seminal work such as Stubbs' (1996 and 2001) and our method largely follows the model of research done at Lancaster University on projects such as RASIM (Baker et al. 2008, Gabrielatos and Baker 2008) and research done by the SiBol group[1] at the universities of Bologna, Siena and Portsmouth (Partington 2010, edited volume).

The research mainly relies on collocation and concordance analysis of the query terms and compares linguistic patterns for the two languages.

5 Definitions

Falsification: according to Popper (1959) falsifiability is an essential quality of hypotheses, and the very scientific status of a theory is its falsifiability, or refutability, or testability. Here we use falsification both to refer to the testing of one's own hypothesis during the analysis and to the researchers' mutual attempt to confute each other's completed analyses, though it is important to keep the distinction between the two types clear.

Reflexivity: we distinguish between two types of reflexivity, personal and epistemological. Epistemological reflexivity 'encourages us to reflect upon the assumptions (about the world, about knowledge) that we have made in the course of the research, and it helps us to think about the implications of such assumptions for the research and its findings' (Willig, 2001: 10), i.e. it is related to the impact of the research process on the outcome. Personal reflexivity, on the other hand, is specifically related to how a person's values, beliefs, interests, identity influence his or her research.

6 Experiment description

The experiment is divided in two phases. For the first phase the two researchers work independently with opposing hypotheses. Researcher A's initial hypothesis is that, despite their similar political leaning, the Italian newspaper is more likely to be bigoted than the British one in treating LGBT related issues. Researcher B's hypothesis is that there won't be such a marked difference.

Each researcher develops a falsifiable account of the patterns of representation. This allows us to discuss the extent to which the initial hypothesis affects the analytic process and the outcome.

In the second phase the two researchers exchange analyses and attempt to falsify the other's results. This allows us to then compare the process of falsification within one's own work with the process of falsifying another researcher's analysis.

7 Collaborative work

It is relevant to point out that the affinity between the two researchers' views made it difficult to find a topic that gave them diverging hypothesis to work from. This case study was chosen precisely because it could provide such a condition.

In CADS (Partington 2009), we tend not to come to the data with a specific hypothesis to test, but we are very likely to have more or less loose expectations. Furthermore, particularly in discourse oriented research, we tend to study things that we care about (in a continuum that goes from vague interest to committed involvement), which makes it important to keep track of our individual influence/bias as researchers (personal reflexivity), on top of the impact the research process itself has (epistemological reflexivity).

In this research project we make the self-reflexivity explicit through the (artificial)

[1] SiBol group website: http://www3.lingue.unibo.it/blog/clb/

construction of reciprocal falsification. We do not claim that this can/should be integrated into research, it is an experiment, but we do maintain that collaborative work has intrinsic added value. Ultimately we see cooperative research as a guarding strategy against corroboration-drive and we argue for a system that encourages academic collaboration[1].

References

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press". *Discourse & Society* 19(3): 273–305.

Baker, P. 2011. "Discourse, news representations and corpus linguistics". Plenary given at Corpus Linguistics 2011 Conference, Birmingham 20-22 July 2011.

Gabrielatos, C. and Baker, P. 2008. "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005". *Journal of English Linguistics* 36(1): 5–38.

Marchi, A. and Taylor, C. 2009. "'If on a winter night two researchers': A challenge to assumptions of soundness of interpretation". *CADAAD Journal* 3(1): 1–20.

Myrdal, G. 1970. *Objectivity in Social Research*. London: Gerald Duckworth & Company Ltd.

Partington, A. 2009. "Evaluating evaluation and some concluding thoughts on CADS". In J. Morley and P. Bayley (eds.) *Corpus-Assisted Discourse Studies on the Iraq Conflict: Wording the War*. London: Routledge, pp. 261–303.

Partington, A. 2010. *Corpora special issue. Modern Diachronic Corpus-Assisted Studies*. Edinburgh: Edinburgh University Press.

Popper, K. 1959. *The Logic of Scientific Discovery*. London: Hutchinson and Co.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. 2001. *Words and Phrases: corpus studies of lexical semantics*. London: Blackwell.

Willig, C. 2001. *Introducing Qualitative research in Psychology: adventures in theory and method*. Open University Press, Buckingham

[1] The contribution of the authors to this research was perfectly symmetric and the author order in this abstract simply alphabetical.

Have – causative, or experiential? A parallel corpus-based study

Michaela Martinková
Palacký University
michaela.martinkova@upol.cz

1 Introduction

Have complemented by a noun phrase (NP) and a non-finite verb form is in Quirk et al. (1985, 1206) ranked among the complex transitive complementation with "coercive" meaning. Huddleston and Pullum argue (2002, 1236) that "*have* is also used with a non-causative 'undergo' sense". This type of *have* is discussed by Quirk et al. (1985, 1412) under the "existential *there*", where the NP in the subject of *have* introduces a participant with "considerable involvement in the existential proposition" (1985, 1411). This reading of *have* is sometimes called 'experiential' (Austin 2004, 77).

The problem arises when it comes to strictly disambiguating *have* as causative or experiential. Only causative, not experiential *have* is felicitous in the imperative and progressive forms (Poldauf 1967, 32). Other criteria are hard to pinpoint. Guilquin in her analysis of causative *have* and *get* in ICE-GB (2003) discards "those constructions that are not causative", but exact criteria are not given. She admits that even the CAUSER "can present various degrees of involvement in the caused event" (2003, 132). In (1), for example, the CAUSER "not only initiates the caused event [...] but is also "acted on" during the operation":

(1) *I had my tonsils removed.*

This paper turns to a multilingual corpus to see whether it "can make meanings visible through translation" (Johansson 2007, 57), more specifically, which meanings of the construction with *have* can be seen "through translation patterns". A Slavic language, namely Czech, was selected since, as Poldauf (1964, 250) argues, there are parallels between the experiential "*have* construction" and a Czech construction with the "non-attached dative", recently analyzed as the "Affected Possessor" construction within the framework of Construction Grammar by Fried (2011). In (2), a Czech equivalent of (1), the dative *mi* translates *I* in the subject of *have*:

(2) *Vytrhli mi mandle.*

'they-pulled-out me_{DAT} $tonsils_{ACC}$'

To apply appropriate terminology, *mi* refers to the "Affected Possessor" (like the English *I* in the

subject of *have*), and the noun *mandle/tonsils* to the "Possessum". However, (1) can also be translated with a causative verb:

(3) *Nechala jsem si vytrhnout mandle.*
I-let$_{PAST}$ PRON$_{REFL.DAT}$ pull-out$_{INF}$ tonsils$_{ACC}$

This paper analyzes the *have* construction with the infinitive (V_{INF}) and the *-ing* form (V_{ING}) and aims to address the following questions:

1. Do the Czech translation equivalents strictly fall into the two categories, experiential and causative? Which of these meanings is more frequent?

2. Do the Czech translation equivalents reflect any difference between the meaning of *have* in the construction with V_{INF} and with V_{ING}?

In more general terms, this study tests the potential of a parallel translation corpus for a contrastive analysis of a very complex linguistic phenomenon with many variables in play: an underspecified notion of affectedness and ownership (e.g. Taylor 1995, Heine 1997, Fried 1999), animacy and inanimacy of participants, contextual dependence, and semantic and syntactic properties of English and Czech verbs.

2 Data and method

The data were taken from Intercorp, a multilingual translation corpus of Czech and 31 languages, namely its Cz-En and En-Cz subpart. A bidirectional parallel subcorpus of fiction was created (2,825,303 words in the En-Cz direction and 866,039 words in the Cz-En direction) and four CQL queries were designed. The results were checked manually to discard spurious tokens.[1]

3 Data analysis

Table 1 presents absolute and relative frequencies (per 0.5mw) of the *have* construction in English source texts (STs) and English target texts (TTs), according to the type of the non-finite verb (V_{INF} and V_{ING}) and the subject of the secondary predication (noun and personal pronoun):

	STs		TTs	
	V_{ING}	V_{INF}	V_{ING}	V_{INF}
noun	57	27	15	5
rel. freq.	10.09	4.74	8.66	2.88
pronoun	30	30	3	8
rel. freq.	5.3	5.3	1.73	4.6

Table 1: Frequencies of *have* construction in Intercorp subcorpus

[1] Those in which the NP following *have* forms a semantic unit with it: *She was simply having trouble concentrating on two things at once.*

Table 1 shows that the *have* construction is more frequent in English STs than in English TTs, which suggests that it tends to be influenced by translation effects. English TTs are thus excluded from the rest of the study.

Table 2 lists Czech translation equivalents of the *have* construction found in English STs:

Czech translation		V_{INF}	V_{ING}
CAUSAT.	causative verb	26	14
	prep. phrase (PP)	0	6
	modal verb	6	1
EXPER.	dative	1	9
mít (*have*) NP+ rel.clause		0	8
zero correspondence		**22**	**46**
whole sentence omitted		1	4

Table 2: Czech transl. equivalents of *have* construction in English STs

4 Discussion of Question 1

The numbers in Table 2 suggest a striking predominance of causative over experiential *have*. This does not correspond to Gilquin (2003), who reports that only 42.5% of tokens of the *have* construction in ICE-GB are causative. The answer is to be found in the large number of zero correspondences, where only the secondary predication, not *have*, gets translated into Czech (39.2% of the tokens with V_{INF} and 54% of those with V_{ING}). A closer analysis reveals that this concerns mainly experiential *have*, the translation of which is more susceptible to translation effects. The Czech dative should have often been used, especially if the "Possessum" is high on the "possessibility hierarchy" scale. This applies to (3), where the dative evoking "affectedness" of the "Possessor" in (3b) is more appropriate than a possessive pronoun used in (3c):

(3) a. *men starving to death, yet having tumors thriving inside them* [Hailey]
b. *jim v těle bují nádory*
'them$_{DAT}$ in body thrive tumours$_{NOM}$'
c. *v jejich těle bují nádory*
'in their body thrive tumours$_{NOM}$'

If the "Possessor" is an obligatory argument of the Czech verb, it must be expressed in the case form required by the verb, i.e. dative is ruled out and "affectedness" of the "Possessor" is lost in Czech:

(4) a. *they didn't have scars burning on their foreheads* [Rowling]
b. *je na čele nepálila žádná jizva*
'them$_{ACC}$ on forehead burned$_{NEG}$ no scar$_{NOM}$'

In some of these cases, a more expressive Czech

verb compensates for the loss of "affectedness", or a valency frame allowing the non-attached dative is selected.

Not all the tokens of untranslated *have*, however, are experiential. Causative *have* was omitted if the verb in the secondary predication was a verb of "introducing":

(5) a. *She had someone introduce us* [Lindsay]
b. *Někdo nás představil.*
'someone$_{NOM}$ us$_{ACC}$ introduced'

Truly problematic are sentences in which the subject of *have* is the "Affected Possessor" and the CAUSER at the same time. The Czech translator then must select one of the readings. This is the case of (6), where *have* was judged as causative by the translator (see 6b), even though its experiential reading can hardly be ruled out (see 6c):

(6) a. *Leonora had guests coming.* [Siddons]
b. *Leonora pozvala nějaké hosty.*
'Leonora$_{NOM}$ invited some$_{ACC}$ guests$_{ACC}$'

c. *Leonoře přijdou hosté.*
'Leonora$_{DAT}$ will-come guests$_{NOM}$'

In some cases, the Czech literal equivalent of *have* followed by a noun postmodified by a relative clause is used, which however rules out the causative reading.

5 Discussion of Question 2

Table 2 suggests that the experiential reading of *have* is more common in the construction with V_{ING} than with V_{INF}; there is just one dative construction equivalent to *have* NP V_{INF} as compared to nine for *have* NP $V_{ING.}$ Translations with a causative verb cover 46% of all translations of *have* NP V_{INF}, but only 16% of *have* NP V_{ING}. Modal verbs expressing obligation are used in the Czech translation of *have* NP V_{INF} rather than *have* NP V_{ING}. Causative *have* in the construction with V_{ING} tends to introduce an inanimate CAUSE, which tends to be rendered in Czech as a PP.

6 Conclusions

While it is possible to identify Czech equivalents of the *have* construction as clearly experiential or causative (Table 2), due to a high number of zero correspondences it cannot be stated with confidence that the Czech equivalents fall neatly into the two categories. Sometimes the affected participant in the experiential *have* construction has to be expressed in the case form required by the Czech verb (dative is ruled out), sometimes it is lost due to translation effects. Causative *have* is lost in translation if a verb of "introducing" is used in the secondary predication. The analysis identified cases of a unified causative and experiential reading of *have*, in which its subject is not only the CAUSER, but is at the same time affected by the activity expressed in the secondary predication – the CAUSER may be causing the event exactly because he/she has some interest in it. This double interpretation of *have* can be due to the fact that *have* is just a grammatical verb linking a secondary predication to the subject. Czech then has to disambiguate, i.e., either use a dative construction, or a causative verb/modal verb/PP.

References

Austin, F. 2004. "Points of Modern English Usage LXXX". *English Studies* 85 (1): 77-88.

Czech National Corpus – InterCorp, Institute of the Czech National Corpus, Prague. Accessible online: http://www.korpus.cz.

Fried, M. 1999. "From Interest to Ownership. A Constructional View of External Possessors". In Payne, D. L. and Barshi, I. (eds) *External Possession*. Amsterdam: Benjamins.

Fried, M. 2011. "Plain vs Situated Possession in Czech: A Constructional Account". In W. McGregor (ed.) *Expressions of possession.* Mouton de Gruyter.

Gilquin, G. 2003. "Causative Get and Have: So Close, So Different". *Journal of English Linguistics* 31: 125-148.

Heine, B. 1997. *Possession: Cognitive Sources, Forces, and Grammaticalization*. Cambridge: CUP.

Huddleston, R. D. and Pullum, G. K. 2002. *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Johansson, S. 2007. "Seeing through Multilingual Corpora". In R. Facchinetti (ed.) *Corpus Linguistics 25 Years On.* Amsterdam – New York: Rodopi.

Poldauf, I. 1940. "Podstata anglického pasiva a anglické vazby zájmové účasti". *Časopis pro moderní filologii* 26: 358-363.

Poldauf, I. 1964. "The Third Syntactical Plan". *Travaux linguistiques de Prague* 1:241-55.

Poldauf, I. 1967. "The Have Construction". *Prague Studies in English* 12: 23-40.

Quirk, R. et al. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Taylor, J. R. 1995. *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Clarendon Press.

Annotating translation errors in Brazilian Portuguese automatically translated sentences: first step to automatic post-edition

Débora Beatriz de Jesus Martins
Federal University of São Carlos
debora.martins@dc.ufscar.br

Lucas Vinicius Avançco
University of São Paulo
avanco@grad.icmc.usp.br

Maria das Graças Volpe Nunes
University of São Paulo
gracan@icmc.usp.br

Helena de Medeiros Caseli
Federal University of São Carlos
helenacaseli@dc.ufscar.br

1 Introduction

Machine Translation (MT) is one of the oldest and most important areas of Natural Language Processing. Although extensive research has been conducted in this field, current state of the art Phrase-Based Statistical Machine Translation (PB-SMT) systems (Och and Ney 2004; Koehn et al. 2007) are not yet able to deliver a perfect translation. Nevertheless, the output of a PB-SMT system can be used as the input for a manual post-edition process, which is faster than a full manual translation.

According to Krings (2001), post-edition can be defined as the process of modifying the MT output so that it becomes acceptable for a given purpose. Whereas the source text to be transformed into the target language is the only input for the traditional translation process, post-edition involves the comparison of two inputs: the MT output and the source text. As stated by Specia (2011), post-edition of MT output can be successfully incorporated into the translation process, assisting in minimizing the time and costs involved.

At present, the most common form of post-editing is one executed by human translators, whether professional translators or users of translation systems. One drawback of this manual post-processing is that it incurs costs and requires time. As an attempt to overcome these problems, some methods have been proposed for automated post-editing (APE) such as: (Béchara et al. 2011), (Potet et al. 2011), (Lagarda et al. 2009), (Simard et al. 2007), (Elming et al. 2006) and (George and Japkowicz 2005).

This abstract presents the first steps towards the construction of an APE system for the PB-SMT system of PorTAl[1] (Vieira and Caseli 2011). The translator of PorTAl was trained using Moses[2] (Koehn et al. 2007) and the FAPESP Brazilian Portuguese-English parallel corpus (Aziz and Specia 2011). This abstract describes the manual annotation process of some sentences automatically translated by the translator of PorTAl. The purpose is to identify the most common errors in translating English sentences into Brazilian Portuguese. The annotation process used the Blast[3] tool (Stymne 2011).

The knowledge derived from this annotation process – the automatically translated corpus annotated with translation errors – will be used for automatic learning of rules for APE. Part of this corpus (only 180 sentences) was already used in the manual construction of a rule-based post-edition system (Avançco 2012) which reported an improvement of 3.72% in BLEU (Papineni et al. 2002) and 1.23% in NIST (Doddington 2002), two of the main measures used to evaluate MT.

The abstract is organized as follows. Section 2 presents the error typology defined for the annotation process and Section 3 describes how this process was performed. In Section 4 the main results are presented, followed by brief conclusions in Section 5.

2 Error typology

The error categories used in the annotation task described here are based on Popovic and Burchardt (2011) which, in turn, are based on Vilar et al. (2006).

After a previous experiment carried out to analyse the most common errors in a small sample of texts translated from English to Brazilian Portuguese, the categories proposed by those authors were adapted to the categories (A-D) and subcategories presented below:

A. Inflectional errors: the erroneous word has the correct base form, but it is wrongly inflected:

1. Number agreement – agreement errors regarding plural and singular;
2. Gender agreement – agreement errors regarding masculine and feminine;
3. Verb inflectional errors – verb conjugation using a wrong person or incorrect verb tense;
4. Part of speech errors – form of a word incorrectly changed by the translation system, resulting in a different

[1] http://www.lalic.dc.ufscar.br/portal
[2] http://www.statmt.org/moses/
[3] http://www.ida.liu.se/~sarst/blast

grammatical class.

B. Lexical problems: only one word makes up the error:

1. Extra word – a word wrongly added to the MT output;
2. Missing word – a word wrongly omitted from the MT output;
3. Not-translated word – a word that remains in the original language in the MT output;
4. Incorrectly translated word – an incorrect translation of a corresponding word in the source text;
5. Misspelling – a misspelled word in the MT output.

C. N-gram: two or more words make up the error. The words typically form a semantic expression or other form of expression:

1. Missing n-gram – a n-gram wrongly omitted from the MT output;
2. Not-translated n-gram – a n-gram that remains in the original language in the MT output;
3. Incorrectly translated n-gram – an incorrect translation of a corresponding n-gram in the source text.

D. Reordering errors: one or more words which are misplaced in the MT output.

3 The annotation process

The annotated corpus is composed of 1,185 sentences. Each automatically translated sentence in Brazilian Portuguese (Sys) is accompanied by the correct version in Brazilian Portuguese (the reference translation, Ref), and the source sentence in English (Src) as available in the original FAPESP corpus[1].

This set of 1,185 triples (Sys+Ref+Src) was annotated by two native speakers of Brazilian Portuguese with good knowledge of English, following four general rules:

R1. To annotate errors from categories C, D, A and B (in that order[2]);

R2. To annotate the minimum sequence of words that need to be changed in order to make the Sys correct;

R3. Do not annotate errors caused by previously annotated errors;

R4. If there is more than one error in the same word or n-gram, all of them should be annotated.

[1]The corpus annotated here is composed of 1,185 sentences from the 1,314 full teste-a suite, available at: http://pers-www.wlv.ac.uk/~in1676/resources/fapesp/index.html.

[2] The order has been chosen because in previous error annotation experiments it was verified that some categories influence others and, therefore, an order was defined to optimize the annotation process.

Initially, the two annotators collaboratively annotated a set of 54 sentences following these four general rules, and adapting them after discussing any disagreement. Then a further 126 sentences were annotated by each separately in order to measure the inter-annotator agreement, which was greater than 60%. The inter-annotator agreement was calculated as the percentage of errors marked exactly the same way, that is, same error category and same words annotated in Src, Sys and Ref. The calculation counted a sentence without errors as one instance. In sentences where errors happened, each individual error was considered to be a different instance.

4 Results

Table 1 presents the total number (and percentage) of errors annotated by at least one annotator in each category. From the 1,185 sentences, 388 of them did not present any identified error occurrence, so 67.26% of the translated sentences were identified as having at least one error occurrence. On average there are 2.51 errors per sentence where errors were identified.

It is possible to notice that the main error source was the lexical one (category B) with 44.95% of the total amount of annotated errors. These results corroborate previous studies for the same language pair (Caseli 2007).

Error category		Amount	%
A	A1	275	13.75
	A2	250	12.50
	A3	199	9.95
	A4	45	2.25
B	B1	102	5.10
	B2	304	15.20
	B3	198	9.90
	B4	281	14.05
	B5	14	0.70
C	C1	10	0.50
	C2	4	0.20
	C3	142	7.10
D		176	8.80
Total		**2000**	**100**

Table1: Results

Furthermore, inflectional errors (category A) amounted to 38.45% of all errors, thus also representing an important category for automatic post-editing to address in order to improve automatic machine translation. Finally, categories C and D represent only 7.8% and 8.8% of the total of errors respectively.

5 Conclusions

This abstract describes the manual annotation process carried out to identify the most common errors of a PB-SMT system. From the results present here we can conclude that lexical errors are the most frequent ones (44.95% of the total amount of annotated errors), mainly missing (15.20%) and incorrectly translated (14.05%) words. These errors can be automatically post-edited based on context information in Sys, Ref and Src and also on extra resources such as a bigger bilingual dictionary (or phrase-table in this case).

The second most frequent error category was inflectional errors (38.45% of the total amount of annotated errors), mainly number (13.75%) and gender (12.50%) agreement. These errors can be automatically post-edited with less effort, considering only the context information in Sys.

The automatic learning of post-edition rules capable of dealing with these errors is the next step of this research. To do so, superficial and linguistic motivated features will be extracted from the annotated corpus to train machine learning algorithms as described by Elming (2006) and George and Japkowicz (2005) among others.

Acknowledgements

We thank the Brazilian supporting agency FAPESP.

References

Avanço, L. V. 2012. (In Portuguese) Statistical Machine Translation: Post-Edition Rules ICMC-USP, São Carlos, 56p..Brazil..

Aziz, W. and Specia, L. 2011. "Fully automatic compilation of Portuguese-English and Portuguese-Spanish parallel corpora". In *Proceedings 8th Brazilian Symposium in Information and Human Language Technology* (STIL), p. 234-238. Cuiabá, Brazil.

Béchara, H., Ma, Y. and Genabith, J. V. 2011. "Statistical post-editing for a statistical MT system". In *Proceedings of the 13th Machine Translation Summit*, p. 308-315. Xiamen, China.

Caseli, H. M. 2007. (In Portuguese) Indução de léxicos bilíngües e regras para a tradução automática. Tese (Doutorado) – ICMC-USP, Abril 2007. 158 p.

Doddington, G. 2002. "Automatic evaluation of language translation using n-gram cooccurrence statistics". In *ARPA Workshop on Human Language Technology.*

Elming, J. 2006. "Transformation-based correction of rule-based MT". In *Proceedings of the 11th Conference of the European Association for Machine Translation* (EAMT).

George, C. and Japkowicz, N. 2005. "Automatic Correction of French to English Relative Pronoun Translations using Natural Language Processing and Machine Learning Techniques". In *Proceedings of the Computational Linguistics in the North East*. Ottawa, Canada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. 2007. "Moses: open source toolkit for statistical machine translation". In *Proceedings of the ACL 2007 demo and poster sessions*, p. 177-180. Prague, Czech Republic, June 2007.

Krings, H. P. Repairing Texts – Empirical Investigations of Machine Translation Post-Editing Processes. The Kent State University Press, 2001.

Lagarda, A. L., Alabau, V., Casacuberta, F., Silva, R. and Díaz-de-Liaño, E. 2009. "Statistical Post-Editing of a Rule-Based Machine Translation System". In *Proceedings of NAACL-HLT 2009*, p. 217-220. Boulder, Colorado, USA.

Och, F. J. and Ney, H. 2004. "The alignment template approach to statistical machine translation. *Computational Linguistics* (30/4): 417-449.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. 2002. "Bleu: a method for automatic evaluation of machine translation". In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, p. 311-318. Philadelphia, USA.

Popovic, M. and Burchardt, A. 2011. "From human to automatic error classification for machine translation output". In Mikel L. Forcada, Heidi Depraetere and Vincent Vandeghinste (eds.) *Proceedings of the 15th conference of the European Association for Machine Translation* (EAMT), p. 265–272. Leuven, Belgium, May 2011.

Potet, M., Esperança-Rodier, E., Blanchon, H. and Besacier, L. 2011. "Preliminary Experiments on Using Users' Post-editions to Enhance a SMT System". In Mikel L. Forcada, Heidi Depraetere and Vicent Vandeghinste (eds.) *Proceedings of the 15th Conference of the European Association for Machine Translation* (EAMT), p. 161-168. Leuven, Belgium, May 2011.

Simard, M., Goutte, C. and Isabelle, P. 2007. "Statistical Phrase-based Post-editing". In *Proceedings of NAACL-HLT 2007*, p. 508-515. Rochester, NY, USA.

Specia, L. 2011. "Exploiting Objective Annotations for Measuring Translation Post-editing Effort". In Mikel L. Forcada, Heidi Depraetere and Vicent Vandeghinste (eds.) *Proceedings of the 15th Conference of the European Association for Machine Translation* (EAMT), p. 73-80. Leuven, Belgium, May 2011.

Stymne, S. 2011. “Blast: a tool for error analysis of machine translation output”. In *Proceedings of the ACL-HLT systems demonstrations*, p.56-61. Portland, Oregon, USA. July 2011.

Vieira, T. L. and Caseli, H. M. 2011. (In Portuguese) “PorTAl: Automatic Translation Resources and Tools for Brazilian Portuguese ”. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology* (STIL), p. 179-183. Cuiabá, Brazil.

Vilar, D., Xu, J., D’Haro, L. F. and Ney, H. 2006. “Error analysis of statistical machine translation output”. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC), p. 697-702. Genova, Italy.

Corpus-driven terminology and cultural aspects: studies in the areas of football, cooking and hotels

Sabrina Matuda
University of São Paulo
sa_brina1@yahoo.com.br

Rozane R. Rebechi
University of São Paulo
rozanereb@gmail.com

Sandra Navarro
University of São Paulo
sandranavarro04@yahoo.com.br

1 Introduction

This paper sets out to describe how corpora can reveal cultural aspects of special subject fields, which contribute to the construction of comprehensive bilingual glossaries, taking into account the needs of a specific user, the translator. To that aim, we rely on data collected from research projects in three different areas – cooking (Rebechi forthcoming), football (Matuda 2011) and hotels (Navarro 2011).

These three areas carry significant cultural features of the country they refer to. Cooking, for instance, encompasses a series of traditional dishes, ingredients and techniques typical of a given community. Football is the favorite sport in countries like Brazil and England, where it is played by millions of fans in their own distinctive styles. Last but not least, when describing a hotel, one makes reference to kinds of establishments, services, architecture, décor, tourist attractions, sports, geography, history and several other characteristics typical of a region. These cultural features certainly pose a great challenge for translators.

These areas are constantly in need of updated and comprehensive reference materials. This is especially true in Brazil nowadays, where major sports events like the Olympic Games and the World Cup are being organized. Consequently, there is an increasing demand for translations of restaurant menus, hotel websites and football news, to name a few examples. In spite of the great demand, all three areas lack terminological publications, especially bilingual materials which would provide professionals with appropriate collocations and phraseologies, as well as contextualized examples extracted from authentic texts.

2 Corpus-driven translation and terminology

Translation-driven corpora are useful not only for the analysis of typical linguistic behavior but also for individual translation choices (Zanettin 2012). In this research, the study corpora allowed for the extraction of culture-specific terminological units and their equivalents in the target languages.

We also rely on the notion of textual terminology (Krieger and Finatto 2004). Under this perspective, specialized texts are the central object of study in terminology and several elements play an important role in term identification, such as textual type recognition, the culture in which the text is embedded, its purpose as a text, its intended target audience, its systematic traits and terminological density.

Our study also resorts to the notion of *equivalence in context* proposed by Chesterman (1998), according to which source and target units should play the same role within the same context in different languages. For Chesterman (1998), when establishing equivalents, the ideal scenario is to compare contexts rather than isolated terms. For example, take the case of *complimentary breakfast*, a collocation largely used in hotel descriptions. This collocation could be literally translated into Portuguese as *café da manhã de cortesia*, an expression that is possible, but rarely used according to our corpus research. Rather, the equivalent that plays the same role in the same context is *café da manhã incluído na diária* [breakfast included in the daily rate].

3 The corpora

In order to carry out our analysis, we compiled one corpus for each cultural area: cooking, football and hotels. The cooking corpus consists of a comparable and a parallel corpus. The comparable corpus comprises cookbooks originally written in Portuguese and in English and has a total of 252,875 and 428,290 tokens respectively. The parallel corpus contains recipes originally written in Portuguese and translated into English and has approximately 110,000 tokens in each language. For football, a comparable corpus was compiled from texts available on the Internet. This corpus consists of approximately two million words – 1,002,897 in English and 917,073 in Portuguese. Each corpus is divided into four subcorpora: laws of the game, newspaper reports on match results, live minute by minute commentaries by sports journalists and live commentaries by football fans via social media like *twitter* and *facebook*. For the hotels area, we have built a comparable corpus comprising texts extracted from hotel websites in Brazil and the USA. The corpus is divided up into hotel categories and contains 321 texts and 546,106 words in English; 710 texts and 514,449 words in Portuguese.

4 Analysis and results

Using the software *WordSmith Tools 6.0* (Scott 2012), we retrieved keywords from the Portuguese language corpora for cooking and football and from the English language corpus for hotels. By looking at those words in context (KWIC), their lists of collocates and clusters, we retrieved their main terminological patterns. The next step was to establish their equivalents by analysing lists of keywords, collocates, clusters and concordance lines in the target language subcorpora. Following these steps also enabled the identification of several cultural-specific terminological units, a very important feature of this paper.

To illustrate, take the case of the football term *chocolate* in Portuguese and its equivalent *cricket score* in English, used when a team wins by a large number of goals. The Portuguese expression was first used in the 1981 Brazilian championship when a team from Rio beat a team from Porto Alegre 4-0. At that time, one of the commentators played a Mexican song in which the chorus goes "*Toma chocolate / Paga lo que debes*" ('he drinks chocolate, he pays what he owes'). The English expression makes reference to the high scores in cricket and does not make sense if literally translated into Portuguese since cricket is not popular in Brazil. As it can be seen, in both cases, references to the original culture have proved difficult to be preserved.

Similarly, the area of cooking also abounds with cultural-specific terms, which may not have a direct standardized equivalent in a different language. For instance, consider a typical ingredient of Brazilian cooking: *farinha de milho*, a type of coarse meal made from cooked corn. The reference works looked up were not consistent in providing an appropriate English equivalent for the Brazilian term. Actually, most findings ('corn meal', 'corn flour', 'maize flour' etc.) could erroneously lead consultants to believe that it is a synonym for *fubá*, a kind of fine flour made from raw corn which, consequently, demands longer cooking time. By searching recipes in which the ingredient is essential in both the comparable and the parallel corpora, we came up with suggestions of equivalence which, although not standardized, reflected attempts of disambiguation, such as: *Brazilian flaky corn flour*, *flaky cornmeal*, *flaked cornmeal* and *coarse*

cornmeal.

Finally, a corpus-based terminological study of hotel descriptions also enabled the identification of several cultural-specific units. Take the example of *hotel-fazenda* [‘farm hotel’]. This collocation describes a famous typical kind of hotel in Brazil, usually found in the Southeastern states of Minas Gerais and São Paulo. These hotels are typically located in historical farms and are famous for offering typical country food and recreational activities envolving farm animals and local nature. In English, there are several establishments identified as *farm*, *country*, *ranch hotels/resorts/bed and breakfast* and so on, but they do not make up a distinct category as famous as the one in Brazil and references to the local culture are inevitably distinct.

Cultural information as exemplified above contributes to the construction of comprehensive bilingual glossaries aimed at translators. Below, we present an example of an entry from the Brazilian cooking glossary (Rebechi forthcoming), which includes cultural information alongside terms and their equivalents:

> **farinha de milho** *(noun)* flaked cornmeal White or yellow coarse meal made from pre-cooked corn and used in the preparation of *cuscuz* (*paulista*) and some kinds of *farofa*.//Ex.: *Break up the flakes of cornmeal with your hands, forcing it through the sieve. [BCJ]* **Compare with *fubá***

Using corpora proved particularly efficient for identifying culture-specific terms and their equivalents because the researcher deals with authentic texts in the same subject field in two different cultures which have distinct characteristics. Consequently, the corpus reflects two different realities.

5 Concluding remarks

Results show that corpora can expand the scope of terminological research by revealing cultural aspects of a special subject field through its linguistic patterns. We believe the advantage of these findings lies in the possibility to systematically include cultural aspects in reference materials, especially bilingual publications, thereby increasing translator’s cultural awareness.

References

Chesterman. A. 1998. *Contrastive functional analysis*. Amsterdam/Philadelphia: John Benjamins.

Krieger, M. G. and Finatto, M. J. B. 2004. *Introdução à Terminologia: teoria e prática*. São Paulo: Contexto.

Matuda, S. 2001. *A fraseologia do futebol: um estudo bilíngüe português-inglês direcionado pelo corpus*. (Master’s dissertation) – FFLCH/USP, São Paulo. Available online at http://www.teses.usp.br/teses/disponiveis/8/8147/tde-31102011-105346/pt-br.php

Navarro, S. 2011. *Glossário bilíngue de colocações da hotelaria: um modelo à luz da Linguística de Corpus*. (Master’s dissertation) – FFLCH/USP, São Paulo. Available online at http://www.teses.usp.br/teses/disponiveis/8/8147/tde-16082012-122119/pt-br.php.

Rebechi, R. R. (forthcoming) A Linguística de *Corpus* como metodologia para a compilação de um glossário de termos da culinária típica brasileira. In S. E. O. Tagnin and C. Bevilacqua (eds.) *Corpora na Terminologia*. São Paulo: HUB.

Scott, M. 2012. *Wordsmith Tools 6.0*. Oxford: Oxford University Press.

Zanettin, F. 2012. *Translation-driven corpus: corpus resources for descriptive and applied translation studies*. Manchester/Kinderhook: St. Jerome.

Is there a reputational benefit to hosting the Olympics and Paralympics? A corpus-based investigation

Tony McEnery
a.mcenery@lancaster.ac.uk

Amanda Potts
a.potts@lancaster.ac.uk

Richard Xiao
r.xiao@lancaster.ac.uk

Lancaster University

1 Introduction

This paper reports research undertaken by the ESRC Centre for Corpus Approaches to Social Science,[1] on behalf of the UK Department of Culture, Media and Sport, into the impact that hosting the Olympic and Paralympic games had upon the UK in general, and London and the host boroughs of the Games in particular.

2 Research questions

RQ1: (a) to what extent are issues of **disability** covered in the media; (b) how, if at all, has the representation of disability changed in recent years (in terms of volume of coverage, tone and attitudes/perceptions/article content); (c) to what extent is there evidence that the construction has shifted, e.g. from one in which disability is central to one in which disability is co-incidental to an individual's identity? (d) to what extent may the Paralympics have contributed to a general shift of representations and constructions of disability?

RQ2: how did hosting the Olympics and Paralympics impact upon the **reputation** of the UK and London in the UK and beyond, focussing on the UK's reputation as a place to visit, invest in and do business with?

RQ3: to what extent did the Olympics and Paralympics alter perceptions of **East London** as a place to live, work and invest in?

In addition, we consider the question of the representation of the UK in the Chinese media.

3 Method

We use collocations, keywords and semantic field analysis to produce our results. We ensure rigour in our collocation analysis by looking for collocates which occur a minimum of 10 times in the data and which occur with a high degree of statistical significance – well beyond the 99.9% confidence level. This helps to strengthen our confidence in the findings we produce – they are highly unlikely to be due either to chance or to distorting effects produced by a handful of atypical examples.

In our keyword analyses we focus on the 'key' keywords, those which have been shown to be most markedly key. This adds weight to our results – the keyword effects are derived from very large datasets and we focus on are the most key of the observed keywords. As with collocation, the keywords we are looking at are significant well beyond the 99.9% confidence level, meaning that the results are highly unlikely to be the result of random chance.

4 Existing datasets used

For this study we both collected new data and reused existing datasets. Three existing datasets were used as reference corpora:

- The enTenTen08 corpus – a general corpus of English from 2008 (2.8 billion words)
- The enTenTen12 corpus – a general corpus of English from 2012 (11.2 billion words)
- Materials archived for 2012 at the most important news hub in China, People's Net (people.com.cn: this includes the *People's Daily* and many other influential newspapers in China).

5 Specialised datasets constructed

Furthermore, we constructed the following specialized corpora:

RQ1 – UK National Newspaper Disability Corpus (UKDC): A corpus of UK national press articles using at least one of the following terms: *person with a disability*, *disabled*, *wheelchair user*, *wheelchair-user*, *uses a wheelchair*, *handicapped*, *cripple*, *crippled*, *wheelchair bound*, *wheelchair-bound*, *confined to a wheelchair*, *differently able* and *handicapable*. We collected articles in each month from 1/1/05 to 31/12/12; total size 52.8 million words (MW) across 70,667 articles.

RQ1 – Global Media Disability Corpus (GMDC): A companion corpus for the UKDC was also constructed from a broad set of major world news publications. For each month from 1/1/05 to 31/12/12, 100 random articles containing the same search terms governing collection of the UKDC were collected, resulting in a corpus of 6 MW.

[1] ESRC grant reference ES/K002155/1

RQ2 – UK National News mentions of England, London and the UK: To explore RQ2 we constructed a corpus of newspaper articles focused on the topics of the economy, trade and investment. Articles mentioning *England*, *London* or the *UK* were gathered from the UK national press from 1/1/05 to 31/12/12. We collected all of the articles, up to a monthly maximum of 500, for 40.4 MW in total containing 1173, 2828 and 4025 mentions per MW of *England*, *London* and *UK* respectively.

RQ2 – Global Media mentions of England, London and the UK: This corpus was sampled using the same terms as the corresponding UK corpus, from the same newspapers as the GMDC. It is 26.8 MW in size and contains 536, 1821 and 188 mentions per MW of *England*, *London* and *UK* respectively.

RQ3 – The UK National News Mentions of the East End, East London and the Host Boroughs: We gathered newspaper reports in which any of the terms *barking and dagenham*, *greenwich*, *hackney*, *newham*, *tower hamlets*, *waltham forest*, *east end* or *east london* were mentioned, both: 1) in the first half of 2012; and 2) linked to four broad topics: the economy, economic indicators, trade and development trade and investment. Given that we wished to explore the impact of the holding of the Games, we further divided each data set into two periods – the months before the Games (January to June 2012: 1.3 MW) and the months during/after the Games (July to December: 0.6 MW).

RQ3 – The Global Media Mentions of the East End, East London and the Host Boroughs: This is a Global Press counterpart to the preceding corpus, gathered using the same search terms and time parameters, and 2.4 MW in size (January to June: 1.3 MW; July to December:: 1.1 MW).

6 Findings

RQ1 – Overall, we conclude that the 2012 London Paralympics *has* had an effect on the construction of disability in the British press. In the UK press, there is now an increased use of preferred ways of referring to disabled people (e.g. *person with a disability, wheelchair user*, *disabled person*), while the use of dispreferred ways of referring to such people (e.g. *cripple*, *wheelchair bound*, *handicapped*) is in sharp decline. Use of the word *disabled* was also analysed 2,400 concordance lines from 2005-2012. We found that over time, there has been a general tendency toward increasing DESCRIBING uses of the word *disabled* (e.g. in the attributive adjective position, "her disabled daughter"). This language is more empowering and democratic (Mautner 2007), construing disabled people as having many different traits, of which their disability is only one. In the year of the Paralympics, ESSENTIALISING uses of *disabled*, which define people solely on their disability (e.g. "Bruno hated the disabled"), dropped markedly in the UK media, indicating a media consensus that such use is disempowering. In UK press reporting around the Games, differently abled people were represented as leading a more active part in society. British English, when compared to American English, seems to be 'leading the way' in the progressive discussion of disability.

RQ2 – There is evidence to show that hosting the Olympics and Paralympics has impacted positively upon the reputation of the UK and London both in the UK and beyond both in terms of the UK as a place to visit and in terms of the UK as a place to invest in and/or do business with. London has experienced both a sustained and positive association with the Games over the period from 2005, with the association intensifying in 2012. The Games helped present London as a city which is transforming itself in a wholly positive way. There is a markedly increased frequency of reporting on East London in general and Stratford in particular in the context of regeneration and investment and less in terms of poverty and welfare dependence, as seen in pre-Olympics reporting. In general English, the positive associations the Games have brought to London have been strengthened by a link being formed between the Olympics and the Diamond Jubilee.

RQ3 – we believe that the Olympics and Paralympics have altered perceptions of East London in particular as a place to live, work and invest in. This shift is positive, moving discussion of East London away from what seemed to be an almost exclusively negative discussion focussed upon poverty and welfare dependence towards a more positive discussion focussed upon regeneration and investment. The host boroughs gain positive associations via their identification with East London, just as the UK gains a reputational boost, we would argue, by the positive associations that London attracted through hosting the Games.

Finally, in the Chinese language press, the Games seem to switch the discussion of the UK away from a wholly negative discussion focussing upon a faltering economy and difficult military engagements overseas to a more positive discussion of positive economic activity, the Games and tourism. The intensity of the association between the UK and tourism in the

Chinese press in the second half of 2012 is very marked indeed. Equally marked is an intensified association between the UK and economic activity in the context of discussions of tourism.

References

Mautner, G. (2007). "Mining large corpora for social information: The case of elderly." *Language in Society*, 36, pp 51-72.

Take a mirror and *take a look*: Reassessing usage of polysemic verbs with concrete and light senses

Seth Mehl
University College London
seth.mehl.10@ucl.ac.uk

1 Introduction

Recent studies have compared corpus-based frequencies of various senses of polysemic words to the cognitive salience of those senses as determined by elicitation tests (cf. Gilquin 2008; Nordquist 2009). For example, Gilquin (2008) compares the various senses of *take* and *give* in the *Switchboard* corpus of spoken English to cognitive salience of those senses as evidenced by elicitation tests. She concludes that 'while language [in corpora] shows a strong preference for abstract, grammaticalised senses such as the delexical [light] use, the senses most often elicited are more concrete' (ibid: 248). Various researchers have speculated on the theoretical problems that this discrepancy raises for cognitive linguistic frameworks (cf. Gilquin 2006, 2008; Nordquist 2009; Arppe et al. 2010; Geeraerts 2010). Two key questions arise from the established research on the subject. First, do light senses of polysemic verbs occur more frequently than concrete senses if both speech and writing are analysed across multiple varieties and using multiple modes of analysis? And, second, can a correlation ultimately be found between frequency in use and cognitive salience?

In order to address these questions, manual semantic analysis was conducted on the English verbs *take* and *make*, polysemic verbs with light senses (e.g. *take a look*; *make a decision*) and concrete senses (e.g. *take the book*; *make a cake*), in the International Corpus of English (ICE). Speech and writing were examined in data from three regions: Singapore, Hong Kong and Great Britain. Both semasiological and onomasiological analyses were conducted. Findings show that relative preferences for concrete and light senses vary between speech and writing and between regional varieties. Onomasiological analyses reveal closer similarities between corpus frequencies and cognitive salience than previously seen.

2 Modes of analysis

Methodologically, the present study approaches semantics both semasiologically and

onomasiologically. Existing studies by Gilquin (2008) and Nordquist (2009) employ only a semasiological approach, which compares the usage frequencies of various senses of a single word (e.g. the concrete sense compared to the light sense). However, language users do not generally select between a light sense and a concrete sense of a verb in a given situation; rather, they select between a particular sense and a semantic alternate or near-synonym. For example, speakers and writers choose between *make* and *produce*, or *take* and *collect*, in specific contexts. An onomasiological analysis addresses this fact by comparing a particular sense of a word to its respective semantic alternates. [1] Relative to semasiological analyses, onomasiological analyses more accurately reflect psycholinguistic processes and user preferences (Geeraerts et al. 1994) and allow for more reliable statistical modeling (Wallis et al. 2012). Geeraerts (2010) has hypothesized that high relative onomasiological frequencies should correlate with high cognitive salience; he calls this correlation *onomasiological salience*.

3 Semasiological analysis

Semasiologically, the ICE corpora show that usage of concrete and light senses of *make* and *take* varies according to region and lexical item (Figures 1 and 2).

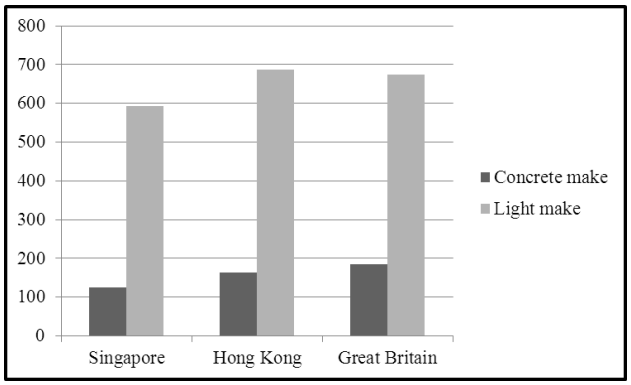


Figure 3. Occurrences of concrete and light make in 3 ICE corpora

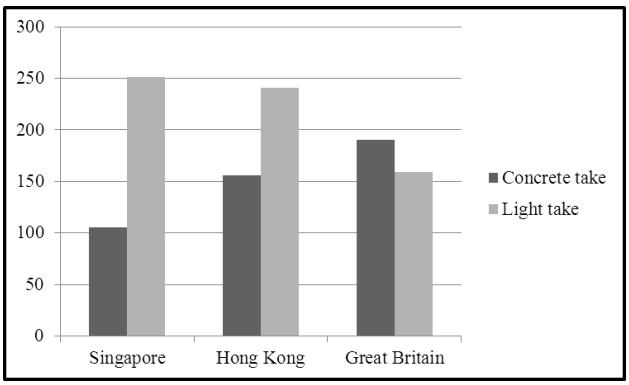


Figure 2. Occurrences of concrete and light *take* in 3 ICE Corpora

Present findings for *make* in the ICE corpora corroborate previous findings on the high relative frequency of the light sense (cf. Jespersen 1954; *Collins COBUILD English Dictionary* 1996; Gilquin 2008). Present findings for *take* in the ICE corpora, however, run counter to previously observed trends. Thus, semasiological data from the ICE corpora reveal more complexity in use than has been previously reported.

Because semasiological data do not closely reflect psycholinguistic processes, it is difficult to draw firm conclusions from Figures 1 and 2. The following sections report on onomasiological analyses, which facilitate stronger conclusions and more firm findings.

4 Onomasiological analysis: Concrete *make*

As shown in Figure 1, concrete *make* is less common than light *make*. Onomasiologically, however, concrete *make* is generally more common than its semantic alternates (e.g. *produce*, *create*, etc.). Figure 3 shows spoken data and Figure 4 shows written data for concrete *make* and its alternates in the 3 corpora.[2]

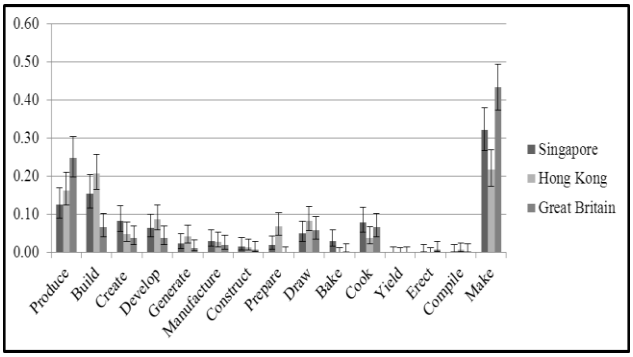


Figure 3. Probability of choosing concrete *make* and semantic alternates in spoken language in 3 ICE corpora

[1] In the present study, semantic alternates of concrete senses are determined by identifying alternate concrete verbs that appear with the same direct objects as concrete *make* and *take*. Semantic alternates of light senses are the verbs derived from the direct objects in the light constructions: e.g. 'to make a decision' is 'to decide'; 'to take action' is 'to act'.

[2] Error bars in all graphs represent Wilson score intervals.

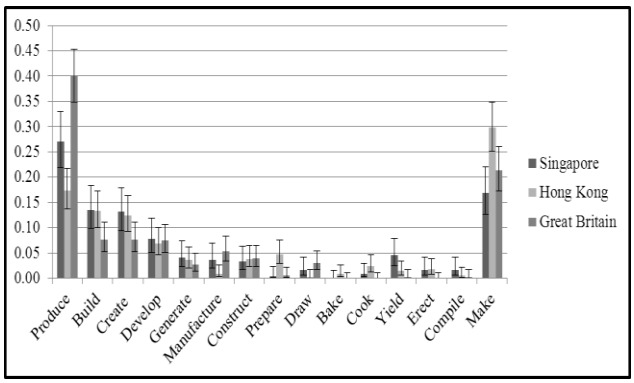


Figure 4. Probability of choosing concrete *make* and semantic alternates in written language in 3 ICE corpora

Significant variation occurs between speech and writing and between regions. It appears that high cognitive salience correlates with high relative onomasiological frequency in spoken language. A significant exception to that correlation is found in Hong Kong usage preferences. If Geeraerts's (2010) hypothesis of onomasiological salience is valid, then Hong Kong English speakers might be expected to exhibit cognitive salience patterns for *make* different to those of speakers in Singapore and Great Britain. Future elicitation testing in each region could pinpoint such patterns and corroborate or refute that expectation.

5 Onomasiological analysis: Concrete *take*

Onomasiologically, concrete *take* tends to be more common than its alternates (e.g. *collect*, *carry*, etc.). Figure 5 shows spoken data and Figure 6 shows written data for concrete *take* and its alternates in the 3 corpora.

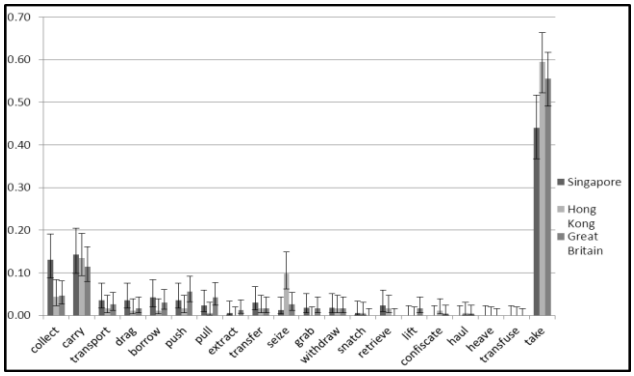


Figure 5. Probability of choosing concrete *take* and semantic alternates in spoken language in 3 ICE corpora

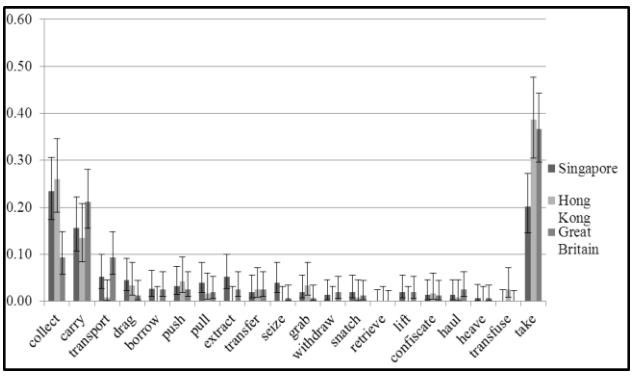


Figure 6. Probability of choosing concrete *take* and semantic alternates in written language in 3 ICE corpora

As with concrete *make*, variation appears between speech and writing and between regions; the high cognitive salience of the concrete sense correlates most strongly with onomasiological frequency in spoken language. According to the onomasiological salience hypothesis, slight variation in cognitive salience patterns might appear between varieties, a possibility that could be corroborated or refuted via elicitation tests in each region.

6 Onomasiological analysis: Light *make* and *take*

Semasiologically, Figures 1 and 2 have shown that light senses tend to be more common than concrete senses. Onomasiologically, light senses tend to be less common than their alternates. Variation appears between regions and lexical items, but very little variation appears between speech and writing. Figures 7 through 12 therefore display aggregated spoken and written data.

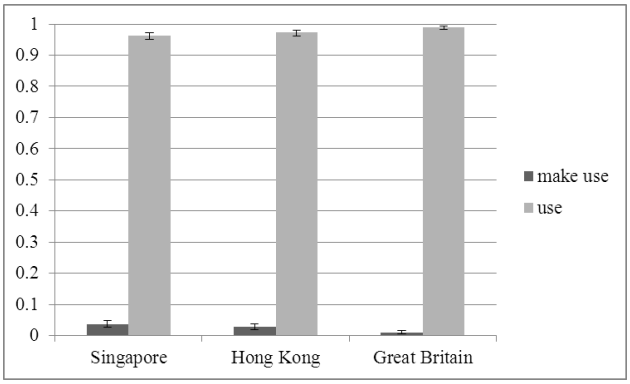


Figure 7. Probability of selecting *make use* and *use* in 3 ICE corpora

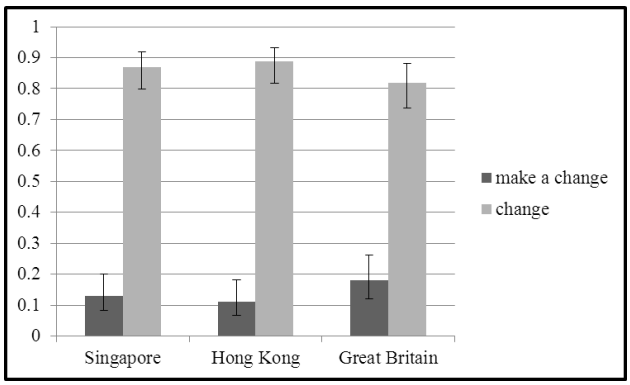


Figure 8. Probability of selecting *make a change* and *change* in 3 ICE corpora

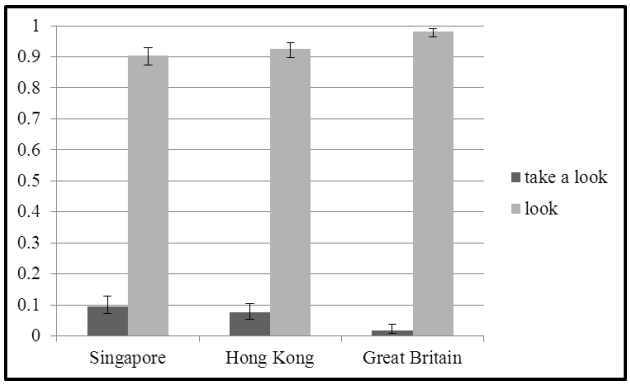


Figure 9. Probability of selecting *take a look* and *look* in 3 ICE corpora

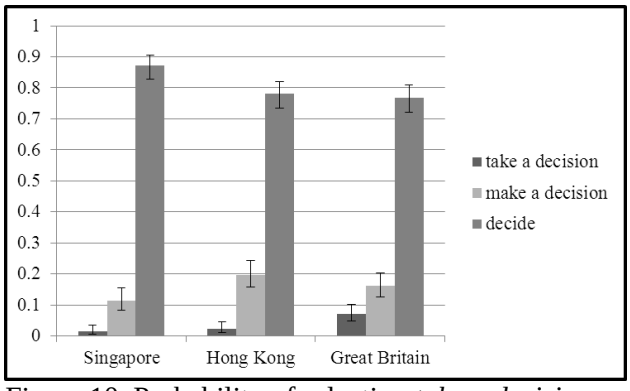


Figure 10. Probability of selecting *take a decision*, *make a decision* and *decide* in 3 ICE corpora

Relatively low onomasiological frequencies for most light constructions correlate strongly with relatively low cognitive salience for light constructions. Exceptions occur with the constructions *take care* and *take action*, as shown in Figures 11 and 12.

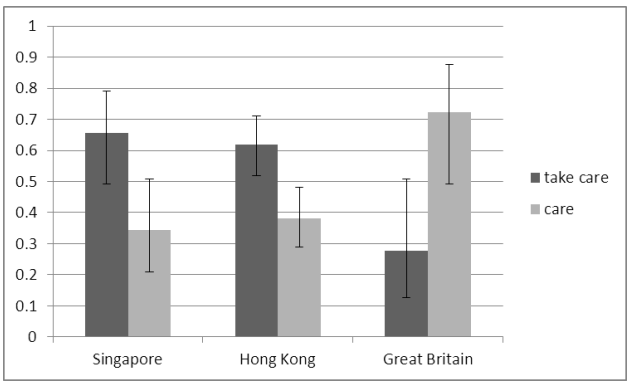


Figure 11. Probability of selecting *take care* and *care* in 3 ICE corpora

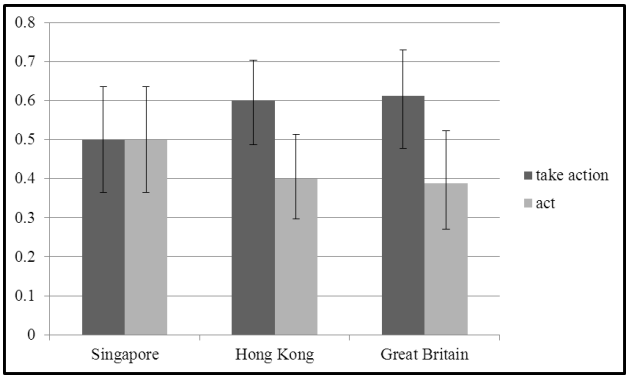


Figure 12. Probability of selecting *take action* and *act* in 3 ICE corpora

According to the onomasiological salience hypothesis, speakers from different regions should exhibit varying cognitive salience patterns for *take care* and *take action*. Further elicitation testing in each region could corroborate that expectation.

7 Conclusions

Contrary to previous findings, semasiologically, light senses are not consistently more common than concrete senses, as shown in Figures 1 and 2. Onomasiologically, concrete senses often tend to be relatively frequent while light senses often tend to be relatively infrequent. However, variation appears and exceptions occur for each lexical item, text type (speech vs. writing) and regional variety.

High cognitive salience trends generally tend to correlate with strong onomasiological preferences. The resemblance is often strongest in spoken language. Geeraerts's (2010) hypothesis might be revised, therefore, to propose specifically that relatively high onomasiological frequencies *in spoken language* correlate with cognitive salience. Spoken language can be seen as less mediated than written language, and therefore might be expected to resemble cognitive salience patterns most strongly. In addition, differing stylistic standards influence psycholinguistic selection processes in speech and writing. Further

elicitation tests, both spoken and written, could address whether cognitive salience for specific lexical items and constructions varies across the three regions (cf. 'cognitive sociolinguistics' in Heylen et al. 2008; Geeraerts et al. 2010).

A mechanism driving the variation (and change) in usage of *make* and *take* in Singapore, Hong Kong and Great Britain has yet to be identified. Semantic change from pragmatic implicature; from historical and sociolinguistic factors of English; or from contact influences are all possibilities.

Regional variation does not seem to conform to existing theories of World Englishes (cf. Kachru 1985; Schneider 2007). No consistent trends across inner or outer circle varieties are apparent, and no consistent trends across established exo-normative or endo-normative categories are apparent. Existing theories of World Englishes tend not to focus on lexical semantic variation and change across global varieties. Further research like the present study can contribute to the construction of new models of lexical semantic variation and change across World Englishes.

References

Arppe, A., G. Gilquin, D. Glynn, M. Hilpert and A. Zeschel. 2010. "Cognitive corpus linguistics: Five points of debate on current theory and methodology". *Corpora* 5 (1), 1-27.

Collins COBUILD English Dictionary. 1996. 2nd edn. London: Collins.

Geeraerts, D. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.

Geeraerts, D., S. Grondalaers and P. Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Geeraerts, D., G. Kristiansen, and Y. Peirsman. 2010. "Introduction: Advances in Cognitive sociolinguistics". In D. Geeraerts, G. Kristiansen and Y. Peirsman (eds) *Advances in Cognitive Sociolinguistics*, 1-19. Berlin: Mouton de Gruyter.

Gilquin, G. 2006. "The place of prototypicality in corpus linguistics: Causation in the hot seat". In S. Gries and A. Stefanowitsch (eds) *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter. 159-191.

Gilquin, G. 2008. "What you think ain't what you get: Highly polysemous verbs in mind and language". In J. R. Lapaire, G. Desagulier and J. B. Guignard (eds) *Du fait grammatical au fait cognitif. From Gram to Mind: Grammar as Cognition*. Bordeaux: Presses Universitaires de Bordeaux. 235-255.

Heylen, K., J. Tummers & D. Geeraerts. 2008. "Methodological issues in corpus-based Cognitive Linguistics". In G. Kristiansen and R. Dirven (eds) *Cognitive Sociolinguistics: Language variation, cultural models, social systems*. Berlin: Mouton de Gruyter. 91-128.

Jespersen, O. 1954. *A Modern English grammar on historical principles, part VI: Morphology*. London: George Allen & Unwin.

Kachru, B. B. 1985. "Standards, codification, and sociolinguistic realism: The English language in the outer circle". In R. Quirk and H. G. Widdowson (eds) *English in the world: Teaching and learning the language and literatures.* Cambridge: CUP and The British Council. 11-30.

Nordquist, D. 2009. "Investigating elicited data from a usage-based perspective". *Corpus Linguistics and Linguistic Theory* 5(1), 105-30.

Schneider, E. W. 2007. *Postcolonial English: Varieties around the world*. Carmbridge: Cambridge University Press.

Wallis, S., J. Bowie and B. Aarts. 2012. "That vexed problem of choice: Reflections on experimental design and statistics with corpora". London: UCL Survey of English Usage. http://www.ucl.ac.uk/english-usage/staff/sean/resources/vexedchoice.pdf. Accessed 1 August, 2012.

A corpus linguistic study of ellipsis as a cohesive device

Katrin Menzel
Saarland University
k.menzel@mx.uni-saarland.de

1 Introduction & Motivation

This paper has a focus on ellipsis as a cohesive device. As a part of the DFG-funded project GECCo – German-English contrasts in cohesion – Towards an empirically-based comparison, ellipsis is analysed quantitatively although it faces two important challenges: a) scarcity of data: We did not find too many occurrences in our corpus of about 1 million words. b) Complex concept: There are several definitions of ellipsis from various perspectives. Ellipsis seems to be a gradual notion with prototypical and less prototypical or marginal cases and very similar other phenomena, such as substitution (Halliday & Hasan, 1976:88). Obviously, the analysis of data and outcome of studies on ellipsis will depend heavily on its definition and subclassification. Ellipsis as a cohesive devise has been studied less intensively than other cohesive phenomena. Corpus linguistic studies on this topic are rare or they usually focus on small subcategories, often in English (e.g. clausal coordinate ellipsis, Muhonen & Purtone, 2012; nominal ellipses after adjectives in English, Günther, 2012). There are no tools yet to efficiently spot and annotate ellipsis automatically (cf. Bos & Spenader, 2011) and it is not possible to query ellipses as such as they are empty elements. However, they occur in the environment of certain syntactical structures or trigger words. Therefore, systematic query patterns to spot ellipses in electronic corpora can be developed. So far there are mainly monolingual accounts of ellipsis. However, in the past few years, some cross-linguistic studies have appeared, e.g. in the framework of transformational grammar (cf. Aelbrecht 2010 on Dutch and English, Merchant & Simpson's 2012 cross-linguistic perspectives on sluicing). In our bilingual comparable and translation corpus GECCo, we can also observe what happens with ellipses in translation, whether they are kept, inserted or replaced, which depends on complex factors such as translation methods and procedures (explicitation / implicitation, degree of modification or faithfulness to the source text) and language-internal aspects. This research should lead to a general description of the way in which ellipsis contributes to text cohesion, based on empirical findings and cross-linguistic comparison.

2 Corpus resources and methods

This paper builds on the SFL-based classification of cohesive devices by Halliday & Hasan (1976) and SFL-based corpus linguistic studies. However, cohesive ellipsis does not seem to be a clearly categorical notion. Probably the most obvious, even if not the most frequent, cases of nominal ellipsis, for example, are those after adjectives. A scale of phenomena that fall under the notion of cohesive nominal ellipses could be as follows: ellipses after adjectives / classifier nouns > numeratives > quantifiers > possessive pronouns > demonstrative pronouns.
The GECCo corpus provides English and German texts of various registers along the written / spoken continuum. The written part of the corpus has English and German original texts that are aligned with their German or English translation. The spoken part is a comparable corpus as translations are not professionally produced for this kind of discourse. GECCo is tagged for: tokens, lemmas, morpho-syntactic information, parts of speech, chunks and sentence boundaries. Annotation of ellipses and their antecedents is currently done with MMAX2, an open source annotation tool. The corpus can be queried with the Corpus Query Processor (CQP) (Evert 2005). It is not possible to query ellipses as such as they are empty elements. However, they seem to occur in the environment of certain syntactical structures or trigger words. Therefore it was possible to formulate CQP-based queries to find potential candidates of ellipsis in the corpus (see Table 1).

Queries to find nominal ellipsis are based on typical structures that might trigger them: e.g. articles / determiners / adjectives / numerals / possessive markers not followed by nouns (cf. Table 1). Query patterns to find VPE with CQP may involve verbless clauses or clauses that include modal, but no lexical verbs, lexical verb ellipsis in comparative constructions (He can run faster than Jane can [ ].), before conditional clauses (e.g. I can [ ] if I want.) etc. Clausal ellipsis can be found by querying adjacency pairs, clauses consisting of one or very few constituents or, in the case of sluicing, with queries of wh-words at the end of a clause. The queries can be adapted to fit German word order patterns. Some structures do not exist in German (operator ellipsis, substitution with do, inclusive imperative followed by verbal ellipsis etc.). The list of potential candidates for ellipsis found by corpus queries had to be disambiguated manually,

particularly to exclude other types of ellipsis and fragments. Additionally, some entire registers were looked through manually to find all cases of cohesive ellipsis. This served as a comparison with ellipses found by CQP queries.

Pattern nominal ellipsis	**CQP query design**
1. nominal ellipsis after article / determiner / numeral / quantifier / possessive marker (+optional adjective)	e.g. in German subcorpora (Stuttgart-Tübingen-TagSet STTS) [pos='adja'][pos='vafin']; (adjective + finite verb); [pos='art'][pos='adja'][pos!='nn|ne']; (article + adjective, not followed by noun/proper noun) in English subcorpora (Penn Treebank tagset): [pos='jj'][pos='vv.*']; (adj. + verb)
2. possessive marker 's not followed by noun	in English subcorpora: [word='s'][pos!='nn|ne']

Table 1: Samples for CQP queries to find nominal ellipsis

3 Results

The CQP queries worked best for all subtypes of nominal ellipsis in both languages. There was a high recall for nominal ellipsis, but unfortunately precision was still relatively low; more sophisticated queries have to be designed in the future. Some subtypes of verbal and clausal ellipsis were easy to query with CQP (e.g. sluicing) others were more difficult to spot.

Nominal ellipsis occurs mainly in certain text types (e.g. texts with many adjectives and nominal style or limited space for printing) but also in the context of certain topics (involving numerals, comparisons, contrasts…). Written discourse in general offers more possibilities for nominal ellipsis due to structural complexity, lexical density, nominalization and longer noun groups. Nominal ellipsis is assumed to be more frequent in German because English can use substitution with *'one'* instead and avoids nominal ellipsis when it could lead to ambiguity due to the morphological characteristics of the language. Surprisingly in our corpus, nominal ellipsis is more frequent in both spoken and written English than in German. Verbal ellipsis often co-occurs with proper names or personal pronouns and therefore also depends on the text topic and the level of interaction in discourse. Verbal ellipsis typically occurs in text types with otherwise rich verb phrase structures to avoid verbal repetition; some subtypes require hypotactic or parallel structures, often involving contrasts between two or more members of a semantic category (e.g. *The parents ate cake, and the children [ ] cookies).* Due to the lack of exact correspondence between the English and German verbal system, there are more differences between English and German verbal ellipsis than with regard to nominal ellipsis.

Filler words, redundancies, anacolutha and less clear sentence boundaries in spoken language make queries for spoken registers more difficult than for written registers. Often larger parts of texts have to be taken into account to determine whether a certain structure is an ellipsis, an anacoluthon or a sentence break, regional variation or simply an error where people forgot to complete a sentence. The special syntax of spoken language and differences between English and German morphology (high frequency of zero derivation / word-class ambiguities in English, declension of adjectives / pronouns as ellipsis remnants in German) have to be considered in queries as well as the particularities of ellipses as syntactically incomplete or – without an appropriate context – even deficient structures. Tagging errors resulting from these untypical syntactic patterns are another aspect that has to be taken into account when formulating CQP corpus queries. Table 2 shows the normalized frequencies of ellipsis subtypes per 100,000 words found in 4 German and English registers of GECCo.

	Nominal ellipsis	verbal	clausal	∑
GO Interview	62.2	9.7	42.2	114.1
EO Interview	129.3	58.0	42.2	229.5
GO Academic	124.4	9.8	43.9	178.1
EO Academic	131.0	29.7	12.4	173.1
GO Fiction	114.2	38.1	51.7	204.0
EO Fiction	154.1	37.8	27.0	218.9
GO Tourism	24.6	13.7	16.4	54.7
EO Tourism	52.9	5.6	0	58.5
Ø of spoken registers	111.73	26.8	35.18	173.71
Ø of written registers	86.45	23.80	23.77	134.02
Ø of GO registers	81.35	17.83	38.55	137.73
Ø of EO registers	116.83	32.78	20.40	170.01

Table 2: Cohesive ellipsis in 4 registers of GECCo (GO = German Originals, EO = English Originals; spoken registers: Interview + Academic lectures; written registers: Fiction/novels + Tourism leaflets)

Cohesive ellipsis seems to be more frequent in spoken than in written language. However, nominal ellipsis, which is the most frequent type of cohesive ellipsis in all registers, is found very often in fictional written texts (due to the high frequency of adjectives / numerals and its similarity to spoken language, although admittedly fiction is a rather heterogeneous register). There would probably also be a high number of nominal ellipses in academic writing as there are generally many NPs in this genre. Similarly, in the spoken register of academic lectures in our corpus there are more cases of nominal ellipsis than in interviews. Clausal ellipsis is typical for dialogues (e.g. in fiction) and spoken language in general, also because there are more clauses in speech. Ellipsis of the lexical verb is typical in English after modal and auxiliary verbs and less frequent in German. For other types of verbal ellipsis such as gapping and sluicing to be possible, relatively long hypotactic sentences are required. These types were found more frequently in written registers, but they were still rare. In our spoken data, sometimes quite the opposite of ellipsis was observed: in structures where ellipsis would have been a possible strategy to avoid repetition, it was not used.

After analysing the phenomenon of ellipsis cross-linguistically in original texts, it might be interesting to look at what happens in translation, e.g. whether there is an influence of the source language and translators leading to less usual constructions in translated texts. It might be possible that English translations of German texts, for instance, include a higher frequency of nominal ellipsis after adjectives where we would normally expect *'one'*, e.g. *The grey fox is not as flamboyant as the red [ ]* or higher frequencies of *'one'* as a substitute where it is not obligatory (e.g. after *'next'*, *'second'*, *'another'*, *'which'*). To give just one example, for nominal ellipsis in EO FICTION (154.1 / 100,000 words), most cases (115.6) were kept in the translation (probably because nominal ellipsis works very similar in both languages), only a few cases were replaced by nouns or pronouns, and 37.84 additional nominal ellipses were inserted in the German translation per 100.000 words. This might indicate that German uses nominal ellipsis more often; on the other hand GO FICTION only has 114.2 cases of nominal ellipsis / 100,000 words.

While it is generally assumed that ellipsis (e.g. exophoric / situational) and fragments are a typical feature of spoken language, it is questionable whether the same is true for ellipsis as a cohesive device. Initially one could think that cohesive ellipsis is very frequently used in spoken language. However, looking through the spoken registers in GECCo, it becomes intuitively obvious that ellipsis as a cohesive device is less frequent than expected in our spoken data (although the total numbers are still a bit higher for the spoken subcorpora than for the written corpus part). In general, it might be the case that cohesive ellipsis is a relatively infrequent phenomenon for a quantitative corpus analysis, despite the fact that the phenomenon as such has been described extensively in the literature, e.g. verbal and clausal ellipsis with their complex system of possible subtypes. The reason for the total numbers of ellipses being lower than expected in the spoken data and not remarkably higher than those found in written texts might be that both prepared academic lectures and prepared interview texts share similarities with written language, that not all less prototypical subcategories were included in the analysis, or that spoken language tends to have a lot of redundancies and repetitions. We are planning to make a comparison between several data sets (including only prototypical categories vs. prototypical + less clear / borderline subcategories). We are also extending our corpus to include more spoken registers to obtain more robust statistical results.

References

Aelbrecht, L. 2010. *The Syntactic Licensing of Ellipsis.* John Benjamins, Amsterdam/New York.

Bos, Johan and Spenader J. 2011. An annotated corpus for the analysis of VP ellipsis. Language Resources and Evaluation 45(4): 463-494.

Evert, S. 2005. *The CQP Query Language Tutorial.* IMS, Universität Stuttgart.

GECCo Project website: http://134.96.85.104/gecco/GECCo/Home.html

Günther, C. *The Elliptical Noun Phrase in English: Structure and Use*. New York: Routledge. 2012

Halliday, M.A.K and Hasan R. 1976: *Cohesion in English.* London: Longman.

Kunz, K. and Lapshinova-Koltunski, E. 2011. *Tools to Analyse German-English Contrasts in Cohesion.* In proceedings of GSCL-2011, Hamburg, Germany.

Lapshinova-Koltunski, E., Kunz, K. and Amoia M. (forthcoming). *Compiling a Multilingual Spoken Corpus.* In Proceedings of GSCP-2012, Belo Horizonte, Brazil.

Merchant, J. and A. Simpson (eds.). 2012. *Sluicing: Cross-linguistic perspectives.* OUP: Oxford.

Muhonen, K. and Purtone, T. 2012. *Rule-Based Detection of Clausal Coordinate Ellipsis.* In: Proceedings of the eighth conference on

International Language Resources and Evaluation (LREC'12) European Language Resources Association (ELRA), Istanbul, Turkey.

Müller C. and Strube M. 2006. *Multi-Level Annotation of Linguistic Data with MMAX2*. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3)

Student perceptions of university instructors: A multi-dimensional analysis of free-text comments on RateMyProfessors.com

Neil Millar
University of Birmingham
n.j.millar@bham.ac.uk

1 Background and aims

This paper presents a large-scale analysis of free-text comments in student evaluations of teaching posted on RateMyProfessors.com (RMP). Student evaluations of teaching (SETs) play an important role in universities and this is reflected in countless published studies on and long-running debate concerning their validity. Recent developments in the way in which SETs are administered and the results are used and disseminated have fuelled this debate. One of the most profound changes has been the growth of review websites on which students anonymously post and read evaluations of their instructors. The largest and most influential of such sites is RMP.

Founded in 1999, RMP now contains c. 15 million evaluations of c. 1.7 million instructors at over 8,000 institutions in the USA, Canada and the UK (RMP 2012). Students who log onto the site anonymously evaluate instructors by providing a numerical score (1-5) for *helpfulness*, *clarity* and *easiness*, as well as optionally providing text comments. In addition, and somewhat controversially, students can optionally award their instructors a chilli mark for '*hotness*'. Evaluations posted on RMP are very influential. Primarily, they inform students' in decisions about classes, but also have implications for institutions and instructors – e.g. scores are factored into the Forbes' annual ranking of US universities (Forbes/CCAP 2011), and there is evidence (albeit anecdotal) that evaluations on RMP are used in staff appraisals (Sanders et al. 2011).

To date, there have been no large-scale analyses of what is perhaps the most informative (and controversial) component of this evaluation system – the unregulated free text comments. Seeking to inform debate surrounding the validity of SETs posted on RMP, the research reported here used a factor analytic approach to objectively group adjectives occurring in text comments on RMP, and, thus, identify dimensions along which

university students commonly perceive their instructors.

2 Method

The framework for analysis started with the assumption that text comments made on RMP will reflect students' perceptions of the most salient and personally relevant characteristics of their instructors. In much the same way that numerical rankings show internal reliability (Bleske-Rechek and Fritsch 2011), we might expect that students will, for a given instructor, also tend to agree on the most comment-worthy characteristics. This agreement will be evidenced by the use of over-lapping linguistic forms to index comparable characteristics. For example, comments about instructors generally perceived as 'helpful' are likely to contain adjectives such as *helpful*, *available*, *willing* and *approachable*, while comments about instructors perceived as 'funny' instructor would contain adjectives such as *funny*, *hilarious*, *witty* and *amusing*. Principal Components Analysis (PCA) was used to identify how adjectives naturally cluster based on their use over a large number of individual instructors. This builds on Chung and Pennebaker's (2008) implementation of PCA, an approach which they term the meaning extraction method, and shares features with Latent Semantic Analysis (Foltz 1996).

A corpus of text comments about instructors employed at the top-ranking 200 US universities was constructed by downloading publically available data from RMP. This contained 467,904 ratings distributed over 50,316 individual instructors – c. 25 million words. Text comments were annotated for parts of speech, and this annotation was used to extract adjectives in an attributive or predicative relationship with a noun or pronoun referring to the instructor (e.g. *he is helpful*, *she is a great teacher*). The adjectives were checked, spelling mistakes/variations were standardized, and adjectives preceded by adverbs of negation were annotated as such. The top 186 most frequent adjective types were retained for analysis – these accounted for 92% (277,084) of all adjective tokens extracted from the dataset. This data was tabulated to produce a matrix containing ones and zeros to represent the presence or absence of an adjective type in comments about a given instructor. Simple principal components analysis with varimax rotation was carried out on this large binary matrix treating adjective types as variables and individual instructors as observations.

3 Results

Based on a scree plot of the eigenvalues, the first seven principal components (PCs) accounting for 16% of total variance were extracted. Of the 186 adjectives included in the PCA, 93 adjectives with component loadings of 0.2 and higher were retained. In each of the seven components, adjectives belong to a distinct semantic set (e.g. helpful, willing, approachable) clustered intuitively alongside adjectives of general evaluation (e.g. great, awesome, amazing). In each of the components, loadings on the adjectives were predominantly unipolar (i.e. almost all adjectives were positively loaded), reflecting co-occurrence tendencies within comments about the same instructor. The seven components are interpreted as latent dimensions along which students commonly perceive their instructors, and were named accordingly.

The ***HELPFULNESS*** dimension (PC1) brings together positive perceptions of helpfulness and approachability (e.g. *helpful*, *willing*, *caring*). The ***FUNNINESS*** dimension (PC2) indexes positive perceptions of dynamism and humour (e.g. *funny*, *hilarious*, *entertaining*). ***INTELLIGENCE*** (PC4) indexes positive perceptions of engagement and subject knowledge (e.g. *brilliant*, *intelligent*, *interesting*). ***RUDENESS*** (PC3) brings together negative perceptions of helpfulness and approachability (e.g. *rude*, *condescending*, *arrogant*). ***INCOMPETENCE*** (PC5) indexes negative perceptions relating to organizational skills, competence and clarity (e.g. *disorganized*, *confusing*, *not_clear*). ***TOUGHNESS*** (PC6) references perceptions of difficulty (e.g. *tough*, *difficult*, *not_easy*), and the dimension referred to as ***HOTNESS*** (PC7) indexes positive perceptions of instructors' appearance (e.g. *hot*, *gorgeous*, *sexy*).

Secondary analyses were carried out to explore relationships between these dimensions and the numerical scores. Adjectives in *HELPFULNESS*, *FUNNINESS*, *INTELLIGENCE* and *HOTNESS* were generally associated with high mean scores for helpfulness and clarity (and also for easiness, but to a lesser extent). The dimension of *TOUGHNESS* was associated with high scores on helpfulness and clarity, but low scores on easiness. *RUDENESS* and *INCOMPETENCE* were associated with low numerical scores across all categories. Analyses also examined the relationship between the seven dimensions and the Five Factor model of personality traits reported in Saucier and Goldberg (1995). The so-called 'Big Five' dimensions of personality are derived from factor analyses of self- and peer-ratings across

extensive inventories of personality adjectives. Based on component loadings on adjectives (N=55), dimensions in these two datasets correlate intuitively. For example, adjective loadings on *Agreeableness* (Big Five dimension II) correlate positively with those on the *HELPFULNESS* dimension, but negatively with those on *RUDENESS*. *Conscientiousness* (III) correlates positively *TOUGHNESS* with and negatively with *INCOMPETENCE*, while *Extraversion* (I) correlates positively with *FUNNINESS* and *Intellect* (V) positively with *INTELLIGENCE*.

4 Discussion

Comments posted on RMP convey predominantly positive perceptions of their instructors – e.g. adjectives in *HELPFULNESS* and *FUNNINESS* account for 47% and 32.3% of all adjectives extracted from the corpus. Contrary to claims that positive evaluations on RMP may merely reward 'easiness' (e.g. Felton et al. 2004), the dimension indexing *TOUGHNESS* indicate that, for certain instructors and certain students, 'difficulty' (i.e. the opposite) is actually associated with perceptions of good quality. The presence of a dimension indexing appearance (*HOTNESS*) suggests that instructor appearance may induce cognitive bias in the ratings – a so-called 'halo effect'. This dimension, which accounts for only 3.1% of adjective tokens, is probably, in part, elicited by the option on RMP to award a chili mark for '*hotness*'. Strong correlations between the seven dimensions and the 'Big Five' indicate that student evaluations of teaching are influenced by personality, as well as suggesting that the corpus derived dimensions have some psychological validity. These findings are discussed in relation to research on SETs and debates surrounding the validity of RMP.

Methodological implications are also discussed. Despite the prolific nature of SET research, very little attention has been given to students' open-ended text comments. This is, in part, due to the fact summarizing information contained in text is less straightforward than discrete item surveys. PCA (and comparable approaches) represent a useful methodology for automatically extracting meaningful lexical themes from a large number of texts. It is argued that in certain contexts (e.g. corpus-based discourse analysis) this approach may provide a more objective and data-driven alternative to keyword analysis.

References

Bleske-Rechek, A., and Fritsch, A. 2011. Student Consensus on RateMyProfessors. com. *Practical Assessment, Research and Evaluation*, 16(18).

Chung, C. K., and Pennebaker, J. W. 2008. Revealing Dimensions of Thinking in Open-Ended Self-Descriptions: An Automated Meaning Extraction Method for Natural Language. *Journal of research in personality*, 42(1): 96–132.

Felton, J., Mitchell, J. and Stinson, M. 2004. Web-based student evaluation of professors: the relations between perceived quality, easiness and sexiness. *Assessment and Evaluation in Higher Education*, 29(1): 91–108.

Foltz, P. W. 1996. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28: 197–202.

Forbes 2011. Compiling the Forbes/CCAP Rankings. Center for College Affordability and Productivity. *Retrieved from: http://centerforcollegeaffordability.org/uploads/2011_Methodology.pdf* (December 2012).

Sanders, S., Walia, B., Potter, J., and Linna, K. W. 2011. Do more online instructional ratings lead to better prediction of instructor quality? *Practical Assessment, Research and Evaluation*, 16(2): 1-6.

Saucier, G., and Goldberg, L. 1996. Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality*, 10(1): 61-77.

RMP, 2012 About RateMyProfessors.com. *Retrieved from: http://www.ratemyprofessors.com/About*. (December 2012)

Hierarchical cluster analysis of nonlinear linguistic data

Hermann Moisl
Newcastle University
hermann.moisl@ncl.ac.uk

1 Introduction

Cluster analysis (Gan et al. 2007) is a family of mathematical methods for the discovery of structure in data by identification and graphical display of proximity relations among data objects. As digital language corpora become increasingly important in linguistics, its application to data derived from such corpora becomes ever more relevant to linguistic research.

Hierarchical analysis is a widely used clustering method. Textbook accounts and current implementations of it all assume that the relationships among the variables describing data objects are linear, and cluster those objects on the basis of linear measurement of proximity among them. It has, however, become increasingly clear that nonlinearity pervades natural processes (Bertuglia and Vaio 2005). This nonlinearity can manifest itself in data describing these processes, and hierarchical analysis of nonlinear data based on linear measurement can give results that are inaccurate in proportion to the nonlinearity. Linguistic communication among humans is generated by a natural process known to be highly nonlinear, the brain (Stam 2006). Any analysis of data derived from language use must therefore consider the possibility that nonlinearity will be present.

This discussion aims to extend the applicability of hierarchical cluster analysis to nonlinear linguistic data.

2 Linearity and nonlinearity

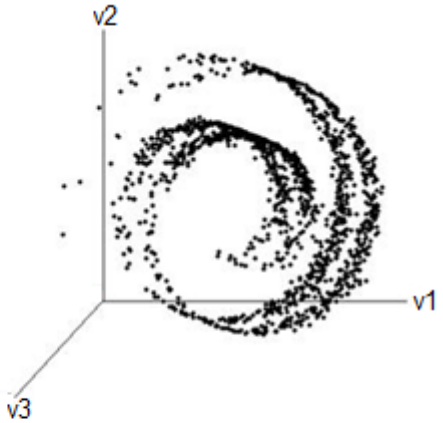
Linear natural processes have a constant proportionality between cause and effect. If a ball is kicked *x* hard and it goes *y* distance, then a 2*x* kick will appear to make it go 2*y*, a 3*x* kick 3*y*, and so on. Nonlinearity is the breakdown of such proportionality. In practice, the linear relationship increasingly breaks down as the ball is kicked harder and harder. Air and rolling resistance become significant factors, so that for, say, 5*x* it only goes 4.9 *y*, for 6*x* 5.7*y*, and again so on. Such nonlinear effects pervade the natural world and gives rise to a variety of complex behaviours.

Data is a description of objects in terms of a set of variables such that each variable is assigned a value for each of the objects. Given *m* objects described by *n* variables, a standard representation of data for computational analysis is a matrix M in which each of the *m* rows represents a different object, each of the *n* columns represents a different variable, and the value at $M_{i,j}$ describes object *i* in terms of variable *j*. The matrix thereby makes the link between the researcher's conceptualization of the research domain in terms of the semantics of the variables s/he has chosen and the state of the world, and allows the resulting data to be taken as a representation of the domain based on empirical observation.

.

	V1	V2	V3
1	-5.214	7.087	6.729
2	-0.421	8.426	7.896
3	-6.134	5.686	6.088
...	...	...	...
1600	-0.945	11.501	-10.867

a: 1600 3-dimensional vectors



b: The vectors in (a) plotted in 3-dimensional space

Figure 1

Matrices have a geometrical interpretation. For each row vector of M:

- the dimensionality of the vector, that is, the number of its components *n*, defines an *n*-dimensional Euclidean space.
- the sequence of *n* numbers comprising the vector specifies the coordinates of the vector in the space.
- the vector itself is a point at the specified coordinates

The set of row vectors in M define a configuration of points in the *n*-dimensional space called the data manifold. Linear manifolds are shapes consisting of straight lines and flat planes and represent linear data, whereas nonlinear manifolds consist of curved lines and surfaces and represent nonlinear data. Figure 1 gives an example of a nonlinear manifold in three-dimensional spaceCluster analysis represents the interrelationships among objects by grouping them on the basis of relative proximity in the data space. If the data is known to be linear, then a linear metric is appropriate. If it is known to contain nonlinearity, however, a linear or nonlinear metric may be appropriate, depending on the nature of the application. To a geophysicist, for example, the distance between two points on the earth's surface might be appropriately measured linearly, but to a geographer the linear measure would seriously underestimate the true surface distance, as in figure 2.

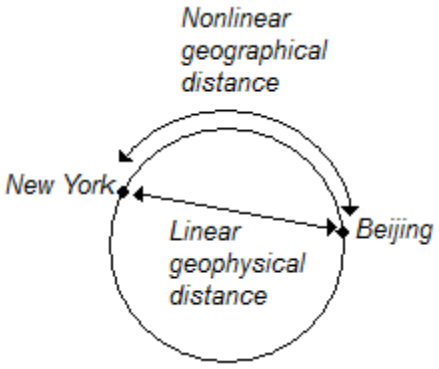


Figure 2

Curvature in a manifold represents the nonlinear aspect of the interrelationship among variables. Linear metrics ignore nonlinearity, making that aspect of the domain inaccessible to the clustering algorithm.

If nonlinearity in the data reflects an aspect of the domain which is salient to the research application, a linear metric misrepresents the relative proximity of data objects and can therefore be expected to generate distorted results; the solution is to base clustering on a nonlinear metric.

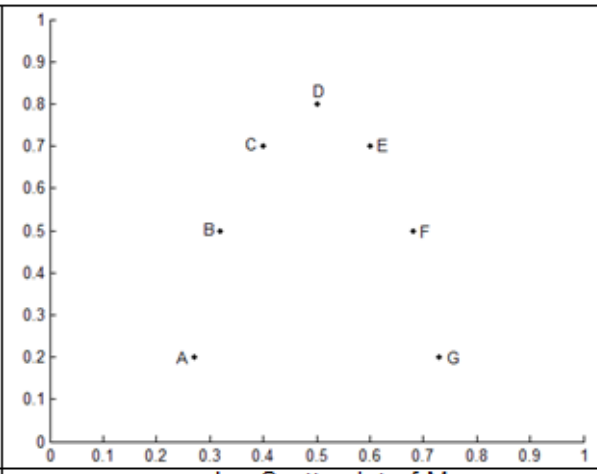
3 A nonlinear proximity metric

Hierarchical methods do not care how the values in the proximity matrices they use to construct cluster trees were derived, and there is consequently no obstacle to using values generated by a nonlinear metric. The metric proposed here is based on approximation of geodesic distance between points in a manifold using graph distance.

Figure 3 shows a small nonlinear matrix M and a scatterplot showing the corresponding manifold shape.

	V1	V2
A	0.27	0.20
B	0.32	0.50
C	0.40	0.70
D	0.50	0.80
E	0.60	0.70
F	0.68	0.50
G	0.73	0.20

a. Data matrix M



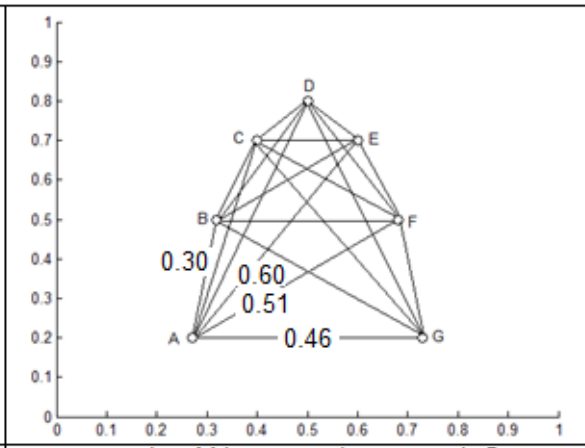
b. Scatterplot of M

Figure 3: Nonlinear data matrix and corresponding scatterplot

Given a maxtrix M with *m* rows and *n* columns, a Euclidean distance matrix D is an *m* x *m* matrix each of whose values $D_{i,j}$ is the Euclidean distance from row vector *i* to row vector *j* of M in *n*-dimensional space. Figure 4a shows D for M together with a graphical representation of the distances.

	A	B	C	D	E	F	G
A	0	0.30	0.52	0.64	0.60	0.51	0.46
B	0.30	0	0.22	0.35	0.34	0.36	0.51
C	0.52	0.22	0	0.14	0.20	0.34	0.60
D	0.64	0.35	0.14	0	0.14	0.35	0.64
E	0.60	0.34	0.20	0.14	0	0.22	0.52
F	0.51	0.36	0.34	0.35	0.22	0	0.30
G	0.46	0.51	0.60	0.64	0.52	0.30	0

a. Euclidean distance matrix D for M



b. M interpreted as a graph G

Figure 4

M is interpretable as a connected graph G each of whose arcs from *i* to *j* is labelled with the Euclidean distance between G_i and G_j, as shown in figure 4b.

A spanning tree for G is an acyclic subgraph of G which contains all the nodes in G and some subset of the arcs of G (Gross and Yellen 2006). A *minimum* spanning tree of G is a spanning tree which contains the minimum number of arcs required to connect all the nodes in G, or, if the arcs have weights, the smallest sum of weights. The minimum spanning tree for G in figure 4b is shown in figure 5, with the arcs comprising the tree emboldened.

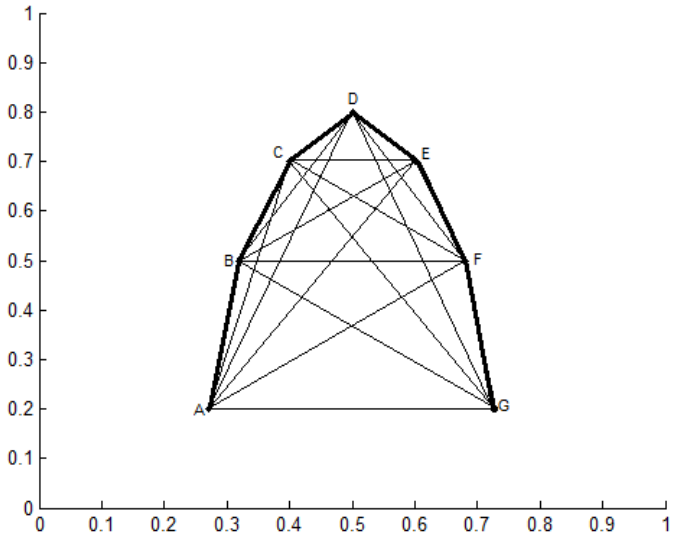


Figure 5

A minimum spanning tree can be used to approximate the geodesic distances between all *m* row vectors of M in *n*-dimensional space using the Euclidean distances because the distance between any two nodes is guaranteed to be minimal. By summing the shortest paths between nodes, a table of graph distances between all data vectors can be constructed: the Euclidean and graph distances between A and B in figure 10 are identical, but from A to C the graph distance is AB + BC rather than the Euclidean AC, and so on. The graph distance matrix and the Euclidean matrix from which it was derived are shown in figure 6.

	A	B	C	D	E	F	G
A	0	0.30	0.52	0.64	0.60	0.51	0.46
B	0.30	0	0.22	0.35	0.34	0.36	0.51
C	0.52	0.22	0	0.14	0.20	0.34	0.60
D	0.64	0.35	0.14	0	0.14	0.35	0.64
E	0.60	0.34	0.20	0.14	0	0.22	0.52
F	0.51	0.36	0.34	0.35	0.22	0	0.30
G	0.46	0.51	0.60	0.64	0.52	0.30	0

a. Euclidean distance matrix for M.

- Sum of distances: 16.27
- Mean distance: 0.39
- Distance A-G: 0.46

	A	B	C	D	E	F	G
A	0	0.30	0.53	0.66	0.80	1.00	1.32
B	0.30	0	0.22	0.35	0.49	0.70	1.00
C	0.53	0.22	0	0.14	0.28	0.49	0.79
D	0.66	0.35	0.14	0	0.14	0.35	0.66
E	0.80	0.49	0.28	0.14	0	0.22	0.52
F	1.01	0.70	0.49	0.35	0.22	0	0.30
G	1.32	1.01	0.80	0.66	0.52	0.30	0

b. Graph distance matrix for M.

- Sum of distances: 22.53
- Mean distance: 0.54
- Distance A-G: 1.32

Figure6

The sum of distances and the mean for the graph matrix are greater than for the Euclidean one, and the graph distance between A and G is almost three times larger than for the Euclidean, which figure 5 confirms visually.

Whereas the linear distance takes no account of manifold shape, the graph approximation of geodesic distance is constrained to follow the shape of the manifold by the need to visit its nodes in the course of tree traversal. This corresponds to approximating the geodesic distance between any two cities on the Earth's surface, say from New York to Beijing in figure 2, by stopping off at intervening airports, say New York -> London -> Istanbul -> Dehli -> Beijing.

4 Case study

The application of a nonlinear metric for hierarchical analysis of corpus-derived linguistic data is exemplified with respect to the phonetic transcriptions for 63 speakers in the *Diachronic Electronic corpus of Tyneside English* (DECTE) . Each DECTE speaker is represented by a 156-element vector which constitutes a description of phonetic usage: each element represents a different phonetic segment in the transcription scheme, and the value at any given element is the frequency with which the speaker uses the associated segment.

Vector index	1	2	3	...	156
Phonetic segment symbol	g	i	t	...	ɜ:
1. nectetlsg01	31	28	123	...	0
2. nectetlsg02	22	8	124	...	0
⋮	⋮	⋮	⋮	⋮	⋮
63. nectetlsn07	19	3	73	...	0

Table 1: Fragment of the NECTE data matrix M

The 63 x 63 Euclidean distance matrix D_{euc} was calculated for M, the minimum spanning tree for D_{euc} was found, and the geodesic distance matrix D_{geo} was derived by tree traversal. The rows of both D_{euc} and D_{geo} were then linearized into vectors of length 63 x 63 = 3969, sorted, and co-plotted to get a graphical representation of the relationship between linear and geodesic distances in the two matrices, as in figure 7.

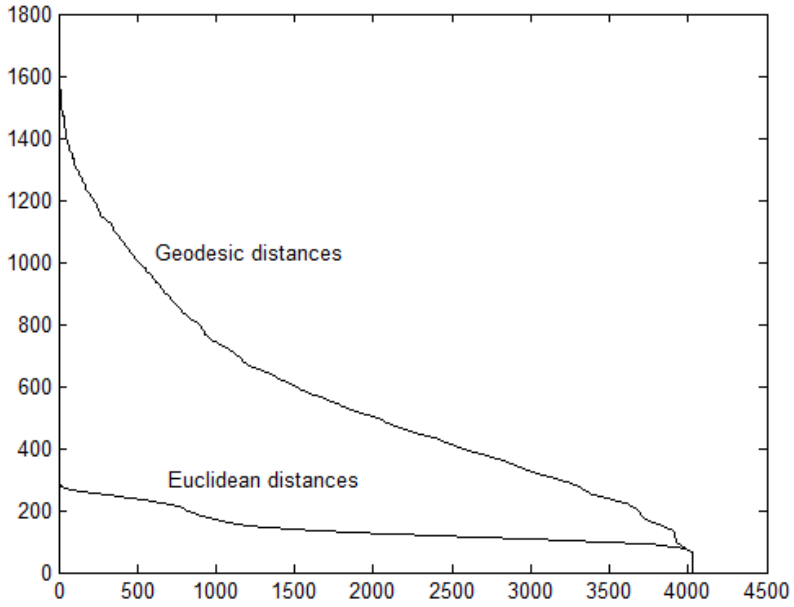


Figure 7

The geodesic distances among the speakers in M are consistently larger than the Euclidean ones; the ratio *mean(D_{geo}) / mean(D_{Euc})* of mean distances, which is 3.89 shows that M contains substantial nonlinearity. D_{euc} and D_{geo} were hierarchically cluster analyzed, with results in figure 8.

Interpretation of (A) – (E) is based on social data associated with the 63 speakers. The two analyses of the nonlinear data manifold M, one based on linear distance measurement and the other based on geodesic, support substantially different sociolinguistic hypotheses about the relationship between phonetic usage and social factors in the Tyneside speech community. As the foregoing discussion has argued, the analysis based in geodesic distance is to be preferred in principle.

References

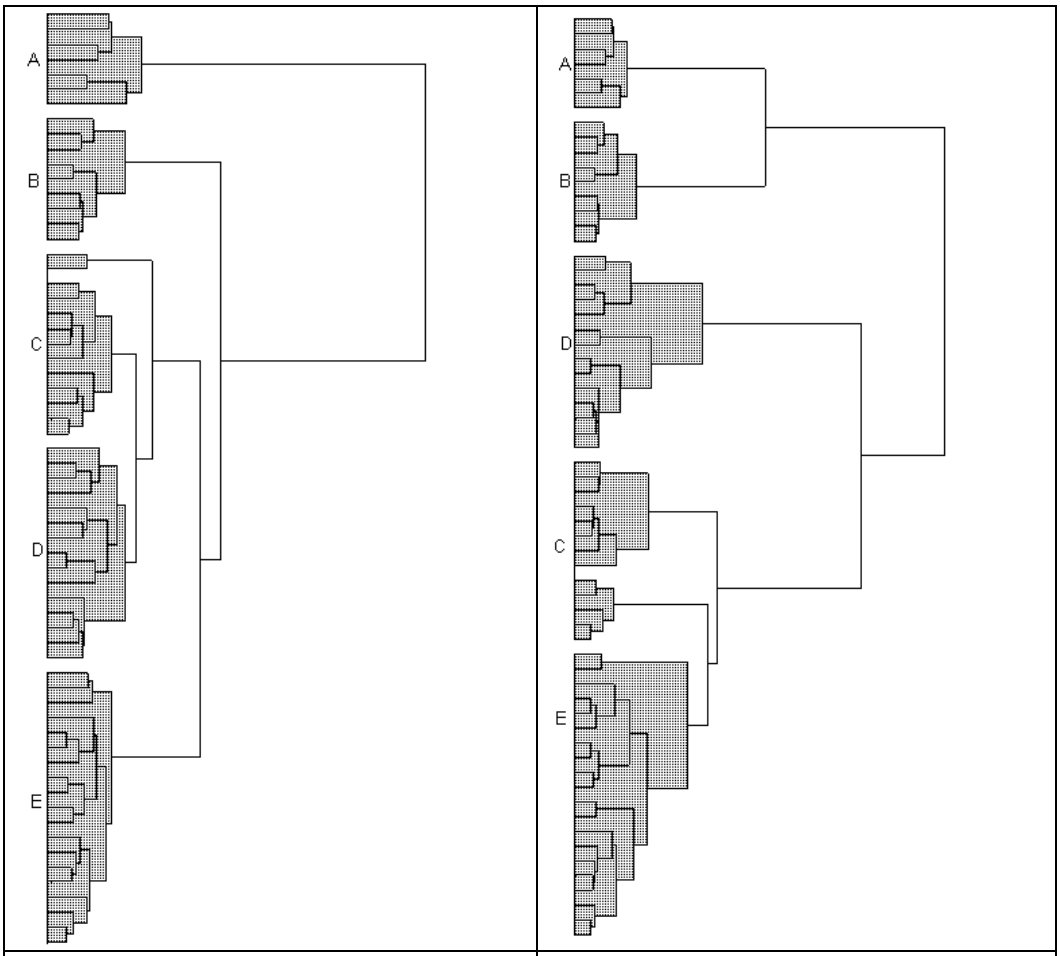
Bertuglia, C. and Vaio, F. 2005. *Nonlinearity, Chaos, and Complexity: The Dynamics of Natural and Social Systems.* Oxford: Oxford University Press.

DECTE: The Diachronic Corpus of Tyneside English: http://research.ncl.ac.uk/decte/

Gan, G., Ma, C., Wu, J. 2007. *Data Clustering. Theory, Algorithms, and Applications,* American Statistical Association.

Gross, J. and Yellen, J. 2006. *Graph Theory and its Applications, 2nd ed.* New York: Chapman & Hall.

Stam, C. 2006. *Nonlinear Brain Dynamics.* Hauppauge NY: Nova Science Publishers.



a. Ward's Method cluster tree for M_{EUC} | b. Ward's Method cluster tree for M_{GEO}

Figure 8

An affix-based method for automatic term recognition from a medical corpus of Spanish

Antonio Moreno-Sandoval
Autonomous University of Madrid
antonio.msandoval@uam.es

Leonardo Campillos-Llanos
Autonomous University of Madrid
leonardo.campillos@uam.es

Alicia González-Martínez
Autonomous University of Madrid
a.gonzalez@uam.es

José M. Guirao-Miras
University of Granada
jmguirao@ugr.es

1 Introduction

Automatic Term Recognition (ATR) aims to develop software to identify a list of candidate words in a text that are likely to be technical terms. After this process, specialists have to assess the results in order to validate the final list of terms.

Current literature on ATR describes basically three approaches: statistical, linguistic and hybrid. First, statistical techniques rely on measuring how distinctive is a word or lemma in a specialized context when compared with a general corpus. This approach is well represented by the log-likelihood statistic (Dunning 1993) included in Wordsmith Tools (Scott 2008), or the logDice metric used in The Sketch Engine (Kilgarriff et al. 2004).[1]

Secondly, linguistic approaches focus on using language resources such as dictionaries, lexicons, and ontologies. For further references of each approach, we refer to Ananiadou and Nenadic (2006).

This paper presents a hybrid approach with two stages. Firstly, automatic methods are applied to construct the list of candidates. Secondly, a list of affixes is used to select proper medical terms, in order to reduce the human assessment.

We conducted an experiment on comparing three different automatic methods to obtain lists of candidates. For that goal, we evaluated the accuracy of the medical terms selected by each approach.

2 Overall description of the method

Our ATR method consists of two phases (see Figure 1):

1. Extracting lists of candidate terms.
2. Matching candidate terms against the affix list.

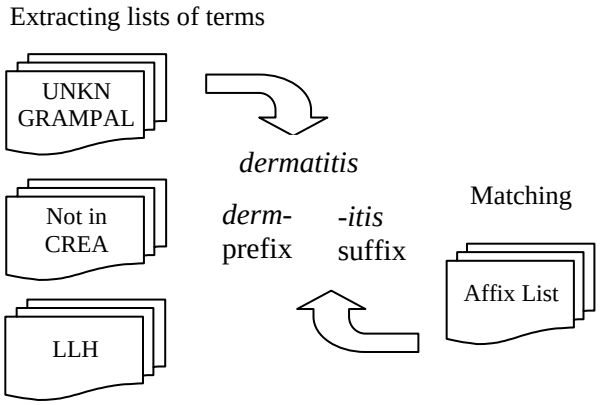


Figure 1. Flow chart of the experiment.

3 The Spanish MultiMedica corpus

The MultiMedica corpus has been compiled for the homonymous project[2] by the Computational Linguistics Laboratory at the Autonomous University of Madrid (LLI-UAM). This is a comparable corpus of Spanish, Arabic and Japanese texts about health topics. One of the foreseen applications is an ATR tool aimed at translators and terminologists in the health domain. The Spanish corpus consists of three resources:

- *Harrison*:[3] it includes professional and scientific texts written by medical doctors, and gathers over 3800 documents
- *OCU-Salud*:[4] it is a collection of journalistic texts written by medical doctors, but edited and reviewed by journalists.
- Website *Tu otro medico*:[5] it assembles encyclopedic articles written by professional doctors for non-specialists.

The corpus was filtered by hand in order to avoid information redundancy. In total, it covers 4200 texts and over 4 million words, and reflects a balanced combination for most medical specialties.

4 Extraction of the three lists

To obtain the lists of candidate terms, we tested three approaches: a tagger-based technique (by

[1] www.sketchengine.co.uk/ [24/05/2013]

[2] http://labda.inf.uc3m.es/multimedica/
[3] http://www.harrisonmedicina.com/
[4] http://www.ocu.org/ocu-salud/
[5] http://www.tuotromedico.com/

using GRAMPAL tagger, Moreno and Guirao 2006), a corpus-based technique (with a frequency list of word forms in the CREA Corpus of Contemporary Spanish) [1] , and a statistical technique (the Log-likelihood statistic, henceforth LLH):

- The "Unknown-by-GRAMPAL" list contains the words that were not recognized by the GRAMPAL lexicon, which contains over 50000 lemmas. This lists includes 22413 tokens.
- The "Not-in-CREA" list is formed by all the words in MultiMedica corpus that are not found in the CREA corpus. This list has 23239 tokens.
- The LLH list was extracted by comparison against the CREA. Only the words ranked over 10 in the LLH statistic are considered. This list assembles 8667 word forms.

Almost 50% of the items in the CREA list are *noisy* words: 350000 out of over 700000 tokens are non-Spanish words (mostly foreign words, but also proper names and misspellings). On the other hand, GRAMPAL can analyze over 500000 correct word forms.

5 The affix list

Linguistic and rule-based methods have already been applied to ATR tasks in the immunological (Ananiadou 1994) and the pharmacological domain (Segura et al. 2008). In our experiment, we used a list of 2168 affixes (considering spelling variations: e.g. *aden-*, *adeno-*). The list includes these data:

- Greek and Latin affixes from the health domain (e.g. *cardio-*, *-itis*); this list has been enriched with very frequent roots (e.g. *pancrea-*) taken from studies on medical terminology (López Piñero and Terrada 2005; Jiménez Arias 2012; Sánchez González 2012). We did not include very general affixes that are not always related to the medical domain (e.g. *pre-*).
- Stems for the recognition of pharmacological substances (e.g. *-cavir*), which were compiled from lists approved by the American Medical Association (AMA) for the nomenclature of clinical compounds,[2] and stems proposed by the World Health Organization (WHOa, WHOb 2011). We collected also affixes that refer to biochemical entities (e.g. *-sterol*).[3] Since most English affixes have a univocal correspondence with the Spanish terms, few needed an adaptation (especially, those ending with vowel; e.g. *–ine* > *-ina*, as in *creatine* > *creatina*).

6 The matching procedure

Prior to the matching procedure, we pre-processed the affix list to generate possible variants of each affix. This stage consisted of three tasks:

- In cases where the affix has an orthographic (or acute) accent, a variant without it is used. The affix may bear accent or not depending on the place of the stress within the word (e.g. in the term *próstata* the prefix is *prós*t-, whereas in *prostático* the prefix is *prost-*).
- When the affix has an epenthetic vowel, two versions of it are used (e.g. from *(e)scoli-* we have *escoli-* and *scoli-*).
- It is important to note that the candidates of the list are non-lemmatized forms. Thus, all inflected variants are generated for suffixes ending in *-o* (e.g. from suffix *-génico* we generate *-génicos*, *-génica*, and *-génicas*).

Subsequently, the affix list is classified in a sublist of prefixes and a sublist of suffixes. Both lists are sorted by length, so the larger affixes are placed at the beginning of the list.

Secondly, we carried out the matching procedure itself. We took each candidate from the three lists and compared it with each affix from the lists. If a substring from the beginning matches one prefix, we stopped the search and marked the prefix. We took the remainder of the candidate string and searched more prefixes until there are no matches. Then, we proceeded in the same manner with the suffixes. In the end, we had a list of prefixes and a list of suffixes for each candidate.

Finally, candidate terms with no affixes were classified as rejected, and items with at least one affix were classified as accepted. From this, we created a list of rejected and a list of accepted candidate terms for the three methods (Table 1).

Results in our corpus show that the list obtained by means of the GRAMPAL tagger has the highest recall, followed by the list obtained by comparing the MultiMedica corpus wordlist with the CREA.

[1] The CREA Corpus contains over 150 million tokens and over 700000 different word forms. Frequency lists are available at: http://corpus.rae.es/lfrecuencias.html

[2] http://www.ama-assn.org/resources/doc/usan/stem-list-cumulative.pdf, http://www.ama-assn.org/resources/doc/usan/new-stem-list.pdf [accessed: 30/12/2012]

[3] We looked up the list of affixes collected by Michael Quinion (2008): http://www.affixes.org [accessed: 02/01/2012]

	Accepted		Rejected	
	Nº	%	Nº	%
Unknown for GRAMPAL	14551	64.92	7862	35.08
Not in CREA	12307	52.96	10932	47.04
LLH	3832	44.21	4835	55.79

Table 1. Accepted and rejected candidate terms.

7 Evaluation

The lists of accepted and rejected terms by each method were manually evaluated in order to confirm the results. We accepted a candidate term if that item was registered in prestigious research articles and medical books[1] or in an authorized medical dictionary (e.g. *Dorland* 2005; or *Diccionario de términos médicos,* Real Academia Nacional de Medicina 2011). Units of measure (e.g. *kg*) or foreign words were not accepted as terms, excepting those borrowings that health professionals tend to use (e.g. *heparin* was rejected, since it is adapted to *heparina* in Spanish; yet *stent* or *bypass* were accepted as terms). After having looked up the candidates in each list, we prepared contingency tables for each method (Tables 2-4).

Not in GRAMPAL	Accepted	Rejected
True	88.56%	48.13%
False	11.44%	51.87%

Not in CREA	Accepted	Rejected
True	83.12%	34.80%
False	16.88%	65.20%

Log-likelihood	Accepted	Rejected
True	91.99%	69.74%
False	8.01%	30.26%

Tables 2-4. Contingency tables.

In the final step of the experiment, we collected a list of all candidate terms by gathering the results from the true accepted and the false rejected terms for each list, and deleting any repeated item. The following table shows the estimation of precision, recall, and F values for each method.

	GRAMPAL	CREA	LLH
Precision	88.56%	83.12%	91.99%
Recall	51.57%	41.44%	14.28%
F measure	65.18%	55.31%	24.72%

Table 5.

[1] We looked up each item on the Google Books corpus online.

8 Discussion

In our experiment, the LLH statistical method stands out slightly above in terms of precision (91.99%), but not in terms of recall, due to the high rate of rejected items that are true terms (69.74%). The highest precision rate in this list may be explained by several reasons. First, the "Not-in-CREA" list is rather noisy for it includes many not recognized verb inflected forms. Secondly, the CREA corpus contains texts from the scientific and medical domain. When comparing the list of words from the MultiMedica corpus against the CREA wordlist, many terms included in the scientific texts from the CREA are not proposed as candidate terms – although they *are* medical terms. Similarly, the list obtained with GRAMPAL does not contain several medical terms which are included in the lexicon of the tagger. For example, GRAMPAL recognizes *vacuna* ('vaccine') as a valid word. Thus, it is not proposed as a candidate term, despite being a medical term.

9 Conclusions and future work

We have performed and evaluated three methods for ATR in the biomedical domain: a tagger-based, a corpus-based, and a statistical approach. The experiment in our corpus has shown that the LLH method achieved a high precision in retrieving candidate terms, at the expense of a low recall. For Spanish, which is an inflecting (or fusional) language, the combination of a tagger-based method with a comprehensive list of affixes provided better results.

Acknowledgments

This research has been supported by the Spanish Government (under the grant TIN2010-20644-C03-03) and by the Madrid Regional Government (under the grant MA2VICMR).

References

Ananiadou, S. (1994) "A methodology for Automatic Term Recognition". *COLING'94 – Proc. of the 15th Int. Conf. on Computational Linguistics:* 1034-1038.

Ananiadou, S. and Nenadic, G. (2006) "Automatic Terminology Management in Biomedicine". In S. Ananiadou and J. McNaught (Eds.) (2006) *Text Mining for Biology and Biomedicine,* pp. 67-97. Boston/London: Artech House.

Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics,* 19(1), 61-74.

Jiménez Arias, Mª. E. (2012) "Afijos grecolatinos y de otra procedencia en términos médicos". *MEDISAN,*

16(6): 1005-1021 [Accessed: 30/12/2012].

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) "The Sketch Engine". In *Proceedings of EURALEX 2004, Lorient, France,* 105-116. http://www.sketchengine.co.uk [Accessed: 02/01/2013]

López Piñero, J. Mª., and Terrada Ferrandis, Mª. L. (2005) *Introducción a la terminología médica*. Barcelona: Masson.

Moreno Sandoval, A., and Guirao, J. M. (2006) "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation". In Y. Kawaguchi, Zaima, S., and Takagaki, T. (eds.) *Spoken Language Corpus and Linguistic Informatics*, pp. 199-218. Amsterdam/Philad.: John Benjamins.

Real Academia Española (2013) Banco de datos (CREA) [online]. Corpus de referencia del español actual. http://corpus.rae.es/creanet.html [Accessed: 02/01/2013]

Dorland (2005) *Diccionario enciclopédico ilustrado de medicina Dorland* (Spanish edition of *Dorland's Illustrated Medical Dictionary*). 30th edition. Madrid: Elsevier, D. L.

Real Academia Nacional de Medicina (2011) *Diccionario de términos médicos*. Madrid: Médica Panamericana

Sánchez González, M.A. (2012) *Historia de la medicina y humanidades médicas*. 2nd edition. Barcelona: Elsevier/Masson.

Scott, M. (2008) WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

Segura Bedmar, I., Martínez, P., and Samy, D. (2008) "Detección de fármacos genéricos en textos biomédicos". *Procesamiento del Lenguaje Natural*, 40, 27-34.

WHO (2011) "The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances" http://apps.who.int/medicinedocs/documents/s19117en/s19117en.pdf [Accessed: 2/01/2013]

WHO (2011) "International Nonproprietary Names (INN) for biological and biotechnological substances (a review)" http://apps.who.int/medicinedocs/documents/s19119en/s19119en.pdf [Accessed: 2/01/2013]

Longitudinal development of L2 English grammatical morphemes: A clustering approach

Akira Murakami
University of Cambridge
a.murakami39@gmail.com

1 Aim of the study

Based on a longitudinal learner corpus, this study aims at disclosing the developmental patterns of grammatical morphemes in English as the second language (L2). More specifically, the following research questions are addressed;

1. Does the longitudinal transition of accuracy within individual learners show systematic patterns such as linear increase or U-shaped development, or does the accuracy randomly fluctuate?
2. Are the patterns different depending on morphemes, learners' native language (L1), and their proficiency?

2 Corpus

The EF-Cambridge Open English Learner Database (EFCamDat) is a learner corpus containing learners' essays written in Englishtown, an online school run by Education First. A typical English course at Englishtown has 16 Lessons with eight Units each. At the end of each Unit is a free composition in which learners are asked to write on a certain topic.

Learners in Englishtown receive feedback from native-speaker "teachers" on each essay. The feedback includes identification and correction of grammatical morphemes, among other things. The present study views the feedback as error annotation and utilizes it in the calculation of accuracy scores explained later.

The EFCamDat includes for each essay such metadata as the learner's country of residence, the topic of the essay, the date and time of submission, and the Lesson and the Unit number at which the essay was written.

3 Target morpheme, L1 groups, and proficiency levels

The study targeted the following six grammatical morphemes; articles, past tense *–ed*, plural *–s*, possessive *'s*, progressive *–ing*, and third person *–s*.

Ten L1 groups were targeted; L1 Portuguese,

L1 Chinese, L1 German, L1 French, L1 Italian, L1 Japanese, L1 Korean, L1 Russian, L1 Spanish, and L1 Turkish. They represent typologically diverse languages and are suited for the investigation of L1 influence.

The Lesson and Unit number the learner wrote the essay at is considered to represent the proficiency level of the learner at the time of writing the essay. According to Englishtown, first three levels are considered as beginner level corresponding to A1 in the Common European Framework of Reference (CEFR), four to six as elementary (A2), seven to nine as intermediate (B1), 10 to 12 as upper intermediate (B2), 13 to 15 as advanced (C1), and 16 as upper advanced (C2).

Approximately 140,000 essays consisting of 10 million words comprised the subcorpus used in the present study.

4 Scoring and data retrieval

The present study analyses accuracy development. As a measure of accuracy, it employs target-like use (TLU) score calculate by the following formulae (Pica 1983);

$$\frac{\textit{number (\#) of correct suppliance}}{\textit{\# of obligatory contexts} + \textit{\# of overgeneralization errors}}$$

Obligatory contexts in the present study refer to the occasions where learners are required to use target morphemes, and overgeneralization errors are the instances where learners incorrectly supplied target morphemes. The number of correct suppliance refers to the number of cases where learners correctly used the morphemes.

In order to obtain TLU scores, the number of obligatory contexts, the instances of overgeneralization, and those of errors were extracted from error-tagged texts. For the identification of obligatory contexts, error-tagged texts were first converted into corrected texts where the corrections of errors were reflected onto the texts. The number of instances of target morphemes in a corrected text was assumed to be the number of obligatory contexts in the corresponding original essay. For the identification of errors, I developed a script that looks at both error tags and the part-of-speech tags in the original and the corrected texts. For example, in the case of an *-s* omission error, the part of speech of the corrected word with *-s* was checked in the corrected text, and if the word was tagged as a verb, the error was considered as an instance of third person *-s* error.

5 Data analysis

Because each essay was relatively short, calculation of TLU scores based on single essays would be unreliable. Therefore, TLU scores were computed over multiple essays written by a learner. For this purpose, the error-tagged essays were first chronologically ordered within each learner based on submission dates. Then TLU scores were calculated in a moving-window fashion over the essays that included at least 15 (for articles and plural *–s*) or 10 (for the other morphemes) obligatory contexts (OCs).

Once TLU scores were obtained, the development of the learners over 10 (for articles and plural *-s*) or 5 (for the other morphemes) windows was analyzed. The learners were then clustered according to their longitudinal developmental patterns of each grammatical morpheme over multiple windows, and the association between the patterns and the learners' L1 and proficiency was explored.

6 Results

KmL clustering (Genolini and Falissard 2010) was applied to the data in order to retrieve typical developmental patterns of morphemes. KmL stands for k-means for longitudinal data and lets similar developmental patterns cluster by themselves in a data-drive manner. KmL has a feature that automatically decides the optimal number of clusters based on Calinski and Harabasz criterion *C(g)* (Calinski and Harabasz 1974). The measure favours the number where there is a large between-cluster variance of and a small within-cluster variance of trajectories, as well as a larger number of trajectories in each cluster.

If KmL is run on the present data as they are, it takes into account the absolute accuracy of each learner and may cluster learners according to their proficiency. As the aim is not to cluster learners by proficiency but by the shape of their accuracy development, all the data points were learner-mean-centred, that is, the mean accuracy value of each learner was subtracted from all data points of the learner. This neutralizes the difference in absolute accuracy and clustering will be based purely on developmental shape, or the deviation from the mean accuracy value of the individual.

Possessive *'s* had too few windows and clustering was not able to be performed on the data. It was removed from the subsequent analysis.

The *C(g)* measure suggested that the optimal number of clusters is two for all the rest of the morphemes. Because third person *–s* included

only one learner in a cluster, it was removed from the subsequent analysis as well.

Figure 1 illustrates the two clusters of developmental patterns of articles. The horizontal axis represents windows and the vertical axis represents TLU scores. Each gray line corresponds to the development of one learner, and thick black lines are locally weighted scatterplot smoothing (LOWESS; Singer and Willett 2003; Larson-Hall and Herrington 2010) lines showing the overall developmental patterns in each group. Different line types (i.e., solid, dashed, and dotted) correspond to the learners with different types of L1s (whether articles are present or absent, or their L1 is Chinese, which is located somewhere between the two groups [Huang 1999; Chen 2004]).

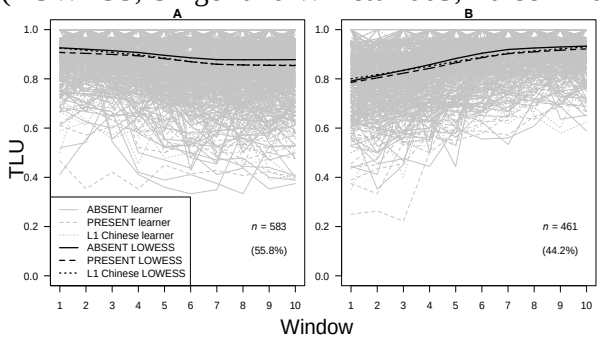


Figure 1: KmL Clustering of Articles

In all the morphemes including articles, the two clusters represented a cluster in which learners' accuracy development is rather flat (as in Cluster A in Figure 1) and that where their development is represented by increasing accuracy (as in Cluster B in Figure 1). The proportion of the learners in each cluster varied across morphemes, however. Besides the commonality, significant individual differences were observed in all the target morphemes. A series of χ^2 tests, G tests, and binary logistic regression analyses suggested that in determining clusters, (i) both L1 and proficiency are significant predictors in articles, (ii) L1 is a significant predictor in plural *–s*, and (iii) neither L1 nor proficiency is a significant predictor in past tense *–ed* and progressive *–ing*.

The association between the clusters was weak, suggesting that the development of multiple morphemes is independent of one another.

7 Discussion and conclusion

As to Research Question 1, some commonalities in the developmental patterns were observed in all the target morphemes. For example, majority of the learners were always classified into the cluster that shows flat development. In this sense, there are certain patterns of development that learners tend to follow. At the same time, however, there was large individual variability that was not explained by L1 or proficiency.

As to Research Question 2, morpheme affects the developmental patterns because the proportion of the learners who show flat development differs across morphemes. L1 can affect the developmental patterns depending on morphemes, but its effect is weak at best. Proficiency can also affect the pattern depending on morphemes, and, if it does, higher proficiency learners tend to show flatter development, possibly due to the ceiling effect.

Together with large individual variability observed, the study demonstrates that the development of morpheme accuracy is a complex process influenced by a variety of factors.

References

Calinski, T. and Harabasz, J. 1974. "A dendrite method for cluster analysis". *Communications in Statistics* 3 (1): 1-27.

Chen, P. 2004. "Identifiability and definiteness in Chinese". *Linguistics* 42 (6): 1129-1184

Genolini, C. and Falissard, B. 2010. "KmL: K-means for longitudinal data". *Computational Statistics* 25 (2): 317-328.

Huang, S. 1999. "The emergence of a grammatical category *definite article* in spoken Chinese". *Journal of Pragmatics* 31 (1): 77-94.

Larson-Hall, J. and Herrington, R. 2010. "Improving data analysis in second language acquisition by utilizing modern developments in applied statistics". *Applied Linguistics* 31 (3): 368-390.

Pica, T. 1983. "Methods of morpheme quantification: Their effect on the interpretation of second language data". *Studies in Second Language Acquisition* 6 (1): 69-78.

Singer, J. D. and Willett, J. B. 2003. *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Exploring intra-author variation across different modes of electronic communication using the FITT corpus

Millicent Murdoch
University of Leeds
ml09m4m@leeds.ac.uk

1 The FITT corpus

The FITT corpus built for this study is a mini, 'corpus of specialised genres' (Handford 2010) comprised of six participants' authentic samples of data, taken from across four modes of electronic communication: **F**acebook, **I**nstant messaging, **T**ext messaging, and **T**weets (FITT). The corpus contains 64,104 words in 2,049 messages. Participants (anonymised) are: Alexa, Geoffrey, Helen, Robert, Tom and Wayne.

In terms of corpus balance and representativeness, since the data represents natural instances of communication for the participants, different preferences for particular media are apparent and frequency of use is also varied across participants, creating an unbalance. However, the corpus is representative of young people's behaviour, as it provides a four month synchronic sample of how these participants use the different channels of communication.

Using a small, specialised corpus allows for a combination of a qualitative, stylistic genre analysis, as well as the more quantitative methods generally employed by corpus linguists, an approach used by, for example, Semino and Short (2004) and Flowerdew (2005).

2 Authorship attribution

Forensic linguists are increasingly concerned with the use of language in electronic formats, due in large part to the rise of criminal activity which takes place in an online capacity, for instance: the distribution of illegal materials such as child pornography, chat-room sexual predators, and online communication channels for terrorist organisations. Because of the ease with which one can gain anonymity in the online world, questions of authorship attribution in such contexts is a growing area of research, both in a forensic capacity, such as the work of Macleod and Grant (2011) on social network activity, and in research which explores online communication discourse, such as text messaging (Tagg 2009).

However, very little has been done to investigate the potential intra-author variations across the range of these different platforms of communication, with the exception of Turell (2010) who explores the concept of 'idiolectal style' between authors' emails and fax messages.

3 Research aims

The main aim of this research is therefore to explore a broad range of online communication channels in order to identify whether stylistic features remain consistent regardless of the means of communication, and to what extent authors adapt their linguistic style to suit the conventions of a particular communicative mode.

4 Corpus-driven methodology

A corpus-driven methodology forms the basis for the analysis of the authors' data. Tagg (2009) argues for a corpus-driven approach to the analysis of text messaging because of the innovative linguistic features it presents and the lack of existing lexical/grammatical frameworks to form the foundation of a wholly corpus-based approach. This argument can also be extended to the other forms of electronic communication under investigation in this research.

Although the work of Tagg (2009) and MacLeod and Grant (2011) has successfully demonstrated the use of corpus methods to provide insight into the linguistic features present for one particular type of electronic communication, no previous research has attempted to identify consistent authorship indicators across all four data types included in the FITT corpus.

5 Preliminary findings

An initial quantitative analysis of the FITT corpus using the Word List function of *Wordsmith Tools* (Scott, 2011) identifies five areas of interest: personal pronouns, direct terms of address, word form variants, semantic lexical groups and non-standard spelling.

Looking at the occurrence of first and second personal pronouns, 'I' and 'You', in the FITT corpus leads to two initial conclusions. Firstly, comparing the usage across the four different communicative platforms, it is apparent that both occur much more frequently in the conversational forms, IMs and Text Messages, than in the micro-blogging forms of Facebook and Twitter. Secondly, there is a great deal of inter-author and intra-author variation in the relative frequencies of 'I' and 'You' across the four modes, with only one author consistently using one more frequently than the other.

Direct terms of address account for 3.97% of words in the FITT corpus. These are made up of 32 variant forms, 17 of which are specific to one author. Of these, four terms, *honey, son, dude* and *G*, occur in the authors' top three most frequently used variants. The form *Forename + Surname* is found to be limited to performing the 'tagging' function in Facebook posts, whilst several pluralised forms, including *guys, girls, kids* and *people* appear only in one author's Tweets. This indicates that the conventions of these modes are having an influence on the linguistic choices the authors make.

An analysis of variants for *yes*, *no* and *ok* reveals quite a distinct level of both inter-author and intra-author variation. For each of these words, only the standard form is used by all of the authors, yet there are 17, 7 and 9 (respectively) variants, 19 of which are specific to one author. Of particular note are: one author's tendency to add prosodic stress to *no* variants, e.g. *noooo, nahhhh*, and the form *okidokey* (as a variant of *ok*) representing the most frequent variant for the author, who uses it exclusively.

Word lists of each of the authors' data reveal clear lexical semantic groups for three of the authors. Alexa's semantic group of 'family' accounts for 3.68% of all her corpus data and includes sub-fields of relations, home, birth, affection and baby. Geoffrey's vocabulary is extremely work oriented and includes sub-fields of IT, locations and business jargon to account for 2.35% of his word list. Finally, 4.31% of Wayne's word list is made up of 'music' related words, including sub-fields of musical genres, promotion, production, and associated artists. These semantic items appear across all four modes of communication, suggesting that these topics are significant to the authors. However, to gain a true insight into authorship features at the level of idiolect, the next stage will be to perform a topic-independent analysis of the corpus.

The final feature which has been looked at in this initial stage of analysis is non-standard spellings. A classification system for types of non-standard spelling has been devised, the top-most level of which consists of eight categories: shortening, lengthening, blends, omitted space, omitted apostrophe, onomatopoeia, accent representation and typos. There is a great deal of inter-author variation, with particular categories being typical for certain authors, whilst absent from the data of others. At the intra-author level, non-standard spellings appear to be a rather consistent feature, with the same categories being used by authors across all of their communicative modes.

6 Conclusion

This research has the potential to contribute not only theoretical knowledge about authorship attribution, but may also impact on the practical application of analytical techniques in real world cases. For instance, if a case of disputed authorship arises regarding an IM conversation, but the only available linguistic data from the suspect is in the form of publicly accessible social networking posts, this research will be able to shed light on the feasibility of using corpora to perform a cross-platform analysis in order to provide accurate and reliable results.

References

Flowerdew, L. 2005. 'An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies.' *English for Specific Purposes*. 24: 321-332.

Handford, M. 2010. What can a corpus tell us about specialist genres? *In* A. O'Keefe and M. McCarthy. (eds) *TheRoutledge Handbook of Corpus Linguistics*. New York: Routledge. pp. 255-269.

MacLeod, N. and Grant, T. 2011. 'Whose Tweet? Authorship analysis of micro-blogs and other short-form messages.' *Proceedings of the International Association of Forensic Linguistics 10th Biennial Conference*. 210-225.

Scott, M. 2011. *Wordsmith Tools.*

Semino, E. and Short, M. 2004. *Corpus Stylistics: Speech, Writing, and Thought Presentations in a Corpus of English Writing*. London: Routledge.

Tagg, C. 2009. *A corpus linguistics study of SMS text messaging*. PhD Thesis, University of Birmingham. Accessed on 27 Dec 2012 from http://etheses.bham.ac.uk/253/

Turell, M. T. 2010. 'The use of textual, grammatical and sociolinguistic evidence in forensic text comparison.' *The International Journal of Speech, Language and the Law*. 17(2): 211-250.

Integrating corpus linguistics and spatial technologies for the analysis of literature

Patricia Murrieta-Flores
Lancaster University
P.Murrieta-flores@lancaster.ac.uk

Ian Gregory
Lancaster University
I.Gregory@lancaster.ac.uk

David Cooper
Manchester Metropolitan University
D.Cooper@mmu.ac.uk

Christopher Donaldson
Lancaster University
Cedonald@stanford.edu

Alistair Baron
Lancaster University
A.Baron@lancaster.ac.uk

Andrew Hardie
Lancaster University
A.Hardie@lancaster.ac.uk

Paul Rayson
Lancaster University
P.Rayson@lancaster.ac.uk

1 Introduction

The purpose of this paper is to demonstrate how the integration of Computational Linguistics and Geographic Information Systems (GIS) can be used in an innovative way, to explore geographically a corpus of traditional literature. The approaches taken in this research enable the identification of unknown patterns and reveal new layers of meaning that might be not discovered through the simple reading of the texts.

Spatial technologies such as GIS have become increasingly used across the arts and humanities with the generalisation of the so-called 'spatial-turn' (Cooper and Gregory 2011). GIS are computer-based systems designed to store, transform, process, analyse and display data from the real world that is spatially-referenced, this is to say, that have assigned geographic co-ordinates (Burrough 1986; Star and Estes 1990; Longley et al. 2011). These systems have been extensively used for the analysis of quantitative data and cartographic sources in disciplines such as Geography and Environmental sciences, and more recently in areas within the humanities such as Archaeology. In fields where the information is traditionally textual-based such as History and Literary studies, different projects have successfully used Natural Language Processing (NLP) techniques to identify place-names in large corpora, assigning coordinates to them through a gazetteer (a process called geo-referencing) (Grover et al. 2010; Gregory and Hardie 2011; Yuan 2010). In the case of Literature, this has allowed the creation of reader-generated mappings of different types of literary geographies (Baer and Hurni 2011; Piatti et al, 2009). The challenge, however, has been not only the integration of unstructured texts into a GIS environment, but also, to explore these sources through spatial analysis. Combining NLP and GIS techniques, the intention of this paper is to illustrate how this interdisciplinary approach is opening up further thinking about the geographies depicted by literary writers.

2 The Lake District Landscape Writings: A working example

The Lake District (north-west England) have been the scene of an exceptionally rich literary tradition that includes guide books, first-person journals, topographic prose texts, poems, novels and plays. These pieces have been extensively influenced by the singularity of the landscape in this region, and therefore, they are highly related to its geography. Although our current work includes the analysis of the corpus of Lake District landscape writings from 1750 to 1900, as an example of that described in the introduction, two famous accounts of tours are explored here: Thomas Grey's tour in the Autumn of 1769; and Samuel Taylor Coleridge's walking expedition of August 1802. They constitute accounts of actual tours and although both are relatively short (approximately 10,000 words), these texts provide ideal material for spatial exploration due to their geographical nature.

In essence, both tours have vast literary significance, as they were of great influence in the construction of the cultural history and identity of the Lake District. Nevertheless, they also are substantially different, and a comparison between their depictions of space and place within the Lakes, may reveal interesting key aspects of the different traditions they came from. In one hand, Gray's work can be interpreted as a proto-Picturesque text, where his representation of spatial experience is related to the characteristics that would become traditionally associated to the touristic way of perceiving the Lakes Landscape. On the other hand, Coleridge's favoured an embodied experience of the Lakes rather than the position of physical detachment from the landscape usually preferred by the Picturesque tourist (Gregory and Cooper, 2009).

3 Mapping the Lakes' literature

After the digitisation and XML tagging of the texts, the Ordnance Survey's 1:50,000 gazetteer was used to identify and assign coordinates to all place-names. This was transformed into a GIS database from which maps addressing different questions were produced. One of the most interesting aspects to explore in these writings was the differences of embodied movement and places experienced by the authors through the landscape. For this, a map of the places visited by both authors was created (Fig. 1). The figure shows not only the routes followed by the authors, but also whether they visited a place or just mentioned it.

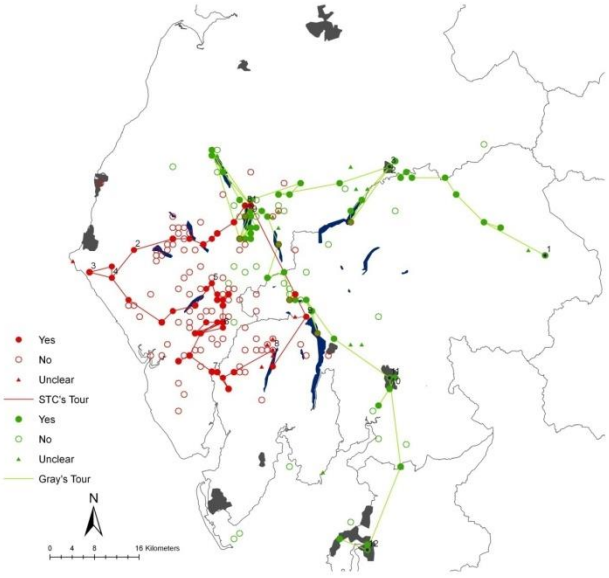


Figure 1: Comparative map of Gray (green) and Coleridge (red) Tours. In full dots sites that were visited by the author and in circles sites just mentioned by the author.

This simple map allows to illustrate how Gray in his 1769 tour, visited mainly places in the eastern side of the Lakes, while Coleridge, would keep his 1802's expedition mostly in the western region. Interestingly, the reason for this spatial difference is possibly related to their own cultural traditions and contexts. Gray's journey is determined by his status as a tourist and outsider, where his itinerary is mostly organised around known touristic bases such as Penrith, Keswick and Kendal. Coleridge on the other hand, visited places according to his identity as environmental insider and Post-picturesque writer (Cooper and Gregory, 2011).

Although useful, these point maps can be difficult to interpret and they are unable to represent whether, for instance, the authors referred to the same place more than once. These problems can be resolved with the use of further spatial techniques such as density smoothing. This technique measures the density of events in a determined location. While figures 2 and 3 identify the places most mentioned by each author, figure 4 shows the surface of the combined places mentioned by both.

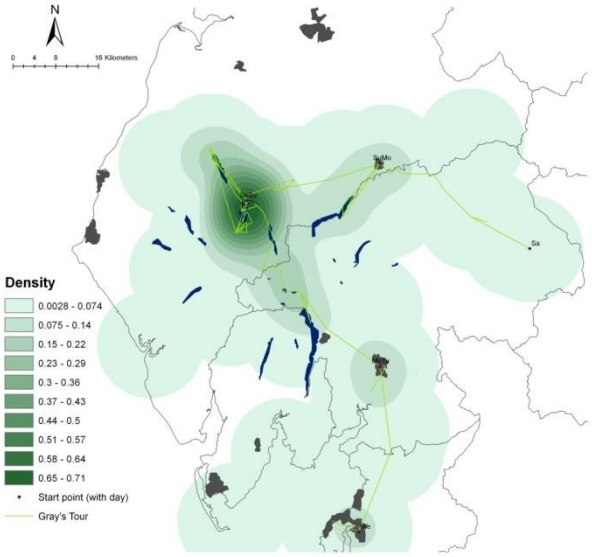


Fig. 2 Density smoothed surfaces of places mentioned by Grey.

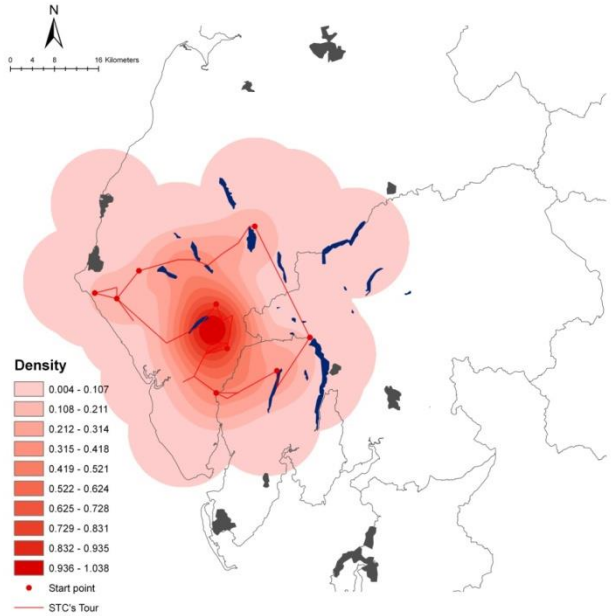


Fig. 3 Density smoothed surfaces of places mentioned by Coleridge.

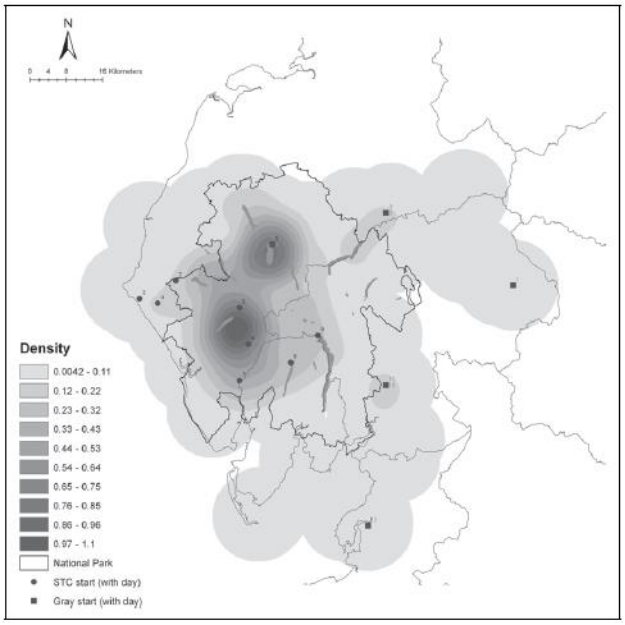


Fig. 4 Density map of both tours.

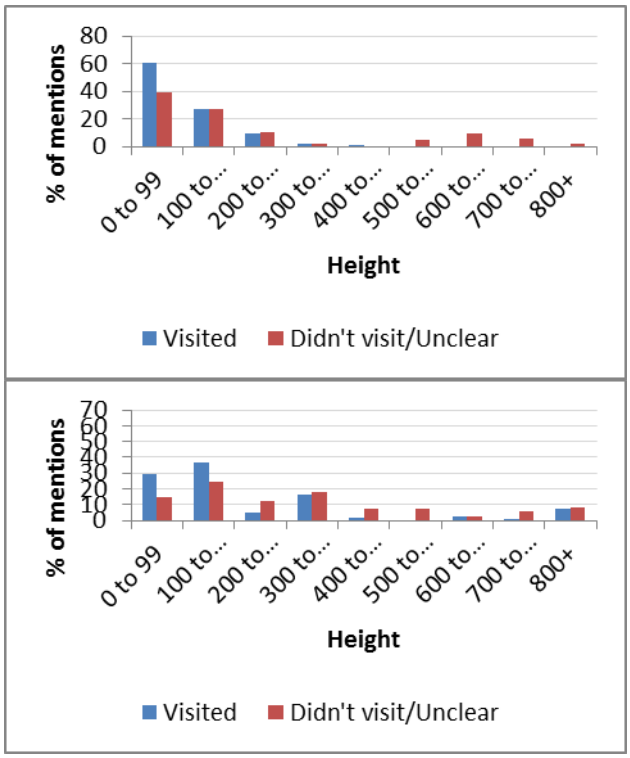


Fig. 5 Frequency distribution of heights of places mentioned by Gray (above) and Coleridge (below)

This technique illustrates that both writers produced quite comprehensive accounts of their experiences within the Lakes (Fig. 2 and 3), but also that they spent significant part of their tours outside the actual Lake District (Fig. 4). The maps also revealed that Grey paid more attention to urban centres where the pattern clusters mainly in the area around the town of Keswick (Fig. 2).

Finally, using a Digital Elevation Model (DEM) or digital representation of the terrain, in this case from the Lake District, a further hypothesis was tested. Grey's proto-Picturesque approach to landscape experience may have influenced his journeys, resulting in the visit of mainly low terrains which concentrated mainly on the lakes. In contrast, Colerigde would be expected to visit higher altitudes, experiencing more adventurous terrains. A comparison of the heights visited (Fig. 5) reveals that this hypothesis might be valid. While Gray's visited mainly places below 200m, Coleridge actually visited places in higher grounds and also paid attention to the ones in between 300 and 600m which Gray basically ignored.

4 Conclusion

We are at early stages in the development of the combination of Computational Linguistics and GIS. Nevertheless, research in this area is proving fruitful already, allowing us to explore literature in a completely new and different way.

Expanding on the work shown here, we are taking these approaches further using: (1) methodologies developed in computational linguistics that enables the automatic identification and extraction of place-names, in addition to collocation analysis; and (2), developing methodologies with spatial statistics. In this manner, using corpus linguistics and spatial methods, this research is starting to reveal the overarching spatial patterns that emerge from mapping at a large-scale the geo-referenced corpus of the Lake District writings composed between 1750 and 1900. With these techniques, the locations that became over and/or under determined by the authors in this period are being identified, pointing to new ideas in terms of perception and construction of social and literary landscapes. In addition to that, the use of these techniques is also answering questions that have been traditionally addressed, but have never been explored. The literary history of the Lake District, as shown here, has been partly shaped by a clear dichotomy between the depictions of space, place and landscape by writers living within the region, and authors who have visited the area as cultural tourists. Using collocation analysis in combination with spatial techniques such as visibility and network analysis, it could be explored whether there was effectively a tendency for writers who lived in the Lake District to concentrate their writings on their immediate locales, and whether it is true that the touristic outsider habitually flit from place to place.

In conclusion, the general purpose of the paper is to show how recent work in the integration of Corpus Linguistics with GIS is opening novel venues for the interpretation of literary work.

Acknowledgments

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850).

References

Baer, H., and Hurni, L. (2011). Improved Density Estimation for the Visualisation of Literary Spaces. The Cartographic Journal, Special Issue Cartographies of Fictional Worlds, 48(4), 309-316.

Burrough, P.A. (1986). Principles of geographical information systems for land resources assessment. Oxford: Clarendon Press.

Cooper, D., and Gregory, I.N. (2011). Mapping the English Lake District: A literary GIS. Transactions of the Institute of British Geographers, 36, 89-108.

Gregory, I.N., and Cooper D. (2009). Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems: A Literary GIS of Two Lake District Tours. International Journal of Humanities and Arts Computing, 3, 61-84.

Gregory, I.N., and Hardie A. (2011). Visual GISting: Bringing together corpus linguistics and Geographical Information Systems. Literary and Linguistic Computing, 26, 297-314.

Grover, C., Tobin, R., Woollard, M., Reid, J., Dunn, S. and Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. Philosophical Transactions of the Royal Society A, 368, 3875-3889.

Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (2001). Geographical Information Systems and Science. New York: John Wiley.

Piatti, B., Reuschel, A.-K., Bär, H. R., Cartwright, W., Hurni, L. (2009). Mapping Literature. Towards a Geography of Fiction. In W. Cartwright, et al. (Ed.), Cartography and Art (pp. 177-192): Wiesbaden.

Star, J., and Estes, J. E. (1990). Geographic Information Systems : An Introduction: Prentice Hall.

Yuan, M. (2010). Mapping text. In D.C. Bodenhamer, J., and Harris, T. (Ed.), The Spatial Humanities: GIS and the future of humanities scholarship (pp. 109-123). Bloomington: Indiana University Press.

Citation in student assignments: a corpus-driven investigation

Hilary Nesi
Coventry University
h.nesi@coventry.ac.uk

1 Introduction

Most investigations of citation practices have concentrated on the output of expert academic writers (for example Hyland 1999; Harwood 2009) or postgraduate research students (for example Thompson 2000; Charles 2006; Pecorari 2006). Borg (2000) and Harwood and Petric (2011) are amongst the very few papers to investigate the way students cite sources in their assessed coursework, but Borg looked at only 16 introductory assignments rather than assignments written in response to a taught module, and Harwood and Petric drew their data from interviews, concerning themselves with student writer attitudes. Neither of these studies adopted a corpus approach.

Generally the assumption amongst applied linguists seems to be that the successful thesis and the published research article, in appropriate disciplines, constitute good models for all non-expert academic writers, and that citation practices identified in these genres can be usefully transferred to undergraduate and masters level coursework.

In practice, however, students taking taught courses rarely aim to present new theory or original research, and usually cite for different and possibly less interesting reasons than their more academically-advanced counterparts. They do not usually need to argue for the centrality of their topic (as this has often been decided for them), nor do they need to identify a gap in the existing research or 'occupy a niche'. Instead they need to display the fact that they have read and understood a range of appropriate academic texts, and that they are conversant with academic citation conventions. Students in search of writing skills support would be best served by descriptions of the practices of successful writers undertaking the same sort of assessment tasks as themselves.

This paper presents the initial findings from a corpus-driven investigation of citation practices in proficient university assignments. It shows the frequency of different citation patterns, and lists some common reporting verbs used by students.

2 Method

The corpus under investigation was the British Academic Written English (BAWE) corpus[1], a 6.5 million word collection of proficient student writing across more than 30 disciplines, and at four levels of study (first year undergraduate to taught Masters level). All the assignments were produced for assessment purposes as part of degree programmes, and were awarded good marks by course tutors. The corpus is part of speech tagged with the UCREL CLAWS7 tagset[2].

The corpus was accessed via *Sketch Engine*[3]. Corpus Query Language (CQL) was used to identify as many instances of citation as possible. As there is no tag which identifies referencing to other sources, the corpus was interrogated using the following six queries:

1. [tag = "MC|MCMC"] [word = "\)"] [tag = "VV|VO|VVD|VVZ|VM"] [tag = "VVI"]? [tag = "CST"]
2. [lemma ! ="such"] [lemma="as"] [tag = "NP.?"] [] {0,5}[word = "\("] [tag = "MC.?"] [word = "\)"] [tag = "VV.?|VO|VM|VB.?| VD.?|VH.?"]
3. [lemma ="such"] [lemma="as"] [tag = "NP.?"] [] {0,5}[word = "\("] [tag = "MC.?"] [word = "\)"] [tag = "VV.?|VO|VM|VB.?| VD.?|VH.?"]
4. [word = "\("] [tag = "NP.?"] [tag = "MC.?"] [] {0,5} [word = "\)"]
5. [word ="ibid"]
6. [word ="op"][word = "\."][word ="cit"]

The query [textpart != "bibliography"] was added to all of these to exclude instances within the bibliography section of assignments.

The first of these queries identifies the pattern: *number + end bracket + lexical verb (or modal verb + infinitive) + 'that'*. Examples of this pattern are 'Flavell (1977) suggested that stages in development must show distinct, qualitative changes' and 'research such as that carried out by Ferrari et al. (referenced in 9) showed that….'

The query excludes present and past participles of the lexical verb, thus avoiding patterns such as 'Assessment of sea-level change has shown variation between regions partly due to erosion and aggradation (Knapp 1997: 155) meaning that an assumption of a certain rise/fall in one area cannot be extended a priori'. It also excludes singular cardinal numbers (MC1), thus avoiding the pattern 'This light-response curve (fig. 1) shows that….'. There are eight such examples in the corpus. Unfortunately the query also excludes two genuine citations:

1. Mott G. (1997, p. 1) highlights that
2. Bhardwaj et al (16, reviewed in 1) demonstrated that

The second of these queries identifies the pattern: *as (but not 'such as')+ proper noun + up to five words + number within brackets + lexical, auxiliary or modal verb*. It yields examples such as 'as Young (1992) notes' and 'The findings of Bowen and Siegel (1970) as well as Greene (1972) showed a relatively strong correlation', but excludes expressions identified in Query 3, where sources are cited as examples, for example 'Radical feminists such as Andrea Dworkin (1976) have broadened the definition of violence', and 'behaviourists such as Skinner (1938) provide modern psychologists with principals'.

The fourth query simply searches for non-integral citations, i.e. proper names and numbers within brackets, such as '(McCracken 1990:24)'.

Queries 5 and 6 search for the use of *ibid* and *op.cit* in the corpus.

Between them, these queries were designed to capture all instances where sources are named within the text or in footnotes, rather than via the Vancouver "author-number" system. However inevitably some instances were missed, and some stretches of text which conformed to the pattern but which were not citations crept in.

The searches were conducted across the entire corpus, and then in the four disciplinary groupings (Arts and Humanities (AH), Life Sciences (LS), Physical Sciences PS) and Social Sciences, (SS)). The distribution of reporting verbs across individual disciplines and genres was also of interest, but this aspect of the study is beyond the scope of this paper.

3 Results

The results from the six queries are shown in Table 1. They indicate a marked preference for the non-integrated citation (Query 4). The highest numbers were in the Social Sciences, and the lowest in the Physical Sciences where the "author-number" system is generally preferred.

In Queries 1 and 2 the tenses of reporting verbs were recorded. The results for VVZ (the 's' form of the lexical verb) and VVD (the past tense of the lexical verb) are shown in Table 2. Other forms occurred much more rarely.

Some of the commonest reporting verbs are listed in Table 3, in alphabetical order.

[1] See www.coventry.ac.uk.bawe
[2] The tagset is listed at http://ucrel.lancs.ac.uk/claws7tags.html
[3] The BAWE corpus in Sketch Engine is freely accessible at https://ca.sketchengine.co.uk/open/

Query	AH	LS	PS	SS	Total
1.	24.5	40.5	5.8	46.5	117.3
2.	7.4	4.4	1.2	15.7	28.8
3.	0.4	0.2	0.1	1.3	2.0
4.	160.9	80.4	10.3	209.3	461.4
5.	24.6	2.4	-	24.8	51.8
6.	-	-	-	0.1	0.1

Table 1: Distribution of types per million words

Verb form	Query 1	Query 2	Total
VVZ	570	113	683
VVD	409	41	563
Total	979	154	

Table 2: Tenses of reporting verbs

Argue	Believe	Claim	Conclude
Demonstrate	Establish	Explain	Find
Highlight	Hypothesize	Illustrate	Indicate
Infer	Note	Observe	Propose
Reason	Report	Reveal	Show
State	Stress	Suggest	Write

Table 3: Common reporting verbs

4 Discussion

It is perhaps unsurprising that the non-integral citation form was favoured, as this is in many ways the least demanding form. The absence of a reporting verb means that in many cases the writer does not evaluate the source, but merely demonstrates awareness of its relevance. In some sorts of assignment, of course, this approach is entirely appropriate. Students do use a wide variety of reporting verbs, slightly more often in the present tense than in the past. It would appear that the 'ibid' and 'op cit' forms of citation are almost never used by non-research students.

This paper only outlines the initial results of an in-depth investigation of the distribution of citation types across disciplines, but student writers and tutors will be able to interrogate BAWE in *Sketch Engine Open* using the CQL query types provided here, to find for themselves an interesting range of citation practices in the disciplines and genres that interest them most.

References

Borg, E. (2000). "Citation practices in academic writing". In P. Thompson (ed.), *Patterns and perspectives: Insights into EAP writing practice*. Reading, UK: Centre for Applied Language Studies.

Charles, M. 2006. "Phraseological patterns in reporting clauses used in citation: a corpus-based study of theses in two disciplines". *English for Specific Purposes* 25 (3) 310-331.

Harwood, N. 2009. "An interview-based study of the functions of citations in academic writing across two disciplines". *Journal of Pragmatics* 41 (3) 497-518.

Harwood, N. and Petric, B. 2011. "Performance in the citing behavior of two student writers". *Written Communication* 29 (1): 55-103.

Hyland, K. 1999. "Academic attribution: citation and the construction of disciplinary knowledge". *Applied Linguistics* 20 (3): 341-367.

Pecorari, D. 2006. "Visible and occluded citation features in postgraduate second-language writing". *English for Specific Purposes* 25 (1) 4-29.

Thompson, P. 2000. "Citation practices in PhD theses". In L. Burnard and T. McEnery (eds.) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang.

Reporting the 2011 London riots: a corpus-based discourse analysis of agency and participants

Maria Cristina Nisco
University of Naples Federico II
mariacristina.nisco@unina.it

Extensively covered by British newspapers, the riots that occurred in London in August 2011 have been defined as "the most arcane of uprisings and the most modern, […] [whose] participants, marshalled by Twitter, are protagonists in a sinister flipside to the Arab Spring. The Tottenham summer […] is an assault not on a regime of tiranny but on the established order of a benign democracy".[1]

Since the roles assigned to the various participants involved in the so-called 'Tottenham summer' appear of central importance in the interpretation of events, this paper investigates the ways in which they are represented by the British press. It explicitly addresses the attribution of agency enacted by the press through a series of linguistic strategies, something which seems to be pivotal in all discourses surrounding and construing the riots. Indeed, the analysis of the depiction of the social actors can provide useful insights on the representations of both individual and collective identities that are conceptualized and reproduced *in* and *by* discourse (van Leeuwen 1996).

Moving from the assumption that media are people's first contact with the external world (van Dijk 1991), and that newspapers, in particular, create public identities for social groups and individuals through specific textual strategies (Fairclough 1995), this paper focuses on the representation of the participants to the riots as portrayed by four British newspapers, namely the *Guardian*, *The Times*, the *Daily Mail* and the *Daily Mirror*. A corpus of approximately 850 articles has been collected over a period of time spanning from August 1st to December 31st, 2011, and downloaded from the web archive LexisNexis.

The paper attempts a corpus-based discourse analysis (Partington et al. 2004, Baker 2006, Baker et al. 2008, Gabrielatos and Baker 2008, Morley and Bayley 2010) examining key words and clusters in the corpus, thus considering the emerging patterns in the discursive construction of agency. Indeed, a focus on the strongest key words and clusters, combined with concordance analysis, can provide helpful indications on the discursive construal of the actors and causes of the riots. More specifically, the paper offers an account of the main lexico-grammatical devices – such as nominalisations, active/passive voice – as well as the negative portrayal of some participants and the generalised references used to blame certain subjects for the riots, while suggesting particular interpretations of the identities of the different groups involved in the events (Wodak et al. 2009).

While shedding light on the dominant interpretative frameworks, corpus-analysis data allow the identification of the main actors in the news reports: the rioters, the police and other State authorities, the residents and local people whose businesses, shops and homes were damaged (in addition to Mark Duggan, whose shooting sparked the riots).[2] A preliminary analysis of the word list shows that 'police' is the most frequently occurring lexical word within the corpus (freq. 3.772 – 0,66%), followed by 'riots' (freq. 2.511 – 0.44%) and 'people' (2.312 – 0,40%), ranking among the first 35 items, within a majority of function words – which can be regarded as surprising in itself, considering that the corpus was collected by searching for the key word 'riots'. A quick look at the concordance plots of two of the main participants – the police and the rioters – also confirms an utterly prevailing presence of the former within the news reports: the plot of the item 'police' displays an extremely high number of hits, something which does not apply to the items 'rioters' or 'riots', as one would have expected. In this sense, the news reports included in the corpus appear to involve ambivalent relations of power between (conventionally identified) powerful and non-powerful participants to the events.

In order to examine not only the frequency in use of specific items (emerging as key words), but also the phraseological structures in which such items occur, then clusters need to be taken into account too. Analysis of the clusters of the item 'police' reveals an overall clearly attributed agency to the police for the shooting of Mark Duggan – that is usually achieved through nominalisations and explicitly expressed agents in passive sentences. Among the most recurrent clusters: 'was shot dead by police', 'the police shooting of Mark, 'the fatal shooting by police', 'whose shooting by police sparked', 'Duggan was

[1] *Telegraph*, 'London riots: the underclass lashes out', published 08/08/2011.

[2] The software used for corpus analysis is Antconc 3.2.2.

killed by police'.[1] Additionally, further hints are also given by the verbal processes and actions (or rather failure to act) attributed to this actor in particular: 'police apologized to the family', 'failure of police to bring', 'metropolitan police could have managed', 'police had lost control of', 'police response had not been'.

However, detailed investigation has shown that the lexical item 'police' is employed in a variety of contexts not necessarily connected to agency. Therefore, in the attempt to exclusively focus on it, a cluster analysis of the co-occurrences of the items 'police' and 'shoot*' (a verb potentially signalling their agency) was carried out. According to results, with the item 'shot', police is usually the agent (e.g.: 'Duggan had been shot by police') – while only in a few occurrences the agent is Duggan (e.g.: 'Duggan had shot at police') – and, similarly, with 'shooting', agency seems to be almost exclusively attributed to police (e.g.: 'the police shooting of', 'the fatal shooting by police'). Nonetheless, more interestingly, with the item 'shoot' the police appears as the affected participant rather than the agent (e.g.: 'rioters tried to shoot down a police helicopter', 'rioters want to shoot at armed police').

Further analysis of the co-occurrences of the items 'police' and 'kill*' (another significant verb of agency) has shown that the police is invariably identified as the agent ('police kill an alleged crack-dealing gangster', 'police weapons used to kill Duggan', 'guy killed by police', 'the bullet that killed was police', 'the police officer who shot and killed'), seemingly standing in stark contrast to previous trends in the press that often excluded or downplayed the role of police in riots or presented it as a victim (van Dijk 1989).

On the other hand, the analysis of the clusters of the item 'rioters' appears more complex since there is a variety of ways through which such participants can be referred to. Indeed, 'rioters' is often used together with 'looters', or replaced by another frequent epithet, namely 'offenders', which is pre-modified by the adjective 'young', in most cases. As a matter of fact, by pointing to a specific portion of the population, the item 'young' seems to be even more significant in the identification of rioters, as some of the clusters retrieved show (pre-modification playing an important role in the depiction of these actors): 'young people involved in the', 'a group of young men', families with disruptive young people', 'of young, alienated, disaffected youth'.

The items 'rioters' and 'looters' were then examined in relation to the verbal processes with which they were often seen to co-occur: in most of the cases they were material processes, such as 'throw*' ('some of the rioters threw petrol bombs'), 'loot*' ('witnesses described rioters looting'), 'steal*' ('rioters turned to a journalist, stealing'), 'destroy*' ('rioters were destroying their own communities'), 'attack*' ('a student attacked by rioters'), 'ransack*' (as looters ransacked her supermarket'), taken as some among the most representative examples.

A closer investigation of the lexico-grammatical markers of agency attribution and of the processes in which the different social actors are involved (whether they be violent material processes – 'the police officer who shot a guy', 'rioters were throwing missiles and petrol bombs' – or mental or verbal processes – 'the Metropolitan police officers devised a policy', 'police warned that') will help assessing the diverging conceptualizations of agency for the events of the so-called 'Tottenham summer'.

Besides considering data from a wider perspective taking into account the whole corpus, the paper will furthermore highlight the contrasts and similarities between the newspapers under investigation in connection to the different contexts of production and audience they address.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., Gabrielatos, C., McEnery, T., Wodak, R. et al. 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press". *Discourse & Society* 19 (3): 273-305.

Cotter, C. 2010. *News Talk. Investigating the Language of Journalism*. Cambridge: CUP.

Fairclough, N. 1995. *Media Discourse*. London: Edward Arnold.

Gabrielatos, C. and Baker, P. 2008. "Fleeing, sneaking, flooding: a corpus-analysis of discursive constructions of refugees and asylum seekers in the UK press 1996-2005". *Journal of English Linguistics* 36 (1): 5-38.

Hall, S. et al. 1978. *Policing the Crisis: Mugging, the State, and Law and Order*. London: Macmillan.

Morley, J. and Bayley, P. (eds.) 2010. *Corpus-Assisted Discourse Studies on the Iraq Conflict. Wording the War*. London: Routledge.

Partington, A., Morley, J. and Haarman, L. (eds.) 2004. *Corpora and Discourse*. Bern: Peter Lang.

Van Dijk, T.A. 1989. "Race, riots and the press: an

[1] Clusters were identified within a 5L-5R size, minimum frequency 3. Unless otherwise specified, all searches for clusters adopted the same parameters.

analysis of the editorials of the British press about the 1985 disorders". *Gazette* 43: 229-253.

Van Dijk, T.A. 1991. *Racism and the Press*. London: Routledge.

Van Leeuwen, T., 1996. "The representation of social actors". In Caldas-Coulthard, C.R. and M. Coulthard (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge.

Waddington, D. 2007. *Policing Public Disorder: Theory and Practice*. Oxford, Willan.

Wodak, R. et al. 2009. *The Discursive Construction of National Identity*. Edinburgh: EUP.

Semantically profiling and word sketching the Singapore ICNALE Corpus

Vincent B Y Ooi
National University of Singapore
vinceooi@nus.edu.sg

1 Introduction and rationale

This paper has the dual purpose of outlining the compilation of the Singapore component of the ICNALE project (Ishikawa 2011, Ishikawa 2013) and analysing aspects of learner English, in terms of two well-known corpus linguistic tools: Wmatrix for semantic profiling (Rayson 2008) and the Sketch Engine for word sketches (Kilgarriff et al 2004).

Inspired by the influential ICLE project (Granger 2011), the ICNALE project offers a suite of learner corpora with significantly Asian perspectives that complement the largely European ones (offered by the ICLE project). The ICNALE project is specific in its parameters by restricting the remit to only i) tertiary students preferably between the ages of 18 to 24 and ii) controlled topics of 200 to 300 words each on 'part-time job' and '(non-) smoking' respectively.

As one of the two most recent additions (the other being the Philippines), the Singapore learner corpus was compiled in the latter part of 2012. It is important to include Singapore because English in the country is neither totally 'native' nor 'foreign'. For instance, while English is taught as a 'first language' in schools, it is the most frequently spoken language at home among only 32.6 percent of the Chinese, 17.0 percent of the Malays and 41.6 percent of the Indians that comprise the major ethnic groups in the country.[1]

Both Wmatrix and Word Sketch provide a degree of sophistication beyond the normal search interface for raw concordances, collocations and wordlists.

2 The ICNALE Project and the Singapore ICNALE Corpus

Started in 2010, the 3-year ICNALE project now totals more than a million words that span 11 countries and areas. There is an online user

[1] Census of Population 2010, http://www.singstat.gov.sg/pubn/popn/c2010sr1/cop2010sr1.pdf and http://www.singstat.gov.sg/news/news/press31082010.pdf

interface[1] that facilitates access to corpora from the following Asian countries: Hong Kong, Pakistan, mainland China, Indonesia, Japan, South Korea, Singapore, the Philippines, Thailand and the Republic of China (Taiwan). The interface also allows separate comparisons between these corpora and essays written by native speakers and native learners. The following specimen screenshot shows the various functions, with the Keyword-in-Context view for Singapore selected (see Figure 1):

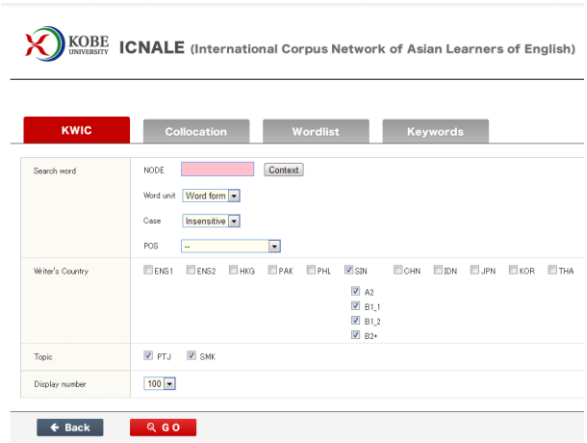


Figure 1: The ICNALE Online Interface

Of the corpora available, the Singapore corpus totals 96733 tokens and 5470 word types. Most of the tertiary participants for this corpus study at the National University of Singapore, followed by those from Nanyang Technological University, SIM University and various polytechnics. Coming from a variety of disciplines (including Psychology, Chemistry, Sociology and Engineering), many of the 200 participants have passed the Cambridge A-level examination with creditable scores in the General Paper. Each participant was asked to write 200-300 words on the topic of 'part-time job', and repeated the assignment on the topic of '(non) smoking'. With the assistance of anti-plagiarism software, essays that copied material from the Web were rejected.

3 The learner corpus

Ishikawa (2011:3) appropriately identifies two central motivations for the study and use of learner corpora: i) studies of "interlanguage" and ii) identifying learner errors, overused and underused patterns. 'Interlanguage' is a concept embraced perhaps especially by EFL (English as a Foreign Language) countries which construe their own production of this second/foreign language as neither their first language nor 'perfectly native'. In 'outer circle' or ESL (English as a Second Language) countries such as Singapore, the concept of 'interlanguage' is much more played down. Notwithstanding this, comparing a learner corpus with an equivalent native one (and, for that matter, between various learner corpora collected in a similar manner) is a worthwhile enterprise: over/under-used patterns and learner errors can be identified so that language teachers and students can focus on reducing their production and reception of these features.

It is sometimes said that non-native learners tend to overuse the term *important* compared with native learners and speakers who have recourse to other adjectives 'such as *critical*, *crucial*, *major*, *serious*, *significant* and *vital*'. (Cheng 2012: 173). It is intriguing that the online user interface shows that, in terms of 'per million word' (PMW) adjusted frequency, Singaporean users achieve 3370.10 compared with native learners at 2574.75, Hong Kong students at 3079.52, China students at 3345.10, and Japanese students at 4775.20 respectively. Does this mean that Singaporean university students are 'less native' than Hong Kong and mainland China students?

4 Applying Wmatrix to the learner corpus

The Wmatrix program (Rayson 2008)[2] represents an online integrated corpus linguistic environment in which texts can be loaded and analysed for word frequency profiles and concordances, or automatically annotated in terms of part-of-speech and word-sense tags (based on 21 semantic categories).

A screenshot of the top 5 semantic categories in the Singapore part-time job sub- corpus (O1-raw frequencies), as compared with the British National Corpus written sampler (as a prototypical example of everyday writing), shows the following categories generated by the software for part-time job (see Figure 2):

O1	%1	O2	%2	LL	
2629	5.70	3691	0.38 +	8015.66	Education in general
1854	4.02	3381	0.35 +	4971.97	Work and employment: Generally
1274	2.76	8327	0.86 +	1134.16	Time: Period
228	0.49	307	0.03 +	708.26	Learning
340	0.74	1060	0.11 +	648.54	Able/intelligent

Figure 2: Top 5 semantic categories in the ICNALE-SIN corpus, compared with the BNC Written Sampler

[1] http://language.sakura.ne.jp/icnale/icnale_online.html

[2] http://ucrel.lancs.ac.uk/wmatrix/

Under "Education in general", significant terms include *college students, university, students, and university education*. This is not surprising, given that the participants are asked to write on the topic of whether a tertiary student should take up a part-time job while in university. In the second and third semantic categories, the concepts of "part-time" and "job" become the most significant ones. In the fourth category, the terms *learning, glean, find out* and *internalising* all point to the value of part-time jobs for one's learning journey. Under the 'Able/intelligent' semantic category, the terms *skills* ('leadership', 'interpersonal' etc) and *ability* are significant.

The screenshot (see Figure 3) somewhat changes with the native learner sub-corpus on part-time jobs (as compared with the BNC written sampler also):

01	%1	02	%2	LL	
1938	4.64	3691	0.38 +	5413.95	Education in general
1361	3.26	3381	0.35 +	3273.72	Work and employment: Generally
5187	12.43	72023	7.44 +	1103.41	Pronouns
198	0.47	307	0.03 +	611.39	Learning
871	2.09	8327	0.86 +	490.71	Time: Period

Figure 3: Top 5 semantic categories in the ICNALE-Native learner corpus, compared with the BNC Written Sampler

Although the top two categories are similar, the third category is now listed as 'Pronouns'. Examining the underlying concordances for this category, we can see that there is a greater willingness among native speakers/learners to use pronouns to personalise the experience, or to use them to generalise hypothetically (such as *If you have a part-time job, you can…*)

5 Word sketching the Singapore Learner Corpus and the Native Corpus

While the Sketch Engine (Kilgarriff et al 2004, Kilgarriff 2009) does not have word-sense tagging that Wmatrix has, it is also 'a corpus tool which takes as input a corpus of any language and a corresponding grammar patterns and which generates word sketches for the words of that language. A word sketch is a 'one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour.' For too long, we can only scan the co-text (or immediate environment surrounding the lexical item) impressionistically or do manual counts painfully. As Kilgarriff (2009) puts it, most language learners are 'scared off' by concordances that become too long or manually difficult to analyze. The Sketch Engine allows the entire input text to be annotated and relevant frequencies derived from the corpus in a much more automatic manner. For instance, in the case of *smoking* (a topical noun in the ICNALE corpus), the word sketch for the native-speaker collective construal of the term is as follows (see Figure 4):

Figure 4: Word sketch for *smoking*, from the ICNALE native speaker corpus

From this sketch, under "object_of", the top verbs associated with 'smoking' include *ban, allow, quit, think, prohibit, see* (as in *I was very disappointed to see smoking banned*) etc.

Compare this screenshot for the same noun in the Singapore ICNALE corpus (see Figure 5):

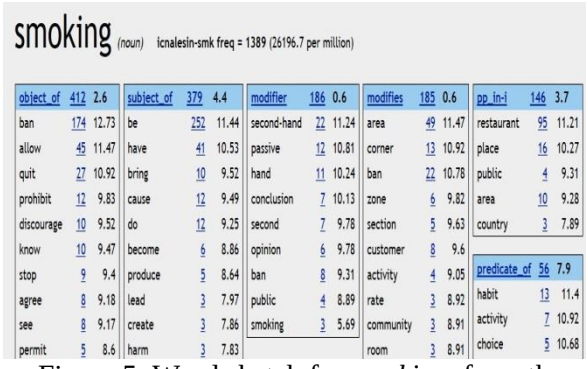


Figure 5: Word sketch for *smoking*, from the ICNALE-SIN corpus

Verbs that collocate with 'smoke' similarly include *ban, allow* and *quit* (as evidence of Singaporeans' native mastery of the language), but then the preferred verbs tend to become more negative: *prohibit, discourage* and *know* (as in *Everyone knows smoking is bad*). This negativity resonates in a public environment in which the government not only encourages its citizens not to smoke but the act is increasingly prohibited in nearly all public spaces.

6 Some conclusions

A learner corpus is an interesting object of study because it 'allows us to quantify the different kinds of mistakes that learners make and can teach us how a learner's model of the target language develops as they progress.' (Kilgarriff 2009). Extending this observation to the ICNALE project, the similar methodology employed for the various countries that exemplify English proficiency in its major levels – ENL, ESL, and EFL – offers systematic insights into the appropriation of lexical priming by different Asian learners (Ooi 2013). Sophisticated corpus tools allow us to trace linguistic variation in a more detailed manner, as demonstrated in this study of the Singapore ICNALE corpus.

References

Cheng W. 2012. *Exploring corpus linguistics: language in Action*. London: Routledge.

Granger, S. 2011. *International corpus of learner English – ICLE*. http://www.uclouvain.be/en-cecl-icle.html

Ishikawa, S. 2011. A new horizon in learner corpus studies: the aim of the ICNALE project, in Weir, G, Ishikawa S, and Poonpon K (eds), *Corpora and language technologies in teaching, learning and research*, Glasgow:University of Strathclyde Press, 3-11.

Ishikawa, S.(ed) 2013. *Learner corpus studies in Asia and the world: Vol 1 Papers from the LCSAW 2013.* Kobe University: School of Languages and Communication.

Kilgarriff, A. 2009. Corpora in the classroom without scaring the students, in *Proceedings of the 18th International symposium on English teaching, Taipei.* http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc

Kilgarriff, A, Rychly P, Smrz P, and Tugwell D. 2004. The Sketch Engine, in *Proceedings of. Euralex*. Lorient, France, July: 105-116. http://www.kilgarriff.co.uk/Publications/2004-KilgRychlySmrzTugwell-SkEEuralex.rtf

Ooi, V. 2013. Lexical priming and Asian learners of English, in Ishikawa (ed), 31-41.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics.* 13:4 pp. 519-549.

Intimations of Spring? Political and media coverage – and non-coverage – of the Arab uprisings, and how corpus linguistics *can* speak to "absences"

Alan Partington
Bologna University
alanscott.partington
@unibo.it

> As one-party states go, Libya is not especially repressive. Gadafy seems genuinely popular […] If [he] is sincere about reform, as I think he is, Libya could end up as the Norway of North Africa (Anthony Giddens, Guardian)

> […] the corpus-based analysis tends to focus on what has been explicitly written, rather than what could have been written but was not or what is implied, inferred, insinuated or latently hinted at (Baker et al 2008: 296)

> [and so] "pragmatic devices and subtle, coded strategies or concepts can not be readily analysed through corpus linguistic means" (Wodak 2007, quoted in Baker et al 2008: 296)

> By contrast […] a critical analysis takes into account absences as well as presences in the data [while] purely descriptive, data-driven approaches are epistemologically inadequate in accounting for the complex linguistic choices made during the processes of production of a text. (Baker et al 2008: 281)

> […] A traditional corpus-based analysis is not sufficient to explain or interpret the reasons why certain linguistic patterns were found (or not found). (Baker et al 2008: 293)

The present paper has two aims. It is firstly an attempt to show how corpus-assisted approach can be helpful in researching a particular historical event and the discussions/representations thereof. Secondly, during the course of this investigation, it become apparent that what was *not* encountered in the data was often as significant as what was present, and it grew ever more apparent that the incorporation of corpus techniques into discourse study offered exceptional opportunities for identifying and quantifying *absences* of various kinds, (see below) as a prelude to their possible

interpretation on the part of the analyst.

A great deal has been written about the media coverage of the Arab uprisings. The current paper looks more generally at the reporting of the Middle East and North Africa (MENA) and compares the way it was discussed in 2011, when the uprisings were taking place, with the year before, one of the aims being to ascertain whether there was any inkling of what was about to occur.

It takes its cue from a parallel research (Partington & Marchi, Forthcoming) into how the uprisings were discussed in the White House press briefings. We noted, in the period immediately before the occupation of Tahrir Square in late January 2011, a distinct *absence* of attention towards the Arab MENA; no statements were made by the administration and no questions were asked by the press about the region (except for a solitary query about the wisdom of the US reopening its embassy in Syria). In other words, both sides, the government and the press corps, seem to have been taken entirely unawares by events.

The next stage of the research is a comparison of how some of the events and actors of the uprisings are reported during 2011 in three media-political sources, White House press briefings, CNN news and, finally, the UK *Guardian* newspaper's own choice of its "best" reports of what it calls "The Arab Spring" (Manhire ed. 2012). Particular attention is paid to evaluation, for instance, how descriptors transform from neutral, even respectful, to highly negative (e.g. "Syria's government" and "President Bashar al-Assad" turn into "vile dictatorship" and "Assad and his Ba'ath party cronies"), and to how and to whom praise and blame is apportioned – and *not* apportioned; that is, when blame is *absent*. We also examine attempts by the US administration in the White House briefings to impose dominant evaluative readings of the MENA events in a process Duguid (2007) calls "forced priming". We then look at the CNN news reports to ascertain whether some of these favoured White House readings are actually adopted or not – are present or absent – in this particular press outlet.

The following part of the examination looks back to 2010, the year before the uprisings occurred with, as already mentioned, the aim of seeing whether the press had any inkling of what was about to occur in the Arab MENA. It is in two sections, moving from the particular to the more general. Firstly, we look at what discussions are present and absent in two UK broadsheet newspapers, the left-leaning *Guardian* and the right-leaning *Telegraph* concerning four of the countries in which uprisings occurred, namely, Tunisia, Libya, Syria and Egypt.

Second, we examine how the Arab MENA and how *Arab* and *Arab*(*s*) in general are discussed We see for instance interestingly different grammatical profiles for the expressions *the Arab world* and *the western world*, the former being very rarely found in primary actor or 'Do-er' role, another significant absence. This was also the year of Wikileaks, covered copiously by the *Guardian* which played an active role in the dissemination of the documents. They proved to contain some surprising information, especially concerning the attitudes of political leaders to other Middle Eastern states.

All of this is to be placed in a wider methodological context. One major criticism leveled at corpus linguistics, particularly from the field of critical discourse analysis, is that although CL may be suited to saying things about what is to be found in a data-set, whether the dataset in question be a corpus, a single text or a particular set of texts, it cannot deal with absences, that is, whatever is not to be found therein (see the set of extracts from Baker at al 2008 that preface this abstract). Here, then, following on highly original work on the topic of CL and absences by Taylor (2012), we respond to this charge by showing how corpus-assisted techniques are invaluable in not only revealing and locating absence but in identifying types of absence and in quantifying it, as precursors to the researcher evaluation of the relevance of absences, in ways which are rigorous, reliable and replicable.

In fact, corpus techniques provide an entirely new dimension into research into absence. Adapting the celebrated terms "known unknowns" and "unknown unknowns", not only can linguists who use corpora investigate "known absence", that is, some feature or behaviour they already know or strongly suspect to be absent from a particular data-set, but they can also, given that the machine-driven phases of corpus research famously throw up entirely unforeseen data, investigate "unknown absence", that is, the serendipitous discovery during the course of research that some feature or behaviour is absent from something or somewhere.

These investigations are conducted within the methodological framework of corpus-assisted discourse studies (CADS). These types of research are eclectic and pragmatic in the techniques they adopt since they are goal-driven, that is, the aims of the research dictate the methodology. Nevertheless CADS tend to display a number of common characteristics and tendencies (Partington et al. 2013). These include, firstly, the emphasis on comparison among data-

sets which entails the use of multiple corpora, and the frequent need to compile "bespoke" (or "*ad hoc*") corpora, including concordance-corpora, and the frequent reordering and subdividing of corpora (for instance, by temporal divisions, say, month by month when, as here, tracking historical phenomena). Secondly, they include the interaction of statistical analysis and close reading not only of short concordance lines but of extended co-texts, including entire texts (here, an article or a briefing). Thirdly, where appropriate they include the combination of corpus-generated observation with data from other external sources (Krishnamurthy 1996).

In the course of the study, we will also reflect on some of the issues, pitfalls and temptations encountered when discussing representations of groups and socio-political-economic events. For instance, how often does the association of a group with a certain characteristic have to occur in our data for us to be justified in talking of stereotyping? One constant temptation is over-reading, to read too much into limited quantities of data, often accompanied by its ugly sibling, the *over-dramatising* of findings.

References

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, A. and Wodak, R. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19 (3): 273–305.

Duguid, A. 2007. Men at work: how those at Number 10 construct their working identity. In G. Garzone & S. Sarangi (eds), *Discourse, Ideology and Specialized Communication*, 453-484. Bern: Peter Lang.

Krishnamurthy, R. 1996. Ethnic, Racial and Tribal: The Language of Racism?. In C.R. Caldas-Coulthard & M. Coulthard (eds), *Texts and Practices: Readings in Critical Discourse Analysis*, 129-149. London: Routledge.

Manhire, T. (ed.) 2012. *The Arab Spring: Rebellion, Revolution and a New World Order*. London: Guardian Books.

Partington A. & Marchi, A. Forthcoming. Using corpora for discourse analysis. In R. Reppen and D. Biber (eds.) *Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.

Partington, A., Duguid A. & Taylor, C. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies*. Amsterdam: Benjamins.

Taylor, C. 2012. And there it isn't: (how) can we access the absent using CADS? Talk given at *CADSConf 2012*, September 13-14th, University of Bologna.

Wodak, R. 2007. Pragmatics and Critical Discourse Analysis. A Cross-disciplinary Inquiry. *Journal of Pragmatics and Cognition*, 15 (1), 203–27.

Using corpus data to calculate a rote-learning threshold for personal pronouns: *You* as a target for *They* and *He*

Laura Louise Paterson
University of Leeds
l.l.paterson@leeds.ac.uk

The acquisition plural and singular second-person pronouns represented by the same phonological and morphological form <you> is largely uncontested in the wider literature. The form is thus used for plural and singular coreference predominantly without academic comment. Yet, in contrast, there has been a strong historical resistance, from both scholars and lay persons, to the parallel usage of *they* as both a third-person plural and third-person singular form. In this paper I use a corpus of UK child/carer transcripts compiled from the CHILDES database (MacWhinney 2000) to provide complementary evidence for Stringer and Hopper's (1998) claim that children receive examples of singular *they* in their L1 input. Furthermore, I argue that the occurrence of this form in L1 input leads to its rote learning and integration into children's internal grammars.

Subscribing to the argument that the personal pronoun paradigm has to be rote learned during L1 acquisition (Rispoli 1994), I argue that the acquisition of *they* in its plural and singular/epicene forms parallels the acquisition of singular and plural *you*, thus creating regularity (or syncretism) in the normally irregular personal pronoun paradigm.

Analysing over 6,000 pronominal tokens from four child/carer pairings, based on primary data from Theakston, Lieven, Pine and Roland (2001), I argue that if the ratio of plural and singular *you* received as L1 input is enough to facilitate the acquisition of the two second-person pronouns then there is no reason why this ratio cannot also apply to the third-person forms. Thus, the second-person pronoun distribution in L1 input is used to calculate a threshold level for rote-learned pronoun acquisition. Positively, the data from the CHILDES corpus indicates that singular *they* (in comparison to plural *they*) is given to children in a ratio which surpasses this benchmark figure.

I contrast my analysis of singular *they* with a discussion of why its main opponent for epicene status – generic *he* – cannot represent regularity in pronoun acquisition. Evidence from the CHILDES corpus indicates that children do not receive enough tokens of generic *he* (again using the second-person forms as a threshold) as L1 input to rote learn the form. However, despite a lack of input tokens, the children in the corpus use generic *he* relatively frequently when compared to their production of singular *they* (see Table 1).

% of overall tokens	Child Production	Carer Input
Singular *You*	97.79	99.92
Plural *You*	02.21	00.08
Singular *They*	03.93	01.02
Plural *They*	96.07	98.98
Masculine *He*	87.17	94.88
Generic *He*	12.83	5.12

Table1: Child output of singular *they* and generic *he*

This apparently contradictory finding in the preliminary quantitative data (where output does not reflect input) is explained by my qualitative analysis of the CHILDES data. A closer look at the concordances of generic *he* indicate that, although the references may initially look generic, it is actually impossible to distinguish whether the pronoun has been used gender-neutrally or for masculine reference. Tokens such as *a fly*, *monster*, and *my froggy* do not have lexically specified biological sex and as such natural gender, or masculine reference, cannot be presupposed to apply here. However, this does not mean that generic reference can automatically be assumed, as Wales (1996) notes that children's choice of pronouns for the personification of their toys is not straightforward. Based on the analysis of wider context, and a case study of the use of *bunny*, it is highly likely that all the apparent generic references in the data for *he* are intended to carry a masculine value.

A closer look at the concordances of potential tokens of singular *they* also indicates a problem insofar as antecedents such as *telescope* and *the shop* are not animate, and therefore the children have not applied the [+/- animate] phi-feature to pronominal antecedents. However, in terms of their L1 input, children do hear animate antecedents of singular *they*, such as *Mouse*, *Nobody*, and *Good Dog*. Furthermore, this anomaly in the data can be explained by the fact that the children in the corpus have not yet reached the age where pronoun acquisition is complete (see Owens Jr 2007) and therefore, their pronoun usage may still be fluid as the paradigm is still being formed in their internal grammars.

In addition, with the corpus data informing issues of syntactic theory, generic *he* is also

rejected on the grounds that rote learning this form would require the addition of a new phi feature [+/- generic] to the pronoun paradigm. Due to the closed-class nature of the personal pronoun paradigm, and the fact that children do not receive this form as L1 input (when problematic antecedents have been eliminated), an addition of this kind seems unlikely. However, in contrast, the acquisition/addition of singular *they* requires no such fundamental change. It only requires the application of the [+/-plural] phi feature, which is common to the second person, and indeed the whole personal pronoun paradigm, to be applied to the third-person forms that the children do receive as L1 input.

Thus, the corpus data supports the arguments that children receive plural and singular *they* as L1 input in a ratio similar to that which they receive plural and singular *you*. It also indicates that children do not receive tokens of generic *he* from their primary carers at the threshold level for rote learning pronouns. The production data also looks promising, although due to the age of the children (between 1;8 and 3;0 years old) and their limited production, only a partial analysis can be undertaken here.

References

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk (3rd edition), vol. 1: The format and programs.* Mahwah, NJ: Lawrence Erlbaum Associates.

Owens Jr, R.E. 2007. *Language development: An introduction* (7th edition). Boston: Pearson Education.

Rispoli, M. 1994. "Pronoun case overextensions and paradigm building". *Journal of Child Language* 21 (1): 157-172.

Stringer, J.L. and R. Hopper. 1998. "Generic he in conversation?" *Quarterly Journal of Speech* 84: 209-221.

Theakston, A. L., E. V. M. Lieven, J. M. Pine, and C. F. Rowland. 2001. "The role of performance limitations in the acquisition of verb-argument structure: an alternative account". *Journal of Child Language* 28: 127-152.

Wales, K. 1996. *Personal Pronouns in Present Day English*. Cambridge: Cambridge University Press.

The identification of metaphor using corpus methods: Can a re-classification of metaphoric language help our understanding of metaphor usage and comprehension?

Katie Patterson
University of Liverpool
k.j.patterson@student.liv.ac.uk

1 Introduction

As current research into metaphor moves away from the heavily theoretical confines of conceptual metaphor and the first wave of Cognitive Metaphor Theory (1980), the last decade has seen researchers follow a trend of more usage-based approaches, drawing their methods and their theories from the fields of cognitive and corpus linguistics (Semino et al., 2004; Koller, 2006; Partington, 2006). Rather than deriving examples from theory, corpus methods allow the researcher to study metaphors as they occur in everyday, real-life usage. With context and a wide recourse to the text playing a major role in the understanding of metaphorical language, a usage-based approach draws upon the social and discourse contexts in which metaphors are used (cf. Cruse, 1986).

Such developments however, have brought about a range of methodological issues, ranging from the categorisation of metaphorical language to the identification and extraction of such language from a large corpus. This paper seeks to address such issues by suggesting an alternative, heuristic approach to the classification of metaphor types. The discussion draws on current research into the theory of Lexical Priming (Hoey, 2005) and metaphor comprehension, whilst addressing the issues that in the past have questioned the validity of corpus methods in relation to metaphor.

2 Metaphor

Lexical metaphor concerns the semantic association of words within a given context. Wikberg defines it as "a way of seeing something in terms of something else, a process which involves a linguistic expression referring in an unconventional way to people, animals, things, events or concepts on the basis of some similarity, correlation, or analogy" (2006: 34). Goatly's definition of metaphor consolidates and develops upon Wikberg's, extending it to grammatical

structure also:

> an unconventional act of reference or colligation is understood on the basis of some similarity, matching or analogy involving the conventional referent or colligates of the unit and the actual unconventional referent or colligates (Goatly, 1997: 86).

Thus any entity referred to metaphorically lacks at least one critical feature possessed by the conventional referents of the word. This divergence in relationship can be defined grammatically by colligation, or semantically by collocation, semantic association, and pragmatic association.

As the term metaphor envelops such a wide span of metaphorical language ranging in terms of grammar and lexis, and from idiomatic or fossilised phrases to highly original and unique metaphorical language (Steen, 1995 and Goatly, 1997), there is a need for addressing the concept in terms of a scale of behaviour. Current research refers to a cline of 'metaphoricity' mainly in terms of grammar and lexis (Simon-Vanderbergen, *et al.* 2003). However, this paper argues that from a lexical approach, metaphorical language also operates on various levels of a cline or gradient, ranging from unconventional collocations and semantic associations, to metaphors so fully fossilised in our language, that we need not be aware of the original meaning or association to understand the metaphor.

This paper focuses upon the different behaviour in metaphor types concerning semantic relationships. The main classes of lexical metaphor will be referred to and discussed as *original*, *fossilised* and *semi-fossilised.* It is proposed that the frequency of the use of the metaphor implicates its level of primings. Thus the suggested categorisation of metaphoric language provides a gradient – the importance is placed on establishing at what point along the cline does a metaphor cease to become original and thus become partly or fully fossilised. This paper proposes that evidence of primings within a metaphorical phrase will help to establish its place along such a cline and conversely, that the frequency of a metaphorical phrase will determine the strength of it's semantic associations.

3 Lexical metaphor and priming

Part of the inherent quality of metaphors is that they override some major semantic relationship. This semantic relationship can be defined by collocation, semantic association, and pragmatic association. According to the theory of Lexical Priming (Hoey, 2005), evidence of such categories (amongst others extending to the textual and grammatical dimension of language) is part of what it means to know or understand the meaning of a word or phrase.

Primings of collocation, semantic association and pragmatic association are not permanent features of a word (or set of words). Each use we make of a word, and each new encounter, either reinforces the primings or loosens them. They may accordingly shift in the course of time and use, and subsequently the lexical item/s can shift slightly in meaning and/or function. This may be referred to as drifts in the priming (cf. Hoey, 2005), and allows for the creative use or flexibility of metaphors. In terms of semantic analysability of metaphor, collocation can be found at one end of the scale whilst the most opaque idiomatic metaphors (those which we do not need to derive the origin of in order to understand) lie at the other.

Metaphors that are used frequently enough, will eventually adopt new semantic relationships, which work to make the metaphor idiomatic or fossilised. These relationships that have been established through repetition, will ultimately determine subsequent contextual usage of the phrase.

4 A CORPUS APPROACH

Metaphorical language operates upon a continuum whereby the types of semantic relations can be strong or weak. An original metaphor works by overriding at least one expectation or priming. This could be on the level of a word or a phrase. A more fossilised metaphor is understood by using established 'second level' primings. This means that the phrase is used frequently enough to adopt new expectations or primings in its usage. Such phrases have become so well embedded into our language, that they no longer behave in the same manner as an original, one-of-a-kind, metaphor: there is a certain level of expectation and association in the mind of the reader/listener that makes the phrase recognizable (i.e. the prosody/collocation/association). Such associations create salient meanings: "coded meanings foremost on our mind due to conventionality, frequency, familiarity, or prototypicality" (Giora, 2003: 10). Original and creative metaphors operate by going against such conventions.

Thus, although both original and fossilised metaphors do not conform to certain language conventions – the original definition of a metaphor referred to by Goatly (1997) ('an

unconventional act of reference'), does not apply to fossilised, idiomatic phrases in the same manner as truly original metaphors. It is only through corpus evidence that such theory-based ideas can be validated. Concordance lines of original and commonly used metaphors will be found to show differences in the way that primings operate. These results may in turn, shed light on new approaches to identifying certain types of metaphor within a corpus.

References

Cruse, D. A. 1986. *Lexical Semantics.* Cambridge: Cambridge University Press: Cambridge.

Giora, R. 2003. *On Our Mind: Salience, Context, and Figurative Language*. Oxford: Oxford University Press.

Goatly, A. 1997. *The Language of Metaphors*. London: Routledge.

Hoey, M. 1997. 'From Concordance to Text Structure: New uses for computer corpora'. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), Proceedings from International Conference on Practical Applications in Language Corpora (pp. 3–23). Lodz: Poland, 10–14 April 1997.

Hoey, M. 2005. *Lexical Priming.* London: Routledge.

Koller, V. 2006. 'Of critical importance: Using electronic text corpora to study metaphor in business media discourse'. In Stefanowitsch, A. & Gries, S. (eds.) *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter, pp. 237-267.

Partington, A. 2006. 'Metaphors, motifs and similes across discourse types: Corpus Assisted Discourse Studies (CADS) at work'. In Stephanowitsch, A & Gries, S. (eds.) *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter, pp. 267-304.

Philip G. 2010. Why prosodies aren't always present: Insights into the idiom principle? In Mahlberg M., González-Díaz V., Smith C. (eds). *Proceedings of the Corpus Linguistics Conference CL2009.* Liverpool: University of Liverpool.

Semino, E., Heywood, J., and Short, M. 2004. 'Methodological problems in the analysis of metaphors in a corpus of conversations about cancer'. In *Journal of Pragmatics*, 33 pp. 1271-1294.

Simon-Vanderbergen, A., Taverniers, M. and Ravelli, L. 2003. *Grammatical Metaphor: Views from Systemic Functional Linguistics*. Amsterdam: John Benjamins.

Wikberg, K. 2008. The Role of Corpus Studies in Metaphor Research. In *Selected Papers from the 2006 and 2007 Stockholm Metaphor Festivals*, N.L. Johannesson and D.C. Minugh (eds) pp. 33–48. Stockholm: Department of English, Stockholm.

Stance adverbials in research writing

Matthew Peacock
City University of Hong Kong
enmatt@cityu.edu.hk

1 Introduction

This paper describes a corpus-based analysis of discipline variation in the use of stance adverbials, for example *certainly* and *generally*, and their role in the construction of epistemic stance. The corpus was 600 research articles (RAs) across 12 disciplines: Biology, Business, Chemistry, Computer Science, Economics, Environmental Science, Language and Linguistics, Law, Neuroscience, Physics and Materials Science, Psychology, and Public and Social Administration.

Biber's 2006 definition of stance adverbials is adopted – items which express attitude or assessment towards a proposition. Epistemic stance is defined as the expression of commitment to the truth of propositions (Hyland 1999) through "value judgments… assessments" (Biber 2006). In RAs epistemic stance is part of the crucial function of claiming position in the discourse community of peers.

Biber et al. (1999) divide stance adverbials into three categories:

A. Epistemic. Convey one of six areas of meaning:

(1) Doubt and Certainty (e.g. *perhaps*). Judgments of certainty/ level of probability
(2) Actuality and Reality (e.g. *actually*). The status of propositions as real life fact
(3) Source of Knowledge/Allude to Evidence (e.g. *according to*). The source of information reported in propositions
(4) Limitation (e.g. *generally*). The limitations of propositions
(5) Viewpoint or Perspective (e.g. *in our view*). The viewpoint from which propositions are true
(6) Imprecision (e.g. *kind of*) mark propositions as being imprecise

B. Attitude (e.g. *fortunately*). Convey evaluations towards propositions.

C. Style (e.g. *frankly*). Comment on the manner of conveying propositions.

Only two previous empirical studies were found. Biber et al. (1999) examined stance adverbials in academic prose in the Longman Spoken and Written English corpus. They report frequency (pmw/per million words) in three

categories: Epistemic 3600, Attitude 350, Style 100. Biber (2006) presents results from two corpora, Doubt and Certainty 1950, Attitude 150, Style 700 in textbooks, and 1300, 150, and 350 pmw in "Written course management".

Stance adverbials may play an important role in constructing epistemic stance and therefore in the key RA functions of putting forward claims and propositions. They would thereby be valuable persuasive devices in research writing including RAs. Yet very little previous research has investigated discipline variation.

2 Research method

The aims of this study were to extend previous research on the form, frequency, distribution, and function of stance adverbials across twelve disciplines. The research questions are:

(1) What stance adverbials do RA authors use, and how frequently do they use them, across a range of disciplines? Are there any interdisciplinary differences?
(2) How do stance adverbials function across a range of disciplines?

The corpus was 600 RAs published in 2000-2008, 50 from each discipline. Analysis was done in these steps using WordSmith Tools 4.0 (Scott 2004):

(1) Build a list of stance adverbials. Biber et al. (1999) list 78. 80 further stance adverbials were identified from grammars, thesauruses, and the RAs themselves.
(2) Check the frequency of all items, and disciplinary variation.
(3) Individually check the function of every occurrence of all items.
(4) Statistical significance was set at $p<.05$.
(5) Two evaluators were involved in checking function.

3 Results

Frequency results pmw by semantic category were Doubt and Certainty 825, Actuality and Reality 414, Source of Knowledge/Allude to Evidence 375, Viewpoint or Perspective 6, Imprecision 208, Limitation 1476, Attitude 141, Style 47. One category, Limitation, thus made up 42% of all stance adverbials; Doubt and Certainty plus Limitation, 66%; these two categories plus Actuality and Reality and Source of Knowledge/Allude to Evidence, 89%. The remaining four categories made up only 11%. Many statistically significant discipline differences were found. One broad difference was between non-sciences and sciences: the sciences showed significantly lower frequencies in all categories, with Chemistry showing even lower frequencies. Language and Linguistics, and Law, were both significantly higher in six categories.

Regarding frequency results for individual forms, four notable findings across all disciplines were clear. (1) Authors used a wide range of forms: 118 appear in the corpus. (2) The range of forms was very much greater in two categories, Doubt and Certainty and Limitation, than any other category. (3) Science authors used an equally wide range of forms as non-science apart from in Doubt and Certainty. (4) Just 20 forms make up around 80% of occurrences: *clearly*, *perhaps*, *probably*, *of course*, *possibly*, *certainly*, *obviously*, *most likely*, *indeed*, *actually*, *according to*, *about*, *generally*, *typically*, *usually*, *in general*, *primarily*, *mainly*, *frequently*, and *largely*. Two prominent areas of individual discipline variation were the significantly higher frequencies of a wide range of forms across several categories in Language and Linguistics, and in Law. Individual item results across all disciplines will be presented.

Regarding function, it was found that all items always functioned to construct epistemic stance, and in line with the Biber et al. functional categories.

4 Discussion and conclusions

Limitation, and to a lesser extent Doubt and Certainty, were much more prevalent and therefore more important to RA authors than hitherto suspected. We suggest that the functions expressed in these categories are of particular value to RA authors. The fact that these categories each contain a much greater variety of linguistic forms than other categories lends support to this conclusion.

Regarding individual forms, two findings seem particularly noteworthy. First, the range of forms employed by authors is wide. Second, just 20 forms making up 80% of all occurrences indicates that these are apparently preferred by authors, and the prevailing terminology employed across twelve disciplines.

Regarding the sciences using significantly fewer stance adverbials overall than the non-sciences, Hyland (2008) theorizes that different disciplines vary in argument structure and in how readers might be persuaded, and that disciplines range along a cline with hard knowledge sciences and softer humanities at opposite ends. He portrays sciences as empirical, objective, quantitative, showing linear and cumulative growth, utilizing experimental methods, not relying on rhetoric, and putting greater weight on methods, procedures and equipment: and

humanities as explicitly interpretive and qualitative, and putting greater weight on strength of argument. A closer examination of Science RAs was then undertaken. Authors developed claims in a different way, using less argument. They described their research justifications, methods, and results in a much more narrative and descriptive style, merely describing their steps and findings one by one. Presumably this is sufficient for readers, who do not need to be explicitly told the connections between facts and claims. In the present corpus, the twelve disciplines did range along a cline with sciences and non-sciences at different ends. Biology, Chemistry, Environmental Science, Neuroscience, and Physics and Materials Science authors did tend to rely less on rhetoric and put greater weight on methods, procedures and equipment, while Business, Language and Linguistics, Law, Psychology, and Public and Social Administration tended more towards interpretive and discursive argument.

Closer examination of the corpus was then undertaken to investigate the striking individual discipline differences, which are not easy to explain. Language and Linguistics authors used 46% more stance adverbials than other disciplines, with Doubt and Certainty being 72% higher, and Limitation 42%. Presumably it is correspondingly more important in Language and Linguistics to express judgments of certainty and probability regarding propositions, to signal the limitations of propositions, and to put greater weight on the strength of argument in these areas. In Law, three out of four major categories were significantly higher. Law authors relied on *perhaps* and *of course* to express Doubt and Certainty; on three terms to express Actuality and Reality; and on four terms to express Limitation. Seemingly it is correspondingly more important in Law to express all of these functions. Chemistry showed particularly low frequencies in Doubt and Certainty, and Limitation. A closer examination of Chemistry RAs found that authors do not appear to cover these functions in ways apart from the use of stance adverbials, and it was concluded that they develop arguments in a different way, almost exclusively through descriptions of their research materials, equipment, and findings. Apparently this is sufficient for Chemistry readers.

The next step was to adopt the techniques of semantic preference, the creation of meaning through multiple occurrences of collocates (Hunston 2007), to examine function more closely. Mahlberg (2003, 2007) says that meaning develops across clusters (word groups which always appear in the same order). WordSmith Tools was used to isolate the clusters and collocates associated with the top 20 stance adverbials across all twelve disciplines. Large numbers were found, and will be presented. It is suggested that these patterns represent standard terminology and also an important part of the meaning, and the function, of these stance adverbials.

Stance adverbials appear to play an important role in the construction of epistemic stance in RAs. Authors employ them to express attitudes, value judgments, and assessments towards their claims and propositions, and thereby accomplish the essential function of claiming membership of discourse communities. Additionally, sciences and non-sciences, and certain disciplines, achieve this in significantly different ways. Finally, semantic preference techniques may be a very valuable method for corpus-based research.

Implications for teaching and for further research will also be discussed.

References

Biber, D. 2006. "Stance in spoken and written university registers". *Journal of English for Academic Purposes* 5: 97-116.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Hunston, S. 2007. "Semantic prosody revisited". *International Journal of Corpus Linguistics* 12 (2): 249-268.

Hyland, K. 1999. "Disciplinary discourses: writer stance in research articles". In C. Candlin and K. Hyland (eds.) *Writing: texts, processes and practices*. London: Longman.

Hyland, K. 2008. "Genre and academic writing in the disciplines". *Language Teaching* 41 (4): 543-562.

Mahlberg, M. 2003. "The textlinguistic dimension of corpus linguistics: the support function of English general nouns and its theoretical implications". *International Journal of Corpus Linguistics* 8: 97-108.

Mahlberg, M. 2007. "Clusters, key clusters and local textual functions in Dickens". *Corpora* 2 (1): 1-31.

Scott, M. 2004. *WordSmith Tools* Version 4. Oxford: Oxford University Press.

A pragmatic analysis of imperatives in voice-overs from a corpus of British TV ads

Barry Pennock-Speck
Universitat de València
pennock@uv.es

Miguel Fuster-Márquez
Universitat de València
miguel.fuster@uv.es

1 Indirectness in TV ads

When asked what the purpose of TV commercials is, most people answer that it is to sell goods or services. Although the ultimate aim of an advertising campaign is certainly to get people to spend their money on whatever it is that the ad people are promoting, this view of the function of TV ads is, to put it mildly, rather unsophisticated. Del Saz-Rubio & Pennock-Speck (2009), Pennock-Speck & Del Saz-Rubio (2009) and Pennock-Speck & Del Saz-Rubio (in press) have shown convincingly that although most TV ads do indeed promote goods and services, they often do so very indirectly. However, the profusion of imperatives such as 'call', 'visit', 'try' in voice-overs and on-screen text might be seen to contradict their contention that the discourse of TV ads is as indirect as they contend. Because, one may argue, what could be more direct than an imperative?

In order to prove that TV ads are indirect, in spite of the wealth of imperatives in this genre, we will look at their role in voice-overs the most salient type of verbal language, which, in turn, is by far the most direct and unambiguous semiotic mode unlike other modes such as images, sounds and paralinguistic features of voice.

We will attempt to show that the illocutionary force of the imperatives in TV ads is dilluted (Byrne 1992: 66) mainly because of the specific nature of the relationship between S(peaker) and H(earer) (Brown and Levinson 1987). In the case of TV ads, the S is a combination of the voice-over actors who are mere animators in the Goffman (1981: 44) sense; the authors of the discourse, that is, the admakers; and the principle, i.e., the companies who pay for the ad to be made and aired. H refers to the viewers. In a TV ad S is in fact in no position to give orders to H as "the speaker must enjoy a superior social status relative to the listener to give a felicitous order and advertisers do not enjoy such a social advantage over viewers" (Geis, 1982: 18 –cited in Byrne 1992). Other authors hypothesize that the real function of imperatives is simply to give "a sense of one person talking to another" (Myers, 1994: 47). With regard to the types of verbs used in commands Leech (1966: 110-111) mentions those referring to a) the purchase of the product, 'get', 'buy' and 'ask for'; b) its consumption of the product, 'have', 'try', "use" and "enjoy"; and c) those that refer to the "appeal to notice", 'look', 'see', 'watch' and 'remember' (see also Versegaard & Schröder (1985). The verb 'buy' is eschewed as it has negative associations of parting with one's money.

2 Imperatives in voice-overs

There seems to be a consensus on the function of imperatives so why the need for yet another publication on the subject? The first reason is of an epistemological nature. If we look at what has been written on the imperative in advertising, the first major study of this genre, Leech (1966), is over forty years old. Others publications which mention imperative clauses, such as Verstegaard & Schröder (1985), focus on print ads while Byrne analyzes print and radio ads. Although Myers (1994) does look at TV ads he offers scant evidence for his categorical statements about their. O'Neill & Casanovas-Catala's (1997) study is one of the few to look exclusively at imperatives in both British and Catalan TV ads but was based on a very small corpus of only fifty ads with no information on how the ads were selected. Given this situation, it seems to us that a more systematic study of British TV ad discourse and imperatives in this genre was needed.

3 Corpus, methodology and discussion

Our corpus is the MATVA (multimodal analysis of TV ads), made up of a total of 1,285 ads recorded on two days in March and two days in June 2009. Of these 594 are non-duplicated and 469 feature voice-overs. The textual part of the corpus is made up of the transcription of voice-overs, on-screen text, dialogue/testimonials and song lyrics. Although massive compared to the collection of ads used in earlier studies, some of which are mentioned below, by corpus linguistics standards our corpus is relatively small. However, as it is based on coherent sampling principles (see Biber 1993), its relatively small size has its benefits for this kind of research. The most important is that there is no danger of our selecting only the data that we are interested in as we analyze quantitatively and qualitatively "all" the ads that include voice-overs in our corpus. To identify the imperatives we tagged our corpus using the UCREL CLAWS7 tagset and located all

the base forms of the verb. Once this was done, we analyzed each instance to detect which were actually imperatives.

The second reason is that our pragmatic analysis of imperatives is based on a new approach to face that, while taking into account Brown and Levinson's (1987) seminal publication on politeness, actually argues against the need for face work to be centred on the concept of politeness–or for that matter–impoliteness. We go back to Goffman's work on face (1955) but redefine it in the light of the work of among others Spencer-Oatey (2000); Eelen (2001); Watts (2003); Arundale (2006); Culpeper (2011), Pennock-Speck & Del Saz-Rubio (2912). In this way, we can account for face work that seems to constitute typical cases of positive and off-record politeness strategies but actually functions to induce a feeling of discomfort in the audience as is the case in the charity ads analyzed in Pennock-Speck & Del Saz-Rubio (2012). Our view of face is perforce psychological because a radically interactional approach (Arundale, 2006) is not tenable in the case of the TV ad genre nor, we would argue, for any other kind of interaction.

4 Preliminary conclusions

Apart from supplying data on the position of the 365 instances of imperatives within the TV ads, the type of ads they are found in, and the kind of verbs found in imperative clauses, our results show definitively that the function of imperatives in TV ads is specific to this genre and more akin to suggestions than directives given the indirect nature of most TV commercials. Moreover, we show how imperatives take into account not only H but also, more controversially, S.

References

Arundale, R. B., 2006. "Face as relational and interactional: A communication framework for research on face, facework, and politeness". *Journal of Politeness Research.* Language, Behaviour, Culture 2 (2), 193-216.

Biber, D. (1993) 'Representativeness in Corpus Design'. *Literary and Linguistic Computing* 8 (4): 243-257.

Brown, P. and Levinson, S. D. 1987. *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Byrne, B. 1992. *Relevance theory and the language of advertising.* CLS Occasional Paper No. 31.

Culpeper, J. 2011. *Impoliteness. Using language to cause offense*. Cambridge: Cambridge University Press.

del Saz-Rubio, M. M. and Pennock-Speck, B. 2009. "Constructing female identities through feminine hygiene TV commercials". *Journal of Pragmatics* 41: 2535–2556.

Eelen, G. 2001. *A critique of politeness theories.* Manchester: St Jerome.

Geis, M. L. 1982. *The language of television advertising.* New York: Academic Press.

Goffman, E. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry: Journal for the Study of Interpersonal Processes*, 18, 213-231.

Goffman, E. 1956. "The nature of deference and demeanor". American Anthropologist, 58, 473-502.

Goffman, E. 1967. *Interactional ritual: Essays on face to face behaviour.* Garden City, New York.

Goffman, E. 1969. *The presentation of self in every day life.* Doubleday: Garden City, New York.

Leech, G. 1966. *English in advertising: A linguistic study of advertising in Great Britain.* London: Longman.

Myers, G. 1994. *Words in ads.* London: Arnold.

O'Neill, M. & Casanovas Catala, M. 1997. The use of imperative in Catalan and English advertisemens: a pragmatic analysis. *Barcelona English Language and Literature Studies*, 8: 261-280

Pennock-Speck, B. and del Saz-Rubio, M. M. 2009. "Voice-overs in standardized English and Spanish television commercials." Atlantis. *Journal of the Spanish Association of Anglo-American Studies.* 31.1: 111–127.

Pennock-Speck, B. and del Saz-Rubio, M. M. (in press) "A multimodal analysis of facework strategies in a corpus of charity ads on British television". Journal of Pragmatics.

Spencer-Oatey, H. 2000. *Culturally speaking: managing rapport through talk across cultures,* London/New York, Continuum.

Watts, R. 2003. *Politeness*. Cambridge University Press: Cambridge.

A defence of semantic preference

Gill Philip
University of Macerata
g.philip.polidoro@gmail.com

1 Introduction

Less striking than collocation, less enticing than semantic prosody, it would be fair to say that semantic preference is the most neglected component of Sinclair's (1990) "extended unit of meaning". This paper intends to reignite interest in semantic preference, discussing some of the issues concerning repetition of words *vs* repetition of ideas, and making a case for the continued (or renewed) practice of consulting KWIC concordances in addition to – or indeed instead of – the more sophisticated and speedy computational tools which are available to the corpus linguist.

2 Collocation

Collocation in its received sense is the co-occurrence of two word forms at least twice in the data examined. Collocations can of course involve more than two words (e.g. "tall, dark and handsome", "single white female"), and "word form" is often extended to encompass some or all forms of the lemma ("serial killer/s", "double whisky/whiskies"). Strictly speaking, different word forms tend to collocate differently from one another (compare "naked eye" and *"naked eyes"), and when they do share collocates, the meaning expressed can be surprisingly different (compare "ruddy cheeks" and "ruddy cheek"). But when we talk of a "collocation" we implicitly include all the acceptable variants and exclude the unacceptable ones.[1]

Being by definition a visible and countable phenomenon, collocation lends itself to automation. A collocation famously "stares you in the face just as it is" (Firth 1957: 14). What this means in terms of the KWIC concordance is that collocations repeated down a page are identifiable as blocks separated by an invisible line: the white space of word boundaries. In terms of decontextualised collocations listings, collocates of a given node can be listed in descending or alphabetical order of statistical significance. What I would like to stress here is that it is a simple computational task to retrieve repeated strings of characters and determine the collocations present in text data. It is a little less simple, but still unproblematic, to retrieve variants of character strings, and therefore flesh out the detail of those collocations. Semantic groupings are another kettle of fish. It is much less simple, and decidedly problematic, to retrieve repeated ideas which may or may not be represented with repeated character strings.

3 Disclaimer

Despite having mentioned the automation of data extraction, the focus of this paper is not to investigate or provide an overview of the state of the art of semantic tagging. The very considerable progress that has been made in this area over the past decade is taken as given and I do not intend to belittle the bewildering complexity that semantic annotation entails. What I do intend to dwell upon is a phenomenon which has emerged in parallel with computational advances: the virtual disappearance of the KWIC concordance in corpus linguistics journals, book series and even at conferences such as this one. The thrust of my argument is that the increasingly sophisticated tools which the average corpus linguist has at his or her disposal are lulling linguists into a false sense of security. If collocations can be extracted automatically, it seems, there is no longer any need to count and measure the data by hand.

While it is clear that "the whole point" of using data in electronic format and running it through concordancing software is to introduce an amount of automation to the analysis, it is also true that it was not intended that the computer should be doing all the analysis. Collocations listings and profiles serve a particular purpose within particular types of language study; but they do not tell the full story and they have to be used as an *aid* to analysis, not a *substitute* for it. This is especially true when the corpus in question is not a general reference corpus but rather a collection of texts which are being analysed using corpus linguistics tools.

4 Semantic preference

Semantic preference is the stepping stone which makes it possible to progress from the concrete realities of collocation to the abstract perception of semantic prosody. Semantic prosody is undeniably more attractive a category in corpus linguistics studies: although counting hits on Google Scholar is a crude measure to use, it is interesting to compare the 1750 hits for "semantic prosody" with 1050 for "semantic preference", not just because of the numerical difference, but also because "semantic prosody" is only used

[1] How exactly we do so is something of a mystery, and beyond the scope of the present paper to discuss.

within corpus linguistics (over 70% of the hits also feature "collocation") yet is 66% more frequent than "semantic preference", despite this latter term having currency throughout the cognitive and linguistic sciences (41% of the hits also feature "collocation") and therefore being the more widely-used of the two. For those scholars whose interest lies in semantic prosody – in pragmatics, in evaluation and in connotation – sketching out the semantic preference is a means to an end rather than worth studying in its own right. Yet for collocations enthusiasts, semantic preference is put together by grouping the recurrent collocates – those extracted by the software – which inevitably means that detail is being lost.

5 Semantic preference in corpora

Why is semantic prosody worth bothering with, then? Insofar as large general reference corpora are concerned, collocation profiles may indeed suffice. But increasingly a corpus is a small collection of texts which are being subjected to corpus-assisted analysis, usually in addition to "manual" analysis; and here semantic preference becomes important. The reason is simple: the shorter the text, the lower the number of collocates extracted via statistical measures, and the lower the frequency of any collocations that are found. A lack of lexical repetition is held to be a feature of good writing. The inevitable corollary is that although word forms may not be repeated, it does not follow that certain notions are not being reiterated in the text: they are simply expressed with different words.

Even when the texts in question are not particularly short, repetition may be absent, or it may be absent at certain (potentially) crucial points in the text. This is true in literary texts, where again repetition is avoided as a matter of good style, but may also be used deliberately in order to fix concepts in the reader's mind. Taking J.K. Rowling's seven-book *Harry Potter* series as an example, the physical attributes of the characters are described in repeated formulaic chunks which undergo little if any modification over the course of the 198 chapters, e.g.:

"greasy black hair"= Severus Snape

"pale, pointed face" = Draco Malfoy

"red eyes like slits" = Voldemort

However, no narrative can survive on formulaic language alone. More subtle forms of reiteration are used to create impressions in the reader's mind, the lack of lexical repetition preventing the reader from being able to pinpoint where his or her interpretation stems from. And it is in such places in a text that semantic preference takes precedence over collocation, and the use of KWIC concordances becomes essential. The semantic preference is built up by observing and grouping single instances of words with similar meanings, or which in context appear to form a coherent group, e.g. "dishonesty".

had not answered honestly
guilty secrets
lies
The Life and Lies of Albus Dumbledore

Figure 1: HP7, Ch2 "dishonesty"

Once the semantic preference identification procedure gets under way, it becomes apparent that determining similarity is not always straightforward. Bottom-up semantic groupings are rather more complex than top-down ones, and can be unpredictable. In some cases, formal semantics prevails, in others, there is sufficient sharing of attributes for group membership to be considered, (c.f Hanksand Ježek 2008). Sometimes meaning distinctions merge. In Figure 2, "fall" literally refers to Dumbledore's plummet from a high tower, but simultaneously refers to his death (it is a euphemism for death, but also a metonym in this context).

moments after Dumbledore had fallen
moments after Dumbledore fell, jumped, or was pushed
right after Dumbledore had died,
R- right after Dumbledore ...
you said after Dumbledore ´s funeral
four weeks after Dumbledore ´s mysterious death

Figure 2: HP7 "after Dumbledore's death"

6 Who's afraid of the KWIC concordance?

A call to re-evaluate semantic preference necessarily involves a call to resuscitate the KWIC concordance as an essential and fundamental part of corpus data analysis. (Re)turning to KWIC concordances compels the analyst to reconnect with the original text(s) in the corpus, to engage with " text" as well as "data", and to remember that linguistics is not about data extraction, but about how language works.

References

Firth, J.R. 1957. "A Synopsis of Linguistic Theory 1930-55". *Studies in Linguistic Analysis*. Oxford: Basil Blackwell. 1-32.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge (Mass.): MIT Press.

Hoey, M. 2005. *Lexical Priming: a new theory of words and language*. London and New York: Routledge.

Hanks, P. and Ježek, E. (2008). "Shimmering lexical sets". *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra. 391-402.

Sinclair, J.M. 1996. "The search for units of meaning". *Textus* 9: 75-106.

Automated semantic categorisation of collocates to identify salient domains: A corpus-based critical discourse analysis of naming strategies for people with HIV/AIDS

Amanda Potts
Lancaster University
a.potts@lancaster.ac.uk

1 Introduction

In modern corpus linguistics, bigger is often better. We now have access to reference corpora containing billions of words, and individual researchers routinely collect *ad hoc* corpora of millions or hundreds of millions of words for specific purposes. This technological advancement is a blessing and a curse; while larger corpora contain more examples of both frequent and infrequent patterns to study, the sheer volume of results is often prohibitive to detailed qualitative analysis. This issue is of particular significance to analysts who are interested in combining the power of corpus linguistic tools with the rich scholarly tradition and interdisciplinary flexibility of additional theories emphasising qualitative analysis.

I here demonstrate one method of exploring the representation of social actors in a large opportunistic corpus. Using a corpus-based critical discourse analytical approach with a strong focus on automated semantic tagging of collocates, I compare construal of *AIDS/HIV patients*, *victims*, *sufferers*, and *carriers* in a 161-million-word corpus of American newspaper texts from 1981-2009.

2 Theoretical frameworks

In recent years, several 'schools' of research have found that combining methodological elements of corpus linguistics with a discourse analytical theory "helps researchers cope with large amounts of textual data, thus bolstering…empirical foundations, reducing researchers' bias and enhancing the credibility of analyses" (Mautner 2009: 138), and this synergy is at the centre of a rapidly developing field of research.

A major strength of the corpus linguistics approach to discourse analysis is increased variety and representativeness owing to large but governable sample size. Using statistical measures, Mautner suggests that linguistic 'norms' can be accurately represented, and

deviations from these norms can be more objectively identified (1995: 22). Drawing upon a large enough corpus, this degree of accrued certainty also has implications for the study of "semantic prosody" (or an attitude associated with a lexical item over time and across texts) – an area of interest to discourse analysts previously immeasurable with any statistical significance (Louw 1993).

3 Description of data

Analysis is based upon a 161-million-word opportunistic corpus of texts from 27 major metropolitan and national American broadsheet print newspapers. The corpus was custom-collected using *Factiva* – an online news aggregator database – and contains all articles from 01/01/1981 to 31/12/2009 containing at least one of the following (case-specific) search terms:

1981: *gay* OR *homosexual* OR *opportunistic* OR *cancer* OR *plague* OR *Kaposi's*
1982: *gay* OR *homosexual* OR *opportunistic* OR *cancer* OR *plague* OR *Kaposi's* OR *G.R.I.D.* OR *GRID* OR *A.I.D.S.* OR *AIDS* OR *H.I.V.* OR *HIV*
1983 onwards: *A.I.D.S.* OR *AIDS* OR *H.I.V.* OR *HIV*

The shift in search terms has been designed to compensate for the variety of associated terms and alternative naming strategies evident in the early stages of the epidemic. All returned texts from 1981 and 1982 were manually reviewed for appropriateness as an additional safeguard.

4 Methodology

The method of analysis is taken in four stages: 1) identification of search terms associated with a social group; 2) calculation of collocates these terms; 3) disambiguation of semantic tags and assessment of major semantic categories represented, and finally, qualitative exploration of themes identified as 'salient' in association with the social group under analysis.

The corpus has been indexed into CQPweb (Hardie 2013), a powerful web-based analysis system. CQPweb employs the UCREL CLAWS and USAS taggers to assign part-of-speech and semantic tags to texts, and allows users to query corpora nearly-instantaneously by word, tag, or combination string.

To obtain a variety of high-frequency naming strategies to form the basis of comparative analysis, I have performed a compound search of the corpus for two possible phraseologies:

AIDS/A.I.D.S./HIV/H.I.V/GRID/G.R.I.D. + (any plural common noun)
(any plural common noun) + *of* + *AIDS/A.I.D.S./HIV/H.I.V./GRID/G.R.I.D.*

A frequency list of the plural common nouns in the positions above was generated, and searched for 'human' naming strategies. Once the most frequent of these (*patients*, *victims*, *sufferers*, and *carriers*) were identified, collocates were calculated across both possible phraseologies in CQPweb. The resultant list was output to Wmatrix, another 'fourth generation' web-based analysis system developed at Lancaster University (Rayson 2008) which also utilizes the UCREL USAS system, but additionally allows users to view vertically formatted wordlists containing all candidate semantic tags. I have used these for disambiguation and calculation of proportional preferences of semantic categories as assigned by the USAS tagger, as opposed to considering single collocates or *ad hoc* categories, which could introduce skew or subjectivity to the analysis.

5 Results

Viewing naming strategies through the scope of collocates categorised into proportional semantic categories allows me to identify salient *themes* (rather than words), and narrow the scope of analysis in an empirical, scientific manner. As this process is computer-assisted, it works very quickly and codes quite objectively. The overview results (see Table 1) lend themselves nicely to further qualitative analysis by pinpointing areas of

USAS Semantic Category	patient(s)	victim(s)	sufferer(s)	carrier(s)
A General and abstract terms	12.5%	6.3%	6.7%	14.3%
B The body and the individual	33.3%	16.7%	6.7%	14.3%
G Government and public	2.1%	8.3%	13.3%	7.1%
H Architecture, housing and the home	4.2%	6.3%	13.3%	0.0%
S Social actions, states and processes	10.4%	27.1%	26.7%	21.4%
X Psychological actions, states & processes	2.1%	8.3%	0.0%	28.6%

Table 1: Collocates of naming strategies, categorised into USAS semantic categories, and expressed as a percentage of the overall number of collocates for a given naming strategy. *Note:* This table is abridged; categories not containing >10% of collocates for any given naming strategy are not shown.

similarity and contrast between the naming strategies, particularly when focussing on categories over a certain cut-off point (here: those containing >10% of the total collocates of a given naming strategy are considered most significant).

For instance, *patient(s)* and *carrier(s)* both show a high proportional preferences for collocates belonging to the USAS A Semantic Category: General and abstract terms. In these cases, this indicates an association with these naming strategies with themes of exclusion: being *isolated, shunned, or quarantined.* In contrast, *victim(s)* and *sufferer(s)* have quite a low proportion of collocates of this type. The naming strategy with the strongest collocate preference for the USAS B Semantic Category is *patient(s)*, which speaks to this naming strategy's underlying meaning; what is interesting is that *sufferer(s)* – arguably another 'body' term – has very few B collocates, and instead has high proportions of collocates in Categories G (e.g. *discrimination*) and H (*homeless*), with over one-quarter of all collocates in Category S (*advocate, homosexual*). This indicates that *sufferer(s)* are construed less through their physical, medical experiences, and more through the lens of their interactions in society.

Indeed, Category S: Social actions, states and processes, contains the highest proportion of collocates for both *victim(s)* and *sufferer(s)*, and accounts for over 10% of all collocates for each of the four naming strategies under investigation, making this the most salient category for analysis. Themes such as care and advocacy (*visit, counsel*), commemoration (*memorialize, dedicate*), sexuality (*homosexual, casual [sex]*), and groups/belonging (*fellow, embrace*) feature here.

One final category represents over 10% of the collocates of a naming strategy. X: Psychological actions, states & processes, is favoured by collocates of *carrier(s)* but does not feature prominently in other naming strategies. Having a high proportion of collocates within this category (e.g. *identify, known, suspected*) implicates a negative semantic prosody of detecting or 'discovering' *HIV carriers* – often against their wills – not present in the other naming strategies.

6 Discussion

In the discussion portion of this paper, I summarise the main points of similarity and difference between the naming strategies explored. I also evaluate the method that I have developed and used herein, and make recommendations on its adoption by other corpus-based (critical) discourse analysts. The inclusion of an extra step to manually disambiguate candidate strings of semantic tags is acknowledged as a limitation, both for the extra time needed, and the subjective nature of selecting the most 'appropriate' tags in special cases.

Overall, I evaluate this method as a useful technique of quickly and (essentially) impartially gaining a broad overview of the most salient semantic categories associated with words, terms, or here – naming strategies. In this instance, the method allows for statistically-driven comparative analysis, and could be conducive to detailed and targeted qualitative investigation in a variety of other contexts.

References

Hardie, A. (2013). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*:3, 380-409.

Louw, W. E. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-76). Amsterdam: John Benjamins.

Mautner, G. (1995). *'Only connect.' Critical Discourse Analysis and corpus linguistics*. Paper presented at the UCREL Technical Papers.

Mautner, G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society, 36*, 51-70.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4 pp. 519-549.

Linking qualitative and quantitative analysis of metaphor in end-of-life care

Paul Rayson
Lancaster University
p.rayson@lancaster.ac.uk

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

Veronika Koller
Lancaster University
v.koller@lancaster.ac.uk

Sheila Payne
Lancaster University
s.payne@lancaster.ac.uk

Elena Semino
Lancaster University
e.semino@lancaster.ac.uk

Zsófia Demjén
Open University
zsofiademjen@gmail.com

Matt Gee
Birmingham City University
matt.gee@bcu.ac.uk

Andrew Kehoe
Birmingham City University
andrew.kehoe@bcu.ac.uk

1 Introduction and background

Metaphor in End of Life Care is an ESRC funded project aimed at investigating how metaphor is used to communicate about the experience of end of life care in the UK. The project team have collected a data set of 1.5 million words from semi-structured interviews (300,000 words) and online forum contributions (1.2 million words) by patients, unpaid family carers and senior healthcare professionals in end of life and palliative care in the UK. This paper explores the challenge of exploiting innovative corpus linguistic and natural language processing methodologies alongside manual techniques for the investigation of metaphors in large-scale data sets.

The care and support needs of people approaching the end of life have relevance for every member of society. This has recently been officially recognised by healthcare policy makers and clinicians in the UK in the form of the End of Life Care Strategy (2008) among others. The way in which the experience of end-of-life care is talked about can shed light on people's views, needs, challenges, and emotions, as well as identify areas with a potential for increased anxiety and/or misunderstanding.

This project combines a focus on a socially pertinent issue, with developments in the application of corpus methods to metaphor analysis. In the project as a whole, we are interested in finding out what the use of metaphor by the three stakeholder groups (patients, unpaid family carers and healthcare professionals) suggests about (a) the experiences and needs of the members of these groups and their mutual relationships, and (b) the nature of metaphor as a linguistic and cognitive phenomenon.

Previous techniques for the linguistic analysis of metaphor have been largely manual and qualitative. However, in recent years, the need for quantitative corpus methods in the investigation of metaphorical patterns has been recognized by a number of researchers (e.g. Charteris-Black 2004, Koller 2004, Deignan 2005, Semino 2005, Stefanowitsch and Gries 2006). Existing corpus studies have relied on a combination of manual analysis with the concordancing of selected metaphorical expressions in electronic data (e.g. Skorczynska and Deignan 2006). However, this methodology only allows researchers to find further instances of previously identified expressions. The exploitation of larger corpora for metaphor research is constrained by the fact that the identification of metaphorical expressions in texts has not yet been successfully automated, despite some promising attempts (e.g. Pasanek and Sculley 2008, Berber Sardinha 2010). Existing approaches in the computational/NLP community tend to focus on specific domains or topics.

In order to overcome these challenges, the Metaphor in End of Life Care project employs an adapted version of the UCREL Semantic Analysis System (USAS) tagger [1] (Rayson et al. 2004) embedded within the Wmatrix software[2] (Rayson, 2008) developed at Lancaster University. This tool enables us to identify potential metaphorical expressions in large data sets without being restricted to pre-established lists of lexical items. In addition, an adaptation in the search function (known as broad-sweep searching) allows us to consider all (and not just the most likely) of the semantic fields that a lexical item may be a part of, thereby providing a much more comprehensive set of potentially metaphorical expressions.

2 Linking qualitative and quantitative methods

In this paper we focus on the combination of manual intensive qualitative analysis and quantitative semi-automatic corpus methods for identifying metaphorical patterns associated with end of life care in our sample dataset of 90,000 words. This sample analysis is aimed at

[1] http://ucrel.lancs.ac.uk/usas/
[2] http://ucrel.lancs.ac.uk/wmatrix/

identifying important semantic domains, which will eventually inform the larger analysis of 1.5 million words. Metaphors relevant to end of life care are identified through a well-established analytical method (Pragglejaz Group 2007). Lexical items deemed to be metaphorical are then allocated to broader semantic domains using a data-driven approach. These semantic domains are matched with established USAS categories (Rayson et al. 2004), embedded in Wmatrix in order to enable the exploration of the larger corpus.

For the manual analysis, we have used an existing tool, eMargin (developed at Birmingham City University). [1] eMargin is an online collaborative annotation tool in which users can highlight, colour-code, write notes and assign tags to individual words or passages of a text. The texts and their annotations can be shared amongst groups, facilitating discussions and allowing analyses and interpretations to be combined.

The manual analysis phase of the project involved collaboration between three analysts on the annotation of metaphors in the interview texts. eMargin proved particularly useful for this since it allows analysts to view one another's annotations, make comments and reach majority decisions on tricky cases. eMargin was designed initially for the collaborative close reading of literary texts so does not yet have the ability to read-in semantically- or grammatically-tagged texts, or to produce frequency lists, concordances, collocations and n-grams. eMargin does, however, have an XML-export option, which has allowed us to export the manually annotated data from the tool and load it into Wmatrix for further automatic analysis and retrieval.

In previous research, we have considered the interoperability of tools such as CQPweb, Wmatrix, Intellitext and WordTree along with eMargin (Rayson et al., 2012) in order to more directly link complementary functionality between different tools, but for the moment, we are transferring data from one system to the other.

In this presentation, we will exemplify the linking of qualitative and quantitative analysis by looking at the metaphorical patterns frequently used in the corpus. We present preliminary results in the form of similarities and differences in the semantic/source domains between the three stakeholder groups and how these were analysed manually and what USAS semantic fields they correspond to.

[1] http://emargin.bcu.ac.uk/

Acknowledgements

This research is funded by the UK's Economic and Social Research Council (grant no.: ES/J007927/1). eMargin development has been supported by two JISC grants.

References

Berber Sardinha, T. 2010. "A program for finding metaphor candidates in corpora". The Especialist (PUCSP) 31, 49-68.

Charteris-Black, J. 2004. Corpus Approaches to Critical Metaphor Analysis. Basingstoke: Palgrave.

Czechmeister, C.A. 1994. "Metaphor in illness and nursing: a two-edged sword. A discussion of the social use of metaphor in everyday language, and implications of nursing and nursing education". Journal of Advanced Nursing. Vo. 19, pp.1226-1233

Deignan, A. 2005. Metaphor and Corpus Linguistics. Amsterdam: Benjamins.

Pasanek, B. and Sculley, D. 2008. "Mining millions of metaphors". Literary and Linguistic Computing 23(3), 345-360.

Pragglejaz Group 2007. "MIP: A method for identifying metaphorically used words in discourse". Metaphor and Symbol 22(1), 1-99.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. 2004. The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12

Rayson, P. 2008. "From key words to key semantic domains". International Journal of Corpus Linguistics 13(4), 519-549.

Rayson, P., Sharoff, S., Nesi, H. and Moreton, E. 2012. Increasing Interoperability between Corpus Tools. 33rd ICAME Conference, Katholieke Universiteit Leuven, May 30 – June 3, 2012.

Reisfield, G.M., and Wilson, G.R. 2004. "Use of Metaphor in the Discourse on Cancer". Journal of Clinical Oncology. Vol.22, Iss.19, p4024

Semino, E. 2005. "The metaphorical construction of complex domains: The case of speech activity in English". Metaphor and Symbol, 20(1): 35-69.

Skorczynska, H. and Deignan, A. 2006. "Readership and purpose in the choice of economics metaphors". Metaphor and Symbol. 21(2): 87-104.

Sontag, S. 1991 [1979, 1989]. Illness as metaphor and AIDS and its metaphors. Penguin

Stefanowitsch, A. and Gries, S. (eds) 2006. Corpus-based Approaches to Metaphor and Metonymy. Berlin: Mouton de Gruyter.

Investigating orality in speech, writing, and in between

Ines Rehbein
Potsdam University
irehbein@uni-potsdam.de

Josef Ruppenhofer
Hildesheim University
ruppenho@uni-hildesheim.de

1 Introduction

This paper contributes to the ongoing discussion whether data from newly emerging communication systems in social media like Twitter or Facebook should be considered a type of oral language or as predominantly written. Following the model of Koch and Oesterreicher (1985), computer-mediated communication (CMC) can be described as *conceptually* oral but *medially* written. Ong (1996), on the other hand, defines language data from social media as *secondary literal*.

In Ong (1988)'s model the main properties of oral language are described as additive, aggregative, redundant and ephemeral, while the permanent, organised and mediated nature of written language is highlighted as the main characteristic of literacy. Taking this as our point of departure, we expect that these properties are expressed on a linguistic level in the data and can be quantified in a corpus.

In this paper, we investigate this claim for German and search for markers of orality in corpora of spoken language, written language and social media data. We present a corpus study quantifying the use of different markers on the lexical and structural level and discuss the differences between the registers. Our results give new insights into the nature of written language, spoken language, and of newly emerging registers in between.

2 Markers of orality

Many studies have sought to investigate and quantify differences between written and spoken registers. Biber and Conrad (2009) define the following features as markers for orality in English: adjacency pairs, repetition, self repair, contractions, questions, modal and semi-modal verbs, *that*-complement clauses with *that* omitted, sentence relative clauses, *wh*-complement clauses and finite adverbial clauses. Other work investigates syntactic complexity, with low complexity assumed to be characteristic of spoken communication. A number of measures have been proposed in the literature for assessing syntactic complexity. Examples are the mean length of utterance (MLU), typically measured in morphemes (Miller and Chapman, 1981), the number of node counts, word counts, or high complexity indicators such as subordinating conjunctions or *wh*-pronouns (Szmrecsányi, 2004). Other approaches measure the distance between dependent words (Lin, 1996; Gibson, 1998) or the POS tag sequence cross-entropy (Roark et al. 2007).

In our study, we try to identify markers of orality on different levels. We use features from the literature as well as new features. On the lexical level, we look at filled pauses (*uh, uhm*), repetitions, backchannel signals (*mmh, okay*), question tags and exclamatives. On the syntactic level, we consider (non-canonical) word order, repetitions, and low syntactic complexity, applying some of the measures described above.

There are, however, problems. The MLU, for example, requires one to define the unit of analysis across corpora from different registers, which is not a trivial task. For spoken language, the lack of punctuation makes it hard to decide on sentence boundaries, especially because there are many utterances with no finite verb, so-called non-sentential utterances (NSU) (see, e.g., Fernández and Ginzburg 2002 for a corpus study on NSUs). In social media data, the creative use of punctuation also makes it hard to define the unit of analysis. Issues like this have to be taken into account in order to avoid a biased comparison across corpora.

3 Data

The data we use in our investigation is the Potsdam Twitter Corpus (PoTwiC) (Rehbein, submitted). The corpus includes German Twitter messages from different regions of Germany (as raw text) and is augmented with an (automatically created) layer of normalisation and part-of-speech (POS) tags. At the moment, the corpus contains 7.311.960 tweets with 105.074.399 tokens (and growing). A subpart of the corpus has been annotated manually to provide a benchmark data set for normalisation and POS taggging experiments. The corpus will be made publically available.[1]

For spoken language, we use the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000) and the KiDKo (Wiese et

[1] Due to legal issues, the original tweets can not be distributed. We will therefore hand out the tweet ids together with scripts to retrieve and process the data so that the corpus can be rebuilt.

al., 2012). The TüBa-D/S is a syntactically annotated corpus based on spontaneous dialogues between native speakers of German role-playing business partners. The topic of conversation in the data is restricted to scheduling. All the dialogues have been transcribed and annotated manually. They include around 38,000 sentences (360,000 token). The annotation provides phrase structure trees enriched with grammatical function labels (dependency relations). The Kiezdeutsch-Korpus (KiDKo) contains spontaneous peer-group dialogues by adolescents from multiethnic Berlin-Kreuzberg (around 48 hours of recording) and a supplementary corpus with adolescent speakers from monoethnic Berlin-Hellersdorf (around 18 hours of recording). The linguistic annotation of the corpus is still under construction.

For written language, we use the Huge German Corpus (HGC) (Fitschen, 2004), an automatically processed corpus of German newspaper text with syntactic annotation, including more that 200 million tokens.

As neither the PoTwiC nor the KiDKo corpus provide linguistic annotations reliable enough for performing our study, we use samples of the data in which we manually check the phenomena of interest.

4 Some results

Due to space limitations we cannot present all our results in the abstract. Instead, we focus on some of our findings which will give the reader a taste of what to expect.

On the lexical level, the most surprising finding is the existence of filled pauses in the Twitter Corpus. The current version of the corpus contains more than 10.000 occurrences of *äh* and *ähm*, which is rather unexpected as filled pauses are often considered either as a means of holding-the-floor while planning the next utterance or as markers of repair. Both should not occur in a written register and thus might suggest that the Twitter data is, at least to some extent, oral. Our study shows that filled pauses occur in both corpora as markers of hesitations, corrections, repetitions and unfinished utterances. However, on a more fine-grained level of analysis one can detect crucial differences between the filled pauses in both registers, the ones in the Twitter data being highly edited and often used with humorous intention, which puts their function as markers of orality in dispute. Unfortunately, filled pauses have been removed from the TüBa-D/S so we cannot compare the frequency of filled pauses in the two corpora (Table 1). Not unexpectedly, the HGC displays the smallest number of fillers which mostly occur in reported speech.

Next, we move on to the syntactic level and investigate differences in word order. We look at *weil* (because) which, in standard written text, introduces a subordinate clause with verb-finite word order. In spoken language, however, *weil* is also used for coordinating two matrix clauses with verb-second word order. We search for the pattern *weil Pronoun* V_{finite}. While being rare in all corpora (Table 1), these constructions which are typical for spoken language also occur in the Twitter data, but with a lower frequency. In the HGC, we found the absolute number of 28 instances of this pattern, again most of them in reported speech.

Feature	PoTwiC	KiDKo	TüBa-D/S	HGC
äh/ähm	*1.007	*54.711	N.A.	*0.017
weil P V_{fin}	*0.031	*3.644	*1.798	*0.001
subj.drop	**44	**15	‡2	**9
obj. drop	**5	**9	‡5	**0

Table1: Results for fillers, *weil* with V2 and topic-drop (normalised frequencies per *10.000 tokens/**100 instances/‡whole corpus)

Topic-drop constructions are distributed very differently across the corpora. In the written HGC corpus we observed 9 cases of subject drop and 0 instances of object drop in a random sample of 100 verb-initial clauses. In the Twitter data, we observed many more occurrences of topic-drop. But the Twitter data also seems to differ from the spoken data in the Tüba-D/S with respect to the distribution of topic-drop types. While we counted 44 cases of subject-drop and 5 of object-drop in a random sample of 100 verb-initial clauses from Twitter, we found 5 cases of object drop and only 2 of subject-drop in all of the Tüba-D/S. These last two proportions are significantly different according to a Fisher's exact test. In the KiDKo corpus, subject and object drop were found 15 and 9 times, respectively, in a random sample of 100 verb-initial constructions. Based on Fisher exact tests, the Kidko results cannot be distinguished from those for Tüba-D/S, while being significantly different from the Twitter-results. Overall, the results suggest that Twitter is somewhere in between the written data and the spoken data when it comes to the distribution of subject and object topic-drop constructions. Notably, while Twitter does exhibit more object-drop than written language, it does have less of it than true spoken language.

5 Conclusions

We present a quantitative investigation of markers

of orality in corpora from a spoken and written register, asking whether (and to what extent) microblogging data should be considered conceptually oral language. Our main findings suggest that social media data exhibits many features of oral language. However, as shown by the case of filled pauses, the relevant features are not always used for the same reasons as in actual spoken language. Finally, we suggest that social media data may be able to supply supplementary data on many spoken language phenomena that cannot be adequately studied in the relatively small corpora of spoken language that are publically available.

References

Biber, D. and Conrad, S. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Fernández, R. and Ginzburg, J. 2002. Non-Sentential Utterances: A Corpus Study. *Traitement Automatique des Languages: Dialogue*, 43 (2):13-42.

Fitschen, A. 2004. *Ein computerlinguistisches Lexikon als komplexes System*. PhD thesis. Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.

Gibson, E. 1998. Linguistic Complexity: Locality of Syntactic Dependencies. *Cognition*, 68 (1):1-76.

Koch, P. and Oesterreicher, W. 1985. *Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte*. In: Romanistisches Jahrbuch 36/85, 15-43.

Lin, D. 1996. On the Structural Complexity of Natural Language Sentences. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 1996)*, 729-733.

Miller, J.F. and Chapman, R.S. 1981. The relation between age and mean length of utternace in morphemes. *Journal of Speech and Hearing Research*, 24:154-161.

Ong, W. 1996. Interview in: *Composition FORUM* 7.2: 65-86.

Ong, W. 1988. *Orality and Literacy: The Technologizing of the Word*. New Accents. Ed. T. Hawkes. New York: Methuen.

Rehbein, I. (under review). PoTwiC – the Potsdam Twitter Corpus. Submitted to the special volume on "Non-Standard Data Sources in Corpus-Based Research" 5th edition of ZSM-Studien.

Roark, B., Hosom, J.P., Mitchell, M. and Kaye, J.A. 2007. Automatically derived spoken language markers for detecting mild cognitive impairment. In *Proceedings of the 2nd International Conference on Technology and Aging (ICTA), 1-8*.

Stegmann, R., Telljohann, H., and Hinrichs, E.W. 2000. *Stylebook for the German Treebank in VERBMOBIL*. Technical Report 239. Verbmobil.

Szmrecsányi, B. 2004. On operationalizing syntactic complexity. In: Purnelle, G.; Fairon, C., Dister, A. (eds.), *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*. Vol. 2. Louvain-la-Neuve, Presses universitaires de Louvain, 1032-1039.

Wiese, H., Freywald, U., Schalowski, S. and Mayr, K. 2012. Das Kiez-Deutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. In: *Deutsche Sprache* 40. 97-123.

It is surprising: do participial adjectives after copular verbs form a special evaluative construction?

Olga Richterová
The Institute of Czech National Corpus,
Charles University in Prague
richterova.olga@gmail.com

1 Introduction

When a word form changes its POS category, it signifies a certain meaning change and presents a phenomenon worth looking into. Such is the case of participial adjectives (PAs) forms derived from verbs with the help of a characteristic ending (*-ící/oucí* in Czech, *-end* in German, *-ing* in English)[1].

But how does one access meaning through monolingual corpora? For determining the semantics of individual investigated participial adjectives, this paper uses a thorough analysis of their collocation profiles.

As a part of a larger research project on participial adjectives in the corpora of synchronic written Czech[2] (Richterová 2012), and German[3], (Richterová 2011), one syntactic position, namely the position following a copular verb (in Czech, only "to be" is unanimously classified as a copula), was recognized as the position that attracts a special semantic group of participial adjectives. In a special construction[4], these adjectives seem to realize the language function of evaluation[5]. Even some adjectives perceived as non-evaluative at first glance (*fungující – functioning, working*) were later proven evaluative by the analysis of their collocation profiles.

Furthermore, preliminary research has shown that a similar construction can be found in a language typologically different from Czech – in German. For English, the tendency still needs to be investigated along with the differences in the usage of this construction in various genres and text types. However, the first data from the BNC querying for the construction *It is* + any adjective ending in *-ing* followed by a punctuation mark, look promising. All the adjectives express evaluation:

-ing forms in BNC	Fq	i.p.m.
interesting	19	0.17
surprising	14	0.13
misleading	6	0.05
disgusting	6	0.05
tempting	5	0.04
striking	5	0.04
frightening	4	0.04
exciting	4	0.04
embarassing	4	0.04
disappointing	4	0.04

Table 1: *It is* + English PAs before punctuation

2 Types of participial adjectives in language in general

The first part of the paper describes the general use of participial forms referred to as participial adjectives (PAs). From the perspective of syntax, though, they can be used as adjectives (modifiers): *následující den* (*der folgende Tag, the following day*); nouns (heads): *cestující* (*der Reisende, a traveller*); or they can remain close to verbal use: *hrající si děti* (*die spielenden Kinder, the playing children*).

After the usage of PAs in spoken and written language is understood, it is possible to identify constructions of special significance – such as the copular one.

The query for lemmas ending in *–ící/oucí* gave 350 126 hits in the SYN2010 corpus with the instance-per-million figure being 2 877.

-ící/oucí forms in SYN2010	Fq	i.p.m.
vedoucí (the leader)	14976	123
následující (following)	13290	109
budoucí (future – adj.)	8225	67.6
cestující (traveller)	5457	44.8
vynikající (outstanding)	5437	44.7
rozhodující (decisive)	4993	41
rostoucí (rising, growing)	4731	38.9
odpovídající (respective, answering)	4244	34.9
stávající (current)	4180	34.4
žijící (living)	3714	30.5

Table 2: 10 most frequent PAs in written Czech

Among these participial forms, some are syntactic nouns, and some are adjectives derived from verbs of movement which synchronically

[1] For more details on the diachronic development of Czech participial adjectives, see e.g. Fried 2008.
[2] Corpus SYN2010, 100 million words
[3] Czech-German InterCorp corpus, about 12 million words
[4] The term construction is understood in line with the definition found in Fried and Östman (2004): "Grammatical constructions are symbolic signs and represent the basic building blocks of a linguistic analysis. (...) A construction is an *abstract*, representational entity, (...) the actually occurring linguistic expressions, such as sentences and phrases, are not constructions, but constructs"
[5] Evaluation: "the subjective presence of writers/speakers in texts as they adopt stances towards both the material they present and those with whom they communicate," (Martin, J. R. and White P. R. R. 2005)

describe temporal development or physical movement. Only two forms seem to explicitly express the attitude of the speaker (*outstanding, decisive*).

When an analysis of the collocation profiles is conducted (with the help of a proximity-based P-collocations tool, Cvrček, V. and Vondřička, P., forthcoming), the word *odpovídající* (literally *answering*) is found to also be used in an evaluative way (unlike the above-mentioned *fungující,* it still retains a certain amount of its original verbal character). Such an analysis is carried out for the 80 most frequent participial adjectives, with the aim of assigning semantic categories to them.

Most PAs express temporal meaning, some keep their original, concrete meaning derived from the verb (including those used as syntactic nouns), others denote existentiality (*living, standing, ...*) or relation (*týkající se, související – concerning*).

All in all, only 10 PAs out of the first 80 expressions enable the speaker to express his or her subjective view and evaluate, thus representing 1/8 of the investigated forms. As we will see, the situation will be rather different for the copular position.

3 PAs in the copular position in particular

The hypothesis stating that the semantics of *–ící/oucí* PAs in copulas differs from the semantics of PAs in other syntactic positions can only be proven for copular constructs identifiable via the corpus queries used.

In copular (linking) verbs, "a subject element or a predication adjunct (...) cannot be dropped without changing the meaning of the verb" (Quirk, R. et al. 1990). Due to an awareness of the fact that the elements forming the construct can occur in various syntactic positions (especially in Czech), a great number of options need to be looked into.

For cases when the copula immediately precedes the PA within the same sentence and a punctuation mark follows, highly interesting results are obtained at first. The query delivered 2 737 instances (22.5 i.p.m.) in SYN2010. Out of the first 50 PAs, only 10 can occur in non-evaluative function. Therefore 4/5 of the investigated forms in the analyzed construction are evaluative.

Following an analysis of a number of frequent constructions containing a copula and a PA, the results concerning Czech, presented in the next section, were retrieved.

4 Preliminary results

In order to meet the aim of identifying whole constructions which trigger or make use of the evaluative function realized by a participial construction following a copular verb, it is necessary to look into possible confounding factors, such as the tense, number and genus of the copula or the subject to which the predication refers and identify whether they play a significant role.

The candidate for an evaluative construction identified so far looks as follows: [a vague and general subject: reference to prior discourse, whole context or situation] [3rd-person singular pronoun at sentence-initial position] [3rd-person singular copular verb in present tense] [an evaluative PA]. Typical constructs found in the corpus SYN2010 are:

- To je *rozhodující* / *alarmující* / *osvěžující ...*
- It is *decisive* / *alarming* / *refreshing...*

A slightly different use is found in clauses omitting the pronominal substitute altogether and followed by another subordinate clause:

- Je *překvapující, jak...*
- It is *surprising, how...*

5 Further research

Given the importance of the evaluative function in human communication, it is crucial to address the very same topic from a contrastive perspective. German and English occurrences of PAs after copulas will be investigated and their level of evaluativeness compared to that identified in Czech.

At the same time, it is necessary to pursue further analyses of Czech and to take into account the differences across genres as well as across spoken and written language.

All these issues will be addressed. In addition, a comparison with regular adjectives in copular position provides an avenue for further research.

6 Summary

The above research into participial adjectives in Czech is based on corpus data, thus enabling a rich collocational analysis. The proposed semantic categorization of the data could be verified by an extremely insightful investigation of collocation profiles.

The hypothesis that there is a substantial semantic difference between copularly used PAs and those used as premodifiers or nouns could not be disproved.

Furthermore, a clear tendency of a specific construction to express the evaluative function was identified. *Isn't that fascinating?*

References

Corpus SYN2010. 2010: *Czech National Corpus – SYN2010*. Institute of the Czech National Corpus, Praha. Available online at http://www.korpus.cz.

Cvrček, V. and Vondřička, P. forthcoming. *P-collocations*. Available online at http://korpus.cz.

Fried, M. and Östman, J. 2004. *Construction Grammar in a Cross-Language Perspective.* Amsterdam / Philadelphia: John Benjamins.

Fried, M. 2008. Constructions and constructs: mapping a shift between predication and attribution. In A. Bergs and G. Diewald (eds.), *Constructions and language change*. Berlin: Mouton de Gruyter. Available online: http://ling.ff.cuni.cz/lingvistika/fried/download/Madison_preproofs.pdf

InterCorp 2013: *Czech National Corpus – InterCorp.* Institute of the Czech National Corpus, Praha. Available at: http://www.korpus.cz.

Martin, J. R. and White P. R. R. 2005. *The Language of Evaluation*. New York: Palgrave Macmillan.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. 1990. *A Student's Grammar of the English Language.* London: Longman.

Richterová, O. 2011. Can Translations Help Reveal Categorial Change? In F. Čermák (ed.), *Korpusová lingvistika Praha 2011 – 1 InterCorp.* Praha: NLN.

Richterová, O. 2012. Participial adjectives in Czech and their German equivalents: How to reveal an ongoing grammaticalization process with the help of translations. In *General and specialist translation / interpretation: theory, methods, practice.* Kyiv: National Aviation University.

The empirical trend: ten years on

Geoffrey Sampson
University of South Africa
sampson@cantab.net

Linguistic science of the past half-century has often been distorted through neglect of normal scientific standards of empirical falsifiability.

One motive behind the rise of corpus linguistics has been a reaction against this unempirical research style, which has yielded some unfortunate results. Thus, one striking example of the dangers of the non-empirical approach was discussed by Geoffrey Pullum and Barbara Scholz (2002), who documented how far-reaching claims by generative linguists about "tacit" grammatical knowledge being innate in mankind were based very heavily on a single alleged fact. To acquire the correct rule for forming English questions in the absence of innate knowledge, it was argued, a child would need to hear a certain type of question which was claimed to be extremely rare in practice – so rare that innate knowledge offers the only reasonable explanation for children's success in mastering the rule. Noam Chomsky asserted that "you can easily live your whole life without ever producing a relevant example … you can go over a vast amount of data of experience without ever finding such a case" (Piattelli-Palmarini 1980: 114–115); for him the belief that each child encounters relevant evidence "strains credulity" (Chomsky 1976: 213). The same claim has been repeated by linguist after linguist quoted by Pullum and Scholz, over a period of decades. Yet none of these linguists have ever suggested that they had checked this rarity claim against observational data.

Pullum and Scholz, referring to corpus data, suggested that the question-type might occur more often than generative linguists supposed. Using the demographically-sampled speech section of the British National Corpus, representing casual, spontaneous speech of a cross-section of the British population, I calculated (Sampson 2002: 81) that an average rate at which one could expect to hear the forms at issue must be at least once every few days; the expected number of instances heard in a lifetime would be in the thousands.

Apparently, a revolutionary new theory about human psychology, seen by many outside the discipline as the central finding of modern linguistics, was based in significant part on the fact that linguists relied on their intuitions for information about how speakers use their

language, and those intuitions could be wildly wrong.

If a linguist calls him- or herself a "corpus linguist" it is safe to infer that he or she works empirically, but most linguists do not use this self-description. A paper presented to the 2003 CL conference (published as Sampson 2005) used a quantitative literature survey to examine how far the discipline of linguistics as a whole had been affected by the newer trend towards using corpora and other empirical data sources. The result (Fig. 1, showing proportions of observation-based to intuition-based publications) turned out to be ambiguous. The all-time low level of empirical work around 1970 was not maintained, but after that level rose gently to 1980 it subsequently remained essentially flat except for a blip (which might have been a chance matter) about 1995, and (except for that blip) it never returned to the level of 1950. (The year 1950 was sampled to give a baseline well before the issue of speaker's intuition as a valid data source became part of the discourse of the discipline.)

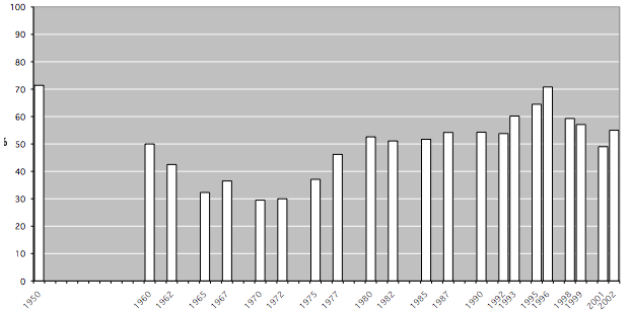


Figure 1. The 2003 figures (smoothed)

The graph might even be read as showing a falling away from empiricism over the very last years surveyed. This seeming trend reversal fell so shortly before the survey date that it was not possible to know whether it was real or a chance fluctuation. But published remarks at the time suggested that influential believers in the superiority of introspection over empirical observation as a research technique were challenging the shift towards empirical work. Reviewing a corpus-based study by Rosamund Moon, Thomas Nunnally (2002: 177) wrote:

> it is intuition that signals ill-formedness, not frequency of formations per million words …

In 2003 Frederick Newmeyer even devoted his Presidential Address to the Linguistic Society of America to arguing against "usage-based" linguistics (Newmeyer 2003: 696). These were the first occasions I had encountered when mainstream generative linguists attacked the corpus-based approach to research as explicitly undesirable (rather than just ignoring it).

By now a further decade of data is available, so it is worth asking whether updating the analysis changes the picture. As part of my research for a forthcoming book on the implications of corpus linguistics (Sampson and Babarczy 2013), I have repeated my survey, updating it and making the sampling somewhat more systematic. The picture has indeed changed: it has become much clearer. Without revisiting the methodological issues I discussed at length in 2003, the present paper considers what has happened to the figures.

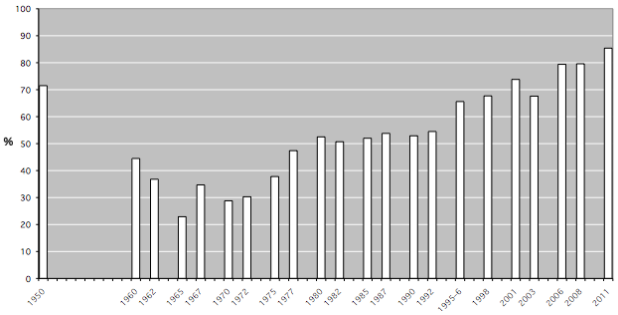


Figure 2. The updated figures (smoothed)

Examining the new figures (Fig. 2), we find that the 1990s "blip" and subsequent downturn in the 2003 paper have washed out. What we now see is an increase, as steady as is plausible for a complex discipline on the borders of humanities and social sciences, from a low of empiricism around 1970 to levels which recently have comfortably exceeded the 1950 baseline. If it is appropriate to discuss the issue as a struggle between opposed methodologies, then it is a struggle which the empiricists are winning. This has not gone unnoticed by observers of the field. Thus in 2008 Brian Joseph, then editor of *Language*, remarked:

> research papers are more experimentally based now than ever before … Also, they are more corpus-based … referees increasingly are demanding such multiple sources of data from authors who fail to offer them … this trend towards "cleaner" and more "natural(istic)" data is all to the good of the field. (Joseph 2008: 687)

Changes of mind by an academic discipline take time to filter out to educated general readers (or even to neighbouring disciplines), so I believe that plenty of members of the public who take an interest in ideas about language are still heavily influenced by unempirical approaches and their "findings". And in this case the time-lag which applies to all new intellectual movements is reinforced by a powerful special factor:

speculative, intuition-based theorizing makes for easier reading than discussion based on detailed empirical data. The latter will inevitably involve numerous complications and qualifications which matter, but which are tedious for non-specialists to wade through (whereas linguists who rely on intuition are free to paint their pictures in broad and simple strokes). Writings addressed by empirical linguists to wider audiences can only do so much to shield readers from those complexities. On student courses, where it is not over-cynical to suggest that many hearers are more interested in acquiring just enough knowledge to get their degree than in chasing the truth down twisting byways, the speculative style of linguistics goes down far better than the empirical style. Claude Hagège noted how language-teachers who study linguistics as part of their professional training find the genre of linguistics criticized here relatively accessible and hence appealing (Hagège 1976: 17 n1); and in practice language-teachers form a main avenue via which academic linguists' ideas disseminate into the wider world.

So it is comprehensible if the general public understanding of language continues to be distorted by the unempirical research style which came into vogue in the 1960s. But the implication of this paper is that things are not always going to be like that. A current of opinion created by a novel academic trend cannot survive indefinitely without ongoing support from the survival of that trend, and unempirical linguistics is not surviving. Max Planck (1949: 33–34) pointed out that a new scientific approach does not win acceptance by convincing opponents, but by the fact that the opponents eventually make way for a new generation. At present, unempirical linguistics is still being written and read. But it seems safe to say that this is a temporary state of affairs.

References

Chomsky, A.N. 1976. *Reflections on language*. London: Temple Smith.

Hagège, C. 1976. *La Grammaire générative: réflexions critiques*. Paris: PUF.

Joseph, B.D. 2008. "Last scene of all …". *Language* 84 (4): 686–90.

Newmeyer, F.J. 2003. "Grammar is grammar and usage is usage". *Language* 79 (4): 682–707.

Nunnally, T.E. 2002. Review of Moon, *Fixed expressions and idioms in English*. Language 78 (1): 172–177.

Piattelli-Palmarini, M. (ed.) 1980. *Language and learning: the debate between Jean Piaget and Noam Chomsky*. London: Routledge.

Planck, M. 1949. *Scientific autobiography and other papers*. New York: Philosophical Library.

Pullum, G.K. and Scholz, Barbara C. 2002. "Empirical assessment of stimulus poverty arguments". In Ritter (2002).

Ritter, Nancy A. (ed.) 2002. *A review of the poverty of stimulus argument*. Special issue of *The Linguistic Review*, 19 (1–2).

Sampson, G.R. 2002. "Exploring the richness of the stimulus". In Ritter (2002).

Sampson, G.R. 2005. "Quantifying the shift towards empirical methods". *International Journal of Corpus Linguistics* 10 (1): 15–36.

Sampson, G.R. and Babarczy, Anna. 2013. *Grammar without grammaticality*. Berlin: de Gruyter.

Identifying discourse(s) and constructing evaluative meaning in a gender-related corpus (GENTEXT-N)

José Santaemilia
Universitat de València
jose.santaemilia@uv.es

Sergio Maruenda
Universitat de València
sergio.maruenda@uv.es

In recent years the enactment of legislation, both in Spain and in the UK, on fundamental gender-related issues – abortion, gender-based violence or homosexual marriages – has given rise to conflicting discourses within both Spanish and British societies that are replicated, amplified, deprecated, perverted or exploited by mass media, political organisations and religion institutions, with a view to demanding either respect or neglect for the minorities to which these legal measures are addressed.

In this paper, we present the work of the research group GENTEXT (*Gender, Language and Sexual (In)Equality*), based at the *Universitat de València*, as part of a larger research project on the elaboration and exploitation of a macro-corpus on gender, social inequalities and political discourse[1]. Our aim is to compile a pragmatic-discursive dictionary with the most relevant terms on social inequality, gender or political issues that feature in our distinct corpora. These corpora will be made public on-line for academic researchers, politicians, educators, experts on gender issues, UE organisations, etc., and our results will be further applied to the teaching and learning of specialised language and translation.

Our research group GENTEXT has set out to document and analyse the concepts, the discursive processes, the ideological tensions and the semantic/pragmatic negotiation of meanings behind this complex discursive reality. For this aim we compiled a 40-million word, comparable (Spanish-English), highly-specialised corpus (GENTEXT-N), made up of Spanish and British dailies embodying liberal vs. conservative positions. The compilation process was carried out using *Nexis UK news databases* and it comprises the period 2005-2012. Our research centres around the analysis of key terms and concepts such as *gender-based* or *domestic violence*, *homosexual(ity)*, *gay* and *abortion*.

On the present stage of this project, we have carried out a keyword analysis that has pointed to the favoured lexical areas and concepts for the selection of entries for the dictionary, based on statistical significance as compared to larger reference corpora. We believe that keywords stand as indicators of discursive or rhetorical strategies, from naming strategies to argumentation, from predication to discursive construction of key gendered or sexualised terms –e.g. *homosexuality*, *family* or *marriage*- used to construct discourses and mould opinions.

Thus, the driving force of these new discourses ensuing social debate is a combination of socio-ideological tensions and semantic/pragmatic negotiation. In this respect, we also advocate for a qualitative analysis of the context, which is inextricably linked to social actors, to historical circumstances, to ideological factors, to power asymmetries, etc. Our basic methodological assumption is that, in order to analyse this complex reality, we need a combination of qualitative and quantitative analyses (as advocated, among others, by Baker and McEnery 2005, Baker et al 2008, Caldas-Coulthard 2010) is essential if we wish to grasp both the linguistic and the ideological underpinnings of the heterogeneous texts we are investigating. Thus, our analyses integrate, on the one hand, critical discourse analysis and Relevance Theory (lexical) pragmatics (Carston 2002; Sperber and Wilson 1998; Wilson 2003; Wilson and Carston 2007) and, on the other, Corpus Linguistics techniques to fully exploit the potentialities of both approaches, thus trying to avoid the oversimplification of ideological bias.

With regard to the latter, we have also researched on semantic/discourse prosodies (Louw 1993; Sinclair 1991, 2000; Stubbs 2001). These help transcend the collocation or even sentential scope to reveal discursive patterns and, consequently, trace evaluative relations (Martin and White 2005). Attitudinal meanings tend to spread out and colour discourse as participants take up a stand oriented to affect, judgement or evaluation (Martin and White 2005: 43). These constitute a network or constellation of discursive concepts which contribute to shaping and (de)legitimising citizens' discourses and rhetorical frameworks within communities of practice.

These (counter-)discourses engage in an ideological colloquy of a large scale: They respond to and affirm ideas and beliefs, they anticipate possible reactions and dissent, they seek endorsement, and so on (Volonishov 1995: 139).

[1] FI2012-39289 – Ministerio de Economía y Competitividad – GEA (GENTEXT+ECPC+ADELEX): Un macro-corpus sobre genero, desigualdad social y discurso político. Análisis y elaboración de materiales didácticos, lexicográficos y computacionales.

Bakhtin (1981: 281) states that discourse is *heteroglossic*, that is it exists against a backdrop of contradictory opinions, points of view and value judgements. Following Martin and White (2005), we are concerned with the way this heteroglossic perspective plays a part in meaning making processes by which participants negotiate attitudinal assessments vis-à-vis the socially constituted communities of shared attitudes and beliefs associated with the positions in the texts, and also work toward obtaining endorsement.

Although we acknowledge the relevance and significance of semantic prosodies in our search for *traces* of ideological discourses in our corpora, in this paper we argue for a more flexible approach to meaning-construction that allows the pragmatic enrichment of concepts in the light of new contextual assumptions. The meaning of these concepts is dynamic and constantly renegotiated in discourse through the exploitation of a whole range of evaluative slants.

We introduce and define the term *discourse constellations* to refer to a form of organising the multiplicity of conceptual representations subject to ideological negotiation and social and political pressure in/between communities of practice. These are nebulous realizations of conflicting ideological concepts/discourses in today's societies and as such they are imprecise and constantly changing, in continuous struggle to become legitimised or *core*, subject to processes of pragmatic adjustment when meaning negotiation comes into play.

References

Bakhtin, M. 1981. *The Dialogic Imagination*. Austin: University of Texas Press.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P. et al. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19(3): 273-306.

Baker, P. and McEnery, T. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics* 4(2): 197-226.

Caldas-Coulthard, C. and Mund, R. 2010. 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society* 21(2): 99-133.

Carston, R. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell Publishing.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP.

Louw, W. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker et al (eds.) *Text & Technology: In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 157-176.

Martin, J. and White, P. 2005. *The Language of Evaluation: Appraisal in English*. London: Palgrave Macmillan.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. 2000. Lexical Grammar. *Naujoji Metodologija* 24: 191-203.

Sperber, D. and Wilson, D. 1998. The mapping between the mind and the public lexicon. In P. Carruthers and J. Boucher (eds.) *Thought and Language*. Cambridge: Cambridge University Press, 184-200.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Voloshinov, V. 1995. *Marxism and the Philosophy of Language, Bakhtinian Thought – an Introductory Reader*. London: Routledge.

Wilson, D. 2003. Relevance theory and lexical pragmatics. *Rivista di Linguistica* 15(2): 273-291.

Wilson, D. and Carston, R. 2007: A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (ed.) *Pragmatics*. Basingstoke: Palgrave, 230-259.

Whittsit, S. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3): 283-305.

Comparing morphological tag-sets for Arabic and English

Majdi Sawalha
University of Jordan
sawalha.majdi@gmail.com

Eric Atwell
University of Leeds
e.s.atwell@leeds.ac.uk

1 Introduction

Part-of-speech taggers are used to enrich a corpus by adding a part-of-speech category label to each word, showing the broad grammatical class of the word, and morphological features such as tense, number, gender, etc. The list of all grammatical category labels is called the tag set. The design of the tag set is an important prerequisite to this annotation task. The task requires a tagging scheme, where each tag or label is practically defined by showing the words and contexts where each tag applies; and a tagger, a program responsible for assigning a tag to each word in the corpus by implementing tag set and tagging scheme in a tag-assignment algorithm (Atwell 2008).

2 Part-of-speech tagging and part-of-speech tag sets for English

Enriching the source text samples of corpora with part-of-speech information for each word, as a first level of linguistic enrichment, results in more useful research resources. English corpora have been developed for a long time and for a variety of formats, types and genres. Several English corpora have been enriched with Part-of-Speech tagging, and a variety of different English corpus part-of-speech tag sets have been developed, including: the Brown corpus (BROWN), the Lancaster/ Oslo-Bergen corpus (LOB), the Spoken English Corpus (SEC), the Polytechnic of Wales corpus (PoW), the University of Pennsylvania corpus (UPenn), the London-Lund Corpus (LLC), the International Corpus of English (ICE), the British National Corpus (BNC), the Spoken Corpus Recordings In British English (SCRIBE), etc (Atwell 2008). The AMALGAM [1] multi-tagged corpus amalgamates all these tagging schemes in a common collection of English texts: in the AMALGAM corpus, the different part-of-speech tag sets used in these English general-purpose corpora are applied to illustrate the range of rival English corpus tagging schemes, and the texts are also parsed according to a range of rival parsing schemes, so each sentence has more than one parse-tree, called "a forest" (Atwell, Demetriou et al. 2000). Figure 1 shows an example a tagged sentence from ALMAGAM. Part-of-speech tag sets and taggers have also been developed for other European languages. The EAGLES, European Advisory Group on Language Engineering Standards project, drew up standards for tag sets, morphological classes and codes for (western) European languages, including EAGLES Recommendations for the morphosyntactic annotation of corpora (Leech and Wilson 1999); a synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: a common proposal and applications to European languages (Monachini and Calzolari 1996); and an EAGLES study of the relation between tag sets and taggers (Teufel, Schmid et al. 1996). Many morphosyntactic schemes for different languages applied the EAGLES guidelines. Example projects are: the MULTEXT project; the GRACE project; the CRATER project; and the morphosyntactic tag set of Urdu. The four projects and the tag set of Urdu are discussed in Hardie (2003 and 2004).

	Brown	ICE	LLC	LOB	PARTS	PoW	SEC	UPenn
select	VB	V(montr,imp)	VA+0	VB	adj	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)	.	.	.	.	.	.

Figure 1. Example sentence illustrating rival English part-of-speech tagging (from the ALMAGAM multi-tagged corpus)

3 Arabic language part-of-speech taggers, tag sets and corpora

Arabic part-of-speech tagging development started more recently. A range of different techniques have been used to solve the problem of part-of-speech tagging of Arabic (Freeman 2001; Khoja 2001; Diab, Hacioglu et al. 2004; Duh and Kirchhoff 2005; Habash and Rambow 2005; Marsi, Bosch et al. 2005; Al-Shamsi and Guessoum 2006; Tlili-Guiassa 2006; Zibri, Torjmen et al. 2006; Alqrainy 2008). Nearly all these Arabic part-of-speech taggers were developed by NLP research groups for their own

[1] Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm

internal use, and are not freely downloadable by other researchers. The taggers use different tag sets, and accuracies are reported on different test corpora.

Arabic corpora started to appear in the late 1980s; A list of Arabic corpora developed outlines their size, type, purpose of development and the materials of construction is found in (Al-Sulaiti and Atwell 2006). Nearly most Arabic corpora have been collected by Arabic corpus linguistics research groups for their own purposes, and are not freely downloadable. The Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell 2004; Al-Sulaiti and Atwell 2005; Al-Sulaiti and Atwell 2006), and the Quranic Arabic Corpus (QAC) (Dukes, Atwell et al. 2010), both developed at the University of Leeds, are the only freely available Arabic corpora on the web. In computational linguistics research, the most widely used annotated corpus of Arabic is the Penn Arabic Treebank (Maamouri and Bies 2004) distributed (at cost) by LDC Linguistic Data Consortium.

This section covers the most important Arabic tag sets and tag set design methodologies. These tag sets are; (1) Khoja's Arabic tag set, (2) Penn Arabic Treebank tag set, (3) ARBTAGS, (4) The Quranic Arabic Corpus morphological tag set, (5) The MorphoChallenge 2009 Qur'an Gold Standard tag set (6) CATiB part-of-speech tag set and (7) the SALMA – Tag Set. The tag sets range from a small set of short tags analogous to BNC or LOB tag sets for English on one hand. On the other hand, longer more detailed morphological tag sets (*e.g.* Penn Arabic Treebank (FULL) tag set) which are analogous to ICE tag set for English.

Khoja's Arabic Tag Set: During early research on developing a part-of-speech tagger for Arabic text, (Khoja, Garside et al. 2001; Khoja 2003) developed a tag set for Arabic which is based on traditional Arabic grammar categories rather than modern European EAGLES standards. Khoja's tag set contains 177 tags; 103 types of noun, 57 verbs, 9 particles, 7 residuals and 1 punctuation. Khoja's tag set includes the morphological features of gender, number, person, case, definiteness and mood.

Penn Arabic Treebank (PATB) Part-of-Speech Tag Set: The most widely used tag set for Arabic is the Penn Arabic Treebank tag set used to annotate the Penn Arabic Treebank (PATB) with part-of-speech tags. The Penn Arabic Treebank model postulates a FULL tag set which compromises over 2000 tag types (Diab 2007). This includes combinations of 114 basic tags listed in the Linguistic Data Consortium (LDC) Arabic part-of-speech/morphological tagging documentation. The FULL tag set exhibits a wide range of morphological features: case, gender, number, definiteness, mood, person, voice, tense, aspect, etc. The LDC also introduced the reduced tag set (RTS) of 25 tags which is designed to maximize the performance of Arabic syntactic parsing. The morphological features marked by the RTS tag set are case, mood, gender, person and definiteness (Diab 2007).

ARBTAGS Tag Set: (Alqrainy 2008) developed a new part-of-speech tag set called ARBTAGS to be used in the development of a part-of-speech tagger. The tag set design followed the criteria proposed by Atwell (2008). ARABTAGS tag set was built on traditional Arabic grammar books to design the tag set. Six morphological features of Arabic words were included: gender, number, case, mood, person and state. ARBTAGS contains 161 detailed tags and 28 general tags to cover the main part-of-speech classes and sub-classes. The 161 detailed tags are divided into 101 nouns, 50 verbs, 9 particles and 1 punctuation mark.

MorphoChallenge 2009 Qur'an Gold Standard Part-of-Speech Tag Set: MorphoChallenge2009[1] Qur'an gold standard is developed using the data of Morphological Tagging of the Qur'an database (Talmon and Wintner 2003; Dror, Shaharabani et al. 2004). It was developed to be used to evaluate morphological analyzers in the Morphochallenge 2009 competition, which aims to develop an unsupervised morphological analyzer to be used for different languages including Arabic. It contains the full morphological analysis for each word, according to the Tagged database of the Qur'an but reformatted to match other Morphochallenge test sets in other languages. The word's morphological analysis is shown after each word where the morphological features are separated by space and "+" sign. These features include the part-of-speech of the word, number, gender, person, case, definiteness, voice and others.

The Quranic Arabic Corpus Part-of-Speech Tag Set: The Quranic Arabic Corpus is a newly available resource enriched with multiple layers of annotation including morphological segmentation and part-of-speech tagging. The Quranic Arabic Corpus tag set adapts historical traditional Arabic grammar, which leads to morphological annotation that uses terminology familiar to many readers of the Qur'an. This terminology enables people with Qur'anic syntax experience to

[1] MorphoChallenge 2009 Qur'an Gold Standard http://www.cis.hut.fi/morphochallenge2009/datasets.shtml

participate in the online annotation to be verified against existing authenticated books on Quranic Grammar (Dukes and Habash 2010).

Columbia Arabic Treebank CATiB Part-of-Speech Tag Set: Another tag set was designed for the part-of-speech and syntactic annotation in the Columbia Arabic Treebank CATiB. A part-of-speech tag set consisting of only six tags is used for the part-of-speech annotation of CATiB. The main reason for using such a small tag set is a tradeoff between linguistic richness and Treebank size. The researchers' assumption for morpho-syntactically rich languages such as Arabic, is that the cost of fine-grain annotation is a slower annotation process, a smaller Treebank and less data to train tools. CATiB is inspired by two ideas. First, it avoids annotation of redundant linguistic information. Second, it uses linguistic representation and terminology from traditional Arabic syntactic studies (Habash, Faraj et al. 2009).

The SALMA – Tag Set: The SALMA Morphological Features Tag Set (SALMA, Sawalha Atwell Leeds Morphological Analysis tag set for Arabic) captures long-established traditional morphological features of grammar and Arabic, in a compact yet transparent notation. First, we introduce Part-of-Speech tagging and tag set standards for English and other European languages, and then survey Arabic Part-of-Speech taggers and corpora, and long-established Arabic traditions in analysis of morphology. A range of existing Arabic Part-of-Speech tag sets are illustrated and compared; and we review generic design criteria for corpus tag sets. For a morphologically-rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. We describe the SALMA Tag Set in detail, explaining and illustrating each feature and possible values. In our analysis, a tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash "-" represents a feature not relevant to a given word. The first character shows the main Parts of Speech, from: noun, verb, particle, punctuation, and Other (residual); these last two are an extension to the traditional three classes to handle modern texts. 'Noun' in Arabic subsumes what are traditionally referred to in English as 'noun' and 'adjective'. The characters 2, 3, and 4 are used to represent subcategories; traditional Arabic grammar recognizes 34 subclasses of noun (letter 2), 3 subclasses of verb (letter 3), 21 subclasses of particle (letter 4). Others (residuals) and punctuation marks are represented in letters 5 and 6 respectively. The next letters represent traditional morphological features: gender (7), number (8), person (9), inflectional morphology (10) case or mood (11), case and mood marks (12), definiteness (13), voice (14), emphasized and non-emphasized (15), transitivity (16), rational (17), declension and conjugation (18). Finally there are four characters representing morphological information which is useful in Arabic text analysis, although not all linguists would count these as traditional features: unaugmented and augmented (19), number of root letters (20), verb root (21), types of nouns according to their final letters (22). The SALMA Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

4 Conclusions

The comparison of the morphological features used in the different tag sets of Arabic shows shared common features such as gender, number, person, case, mood and definiteness. Features such as voice, tense and aspect are included in the PATB FULL tag set. State is included in the ARBTAGS tag set. Diptotic is a feature of the MorphoChallenge 2009 tag set, and verb form and derivation are features of the QAC tag set. Chapter 6 discusses the 22 morphological features of The SALMA Tag Set defines a fine-grained tagging scheme which is describes 22 morphological features of the word's morphemes based on traditional Arabic grammar.

Finally, we examine the range of PoS-tagsets for English corpora, and note an analogous issue. The AMALGAM website[1] details the PoS-tagsets used in Brown, ICE, London-Lund, LOB, Parts, PoW, SEC and UPenn corpora of English. Although English does not have the morphological complexity of Arabic, there are still significant differences between these English corpus PoS tagsets. We see the need for an English tagset where features are clearly enumarated, equivalent to the SALMA tagset for Arabic. Arabic corpus linguists can build on the cultural heritage of formal description of the Classical Arabic of the Quran; however, there is no equivalent key classical text and tradition for English.

References

Al-Shamsi, F. and A. Guessoum (2006). A Hidden Markov Model-Based POS Tagger for Arabic. 8es

[1] http://www.comp.leeds.ac.uk/ccalas/tagsets/tagmenu.html

Journees internationales d'Analyse statistique des Donnees Textuelles.

Al-Sulaiti, L. and E. Atwell (2004). Designing and developing a corpus of contemporary Arabic TALC 2004: Proceedings of the sixth Teaching And Language Corpora conference.

Al-Sulaiti, L. and E. Atwell (2005). aConCorde: Towards a Proper Concordance of Arabic. Corpus Linguistics conference 2005, University of Birmingham, UK.

Al-Sulaiti, L. and E. Atwell (2006). "The design of a corpus of contemporary Arabic." International Journal of Corpus Linguistics 11: 135-171.

Alqrainy, S. (2008). A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging. 2008. Leicester, UK, De Montfort University. PhD: 197.

Atwell, E. (2008). Development of tag sets for part-of-speech tagging. Corpus Linguistics: An International Handbook, Volume 1. A. Ludeling and M. Kyto, Mouton de Gruyter: 501-526

Atwell, E., G. Demetriou, et al. (2000). "A comparative evaluation of modern English corpus grammatical annotation schemes." ICAME Journal, International Computer Archive of Modern and medieval English, Bergen 24: 7-23.

Diab, M., K. Hacioglu, et al. (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. Proceedings of HLT-NAACL

Diab, M. T. (2007). Towards an Optimal POS Tag Set for Arabic Processing. Proc RANLP.

Dror, J., D. Shaharabani, et al. (2004). "Morphological Analysis of the Qur'an." Literary and Linguistic Computing 19(4): 431-452.

Duh, K. and K. Kirchhoff (2005). POS Tagging of Dialectal Arabic: A Minimally Approach. ACL-05, Computational Approaches to Semitic Languages Workshop Proceedings. University of Michigan Ann Arbor, Michigan, USA: 55-62.

Dukes, K., E. Atwell, et al. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. Language Resources and Evaluation Conference (LREC 2010). Valletta, Malta.

Dukes, K. and N. Habash (2010). Morphological Annotation of Quranic Arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta,19-21 May 2010., European Language Resources Association (ELRA).

Freeman, A. (2001). Brill's POS Tagger and a Morphology Parser for Arabic. NAACL 2001 Student Rersearch Workshop, Lancaster University.

Habash, N., R. Faraj, et al. (2009). Syntactic Annotation in Columbia Arabic Treebank. 2nd International Conference on Arabic Language Resources & Tools MEDAR 2009. Cairo, Egypt.

Habash, N. and O. Rambow (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan, Association for Computational Linguistics.

Khoja, S. (2001). APT: Arabic Part-of-Speech Tagger. Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania.

Khoja, S. (2003). APT: An Automatic Arabic Part-of-Speech Tagger. Computing Department. Lancaster, UK, Lancaster University. PhD: 157.

Khoja, S., P. Garside, et al. (2001). A tagset for the morphosynactic tagging of Arabic. Corpus Linguistics 2001. Lancaster University, Lancaster, UK.

Leech, G. and A. Wilson (1999). Standards for Tagsets. Syntactic Wordclass Tagging. H. v. Halteren, KLUWER Academic Publishers: 55-80.

Maamouri, M. and A. Bies (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).

Marsi, E., A. v. d. Bosch, et al. (2005). Memory-based morphological analysis generation and part-of-speech tagging of Arabic. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, Association for Computational Linguistics.

Monachini, M. and N. Calzolari (1996). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages. Pisa, Italy, Istituto di Linguistica Computazionale -CNR.

Talmon, R. and S. Wintner (2003). Morphological Tagging of the Qur'an. In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop. Budapest, Hungary.

Teufel, S., H. Schmid, et al. (1996). Study of the relation between tagsets and taggers. Stuttgart, Germany Institut für maschinelle Sprachverarbeitung, Universität Stuttgart

Tlili-Guiassa, Y. (2006). "Hybrid Method for Tagging Arabic Text." Journal of Computer Science 2(3): 245-248.

Zibri, C. B. O., A. Torjmen, et al. (2006). "An Efficient Multi-agent system Combining POS-Taggers for Arabic Texts." CICLing 2006, LNCS 3878(pp.121-131).

Comparing collocations in the totalitarian language of the former Czechoslovakia with the language of the democratic period

Věra Schmiedtová
Charles University
vera.schmiedtova@ff.cuni.cz

1 Introduction

In 1989 the Czechoslovak Republic underwent major political changes. The period of communist totalitarian rule came to an end and the country returned to democracy. This paper is an attempt to show how political discourse has changed and to what extent these changes are reflected in language. We illustrate these changes through the frequency of collocations of certain words present in the top one hundred most frequently occurring auto-semantic words nouns and adjectives included in the Dictionary of Communist Totalitarianism [1] which is based on the Totalitarianism Corpus[2] . Thís corpus represents the period of communist rule (1948-1989). In order to compare collocations of specific words from the period of totalitarianism and those from the democratic period, we use a newly established corpus of journalistic texts, which represent the period from the return to democracy in 1989 to the present day (Corpus Věra[3]). The two corpora are comparable in terms of size and the type of texts used.

Following software tools are used when working with corpora: Bonito [4] and Sketch Engine [5] . The Dictionary of Communist Totalitarianism, which is based on frequencies in the Totalitarianism Corpus and includes the following words chosen among the top one hundred most frequently occurring words[6]:

nouns*: life, person, organization, society, struggle*

adjectives: *new, high, American, communist*

For comparison we have also chosen words that are not among the first hundred words in the Dictionary of Communist Totalitarianism, but are typical for the communist period:

nouns: *love, happiness*

adjectives: *native, cordial, forged, indeterminable, corrupt, orthodox*

In general usage the words chosen are neutral, but we show how their usage has changed between the two periods – how the key ideas of society have changed.

The following tables show the frequency of collocations in the Totalitarianism Corpus compared with their frequency in the Corpus Věra. As an example we have chosen the noun ***life***, which is part of our basic vocabulary.

	Totalitarianism	Věra
rich	87	6
spiritual	117	11
economic	244	6
beautiful	23	3
cultural	561	43
better	136	23
human	341	164
international	65	0
peaceful/ in peace	87	1
new	336	47
tranquil	31	1
joyful	118	0
contemporary	64	15
social	714	48
happy	404	9

Table 1

2 Evaluation of the tables

Table 1 is made up of words with a positive semantic prosody. (*rich, beautiful, better, new, joyful, happy* life*).* They were used to describe life under communism. The communist ideology presented itself as the way of peace opposed to war, as we can see in the expressions (*tranquil, peaceful* life); it was focused on manufacturing and production, which explains the frequent occurrence of the word *economic;* the workers' movement was *international;* attention was shifted away from the focus on private sphere towards *social* life; *contemporary* life was placed

[1] Dictionary of Communist Totalitarianism (Čermák, F., Cvrček, V., Schmiedtová, V. [eds.]), published Lidové noviny 2010

[2] The Totalitarianism Corpus is a special corpus made up of three samples of scans of Rudé právo, the official daily newspaper of the Communist Party of Czechoslovakia (1952; 1969; 1977) and of 91 scanned propaganda sources from the time. The Totalitarianism Corpus includes around 15 million words.

[3] The Corpus Věra is on a similar scale to the Totalitarianism Corpus, i.e. around 15 million words. It is made up of extracts from some of the main Czech daily newspapers – Hospodářské noviny, Lidové noviny, Deníky Moravia, Deníky Bohemia, Metro, from the period 2000-2009.

4 Rychlý, Pavel: Bonito – graphical user interface to the system Manatee, version 1.80

[5] Sketch Engine, Lexical Computing Ltd, ver: SkE-2.39-2.78

[6] The most frequently occurring auto-semantic words are predominantly nouns and adjectives.

in contrast with life in the past, non-communist society; paradoxically the propaganda puts great emphasis on *human* life, although it was not valued by the regime; the greatest paradox was the promotion of *spiritual* and *cultural* life because both were stifled in particular by censorship and lack of freedom. These examples of collocation frequency reflect the mendacity and manipulative nature of communist propaganda: the abuse of common language to achieve propaganda ends without reflecting reality as people who lived under totalitarian rule knew it and as we are now finding out through the study of the past.

	Totalitarian	Věra
political	515	22
socialist	49	0
within the party	60	0
Lenin's	13	0
party	114	0

Table 2

Table 2 is made up of adjectives which we would expect to occur frequently during the period of totalitarianism. They point to its main priority – to promote a certain political view of the world. We see a dramatic decrease in the use of these collocations in contemporary language and in contemporary society.

	Totalitarianism	Věra
ordinary/everyday	7	79

Table 3

Table 3 indicates a word returning to its normal usage. In the language of communist propaganda nothing ordinary or everyday was emphasized. Everything was geared towards the promotion of communist ideals.

	Totalitarianism	Věra
whole	250	299
sexual	13	17
family	64	44
private	61	43

Table 4

Table 4 is made up of neutral collocations in the language of both periods. The similar frequency of the adjective *sexual* comes as a surprise. We know communist propaganda and ideology to have been very prudish when talking of sex.

3 Evaluation

The language of totalitarianism in the former Czechoslovakia works with the semantic structure of Czech. However, it has a propaganda function, resulting in words that are in general usage being abused for propaganda purposes. It is aggressive, monotonous, frequently repeating certain associations. It adds a positive or negative character to certain words. For example, the word *American* is always negative, the word *Soviet* is always positive. It frequently abuses words with a positive semantic prosody for ideological purposes. It creates new meanings of words through new associations, for example *western = capitalist.* It has a liking for certain areas of meaning, such as: building a better future; the struggle against enemies of the new order; "the democratic character of culture and education", which is a concealed way of referring to censorship in these fields. Euphemisms are often used with a view to concealing reality. This language is not creative; it draws from automated components of language. It often uses set phrases. To this day these word combinations are very often used as quotations in reference to the communist period.

References

Čermák, F., Cvrček, V., Schmiedtová, V. (2010): Dictionary of Communist Totalitarianism, Lidové noviny, Prague

Schmiedtová, V. (2006):-The First Conference of The Slavic Linguistics Society – What did the totalitarian language in the former socialist Czechoslovakia look like? http://www.indiana.edu/~sls2006/page2/page2.htm

Schmiedtová, V. (2007): Totalitní jazyk .v bývalém Československu. Koncept slova práce. In: Totalitarismus 3, sborník z konference, katedra antropologie, FF ZU, s.110 – 116

Schmiedtová, V. (2008): Hodnotící prostředky v totalitním jazyce 1948-1989 v bývalém Československu. In: Totalitarismus 4, sborník z konference, katedra antropologie FF ZU. Plzeň, s. 186 – 196

Schmiedtová, V. (2011): Die Sprache der Propaganda in der Tschechoslowakei 1948-1989 In Brücken, Germanistisches Jahrbuch Tschechien-Slowakei 2011 Nakladatelství Lidové noviny ISBN 1803-456X,s. 93-115

Rychlý, P.: Bonito – graphical user interface to the system Manatee, version 1.80

Rychlý, P. (2007): Manatee/Bonito – A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, 2007. p. 65-70. ISBN

978-80-210-4471-5.

Sketch Engine, Lexical Computing Ltd, ver: SkE-2.39-2.78

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004): The Sketch Engine, EURELEX 2004, Loriet, France, p. 105-116

Linguistic means of knowledge transfer through knowledge-rich contexts in Russian and German

Anne-Kathrin Schumann
Saarland University / University of Vienna
anne.schumann@mx.uni-saarland.de

1 Introduction and related work

Defining statements have long been recognised as a fundamental means of knowledge transfer. Corpus-based research on the description and automated extraction of such statements has produced results for a variety of languages, e. g. English (Pearson 1998, Meyer 2001, Muresan and Klavans 2002, Marshman 2008), French (Malaisé et al. 2005), Spanish (Sierra et al. 2008), German (Storrer and Wellinghoff 2006, Walter 2010, Cramer 2011), Slovenian (Fišer et al. 2010), "Slavic" (Przepiórkowski et al. 2007), Portuguese (Del Gaudio and Branco 2007) and Dutch (Fahmi and Bouma 2006, Westerhout 2009). These studies describe linguistic properties of defining statements in text corpora, lexico-syntactic patterns or extraction grammars. Not all of them report results of extraction experiments, but many of the papers that do so combine linguistically informed extraction methods with machine learning or other filtering methods.

Only few studies, however, provide a systematic description of the linguistic properties of defining statements and statistical studies on this topic seem to be largely missing, while results of small-scale empirical studies suggest that the amount of variation that can be observed in empirical data is not appropriately depicted by the descriptions of defining statements in the literature (Walter 2010). Moreover, the question whether there are universals common to defining statements in different languages remains unresolved.

This paper presents a linguistic analysis of knowledge-rich contexts from a gold standard that was created on the basis of Russian and German web corpora as part of ongoing PhD thesis work. In the following sections, the concept of knowledge-rich contexts is refined and gold standard creation is shortly described. Linguistic analyses and their results are explained.

2 Knowledge-rich contexts and definitions

From a non-scientific perspective, knowledge-rich

contexts (KRCs) can be described as pieces of text that may turn out to be helpful for a conceptual analysis task. For a better understanding of what this means in practice, examples 1 and 2 present a Russian and a German context found in our data.

For a more formal definition of KRCs, it is important to consider that KRC extraction is related to the development of terminological knowledge bases (Meyer et al. 1992) and concept systems. These systems stress the relevance of semantic relations holding between concepts. Consequently, KRC extraction aims at identifying contexts for specialised terms that provide semantic information about the underlying concepts, including information about semantic relations between concepts (e. g. meronymy, causality; definitions of relations for terminological purposes are given in ISO 1087-1: 2000). Moreover, KRCs, as defined in Schumann (2011), are related to a set of minimal validity criteria that, however, are formally much less strict than the criteria usually attached to definitions.

In practice, the boundary between definitions and KRCs is not always clear. Several of the above-mentioned studies employ the term "definition", whereas the types of "definitions" subsumed under this term vary considerably. For our own work, we assume that definitions are subtypes of KRCs which, in turn, echo the categories of "proper definition", "redundant definition", "complete definition" and "partial definition" as introduced by Bierwisch and Kiefer (1969) while covering a larger set of semantic relations, e. g. those relations that are relevant to terminological tasks, and satisfying less strict formal criteria.

3 Gold standard creation

Bierwisch and Kiefer (1969) are among the first to point out that linguistic criteria alone do not fully explain whether a statement can be considered defining or not. Cramer (2011) conducts extensive definition annotation experiments, concluding that the annotators' individual stance towards a candidate statement and the corresponding text, knowledge of the domain and other criteria influence whether a statement is considered defining.

In our own annotation work, KRC candidates were manually selected from two small German and two small Russian web corpora (development and test corpora) in two annotation cycles.

1) Система охлаждения служит для отвода излишнего тепла от деталей двигателя, нагревающихся при его работе.

   [The cooling system serves to remove excess heat from those parts of the engine that heat up during exploitation.]

2) Das Verhältnis Energieertrag ("Output") zu Input wird Leistungszahl genannt.

   [The relation between energy output and input is called coefficient of performance.]

Each candidate was then presented to two independent annotators – MA translation students and experienced translators – together with a term serving as definiendum in order to obtain a judgement on its KRC status. For positive judgements, annotators were also asked to give a simple binary assessment of their annotation confidence (e. g. "confident" vs. "not confident"). However, since this study cannot claim statistical relevance due to the small number of annotators per sentence and the preselection of KRC candidates by the author, qualitative criteria were then applied to select the KRCs that make up the final gold standard. Table 1 gives an overview over the small gold standard corpora and the numbers of KRCs contained by them.

Corpus	Tokens	KRCs
German_development	~ 160,000	337
German_test	~ 170,000	295
Russian_development	~ 99,000	292
Russian_test	~ 75,000	268

Table 1: Gold standard

The results of our (small) annotation experiment seem to support Cramer's (2011) claim that individual criteria of the annotators influence the annotation process, resulting in different rates of acceptance/rejection and varying levels of confidence.

4 Linguistic analyses

Linguistic analyses were performed in order to arrive at a description of the specific linguistic properties of the KRCs in the gold standard. To decide on the linguistic features that may be typical for KRCs, various linguistic approaches were combined:

- approaches to definition and KRC extraction (see section 1) that describe typical linguistic properties of defining statements,
- approaches describing linguistic means of conveying generality (Drescher 1992), since generalisability of the information

was one of the validity criteria for KRC candidate selection,

- Systemic Functional Linguistics (Halliday 1994) because of its holistic, functional approach that may be helpful in capturing phenomena beyond the word level.

On the lexical level, we studied both *POS and lemma frequencies*, comparing the KRC data for each language with an equal amount of randomly selected non-KRCs from the gold standard corpora. Statistical significance tests were used to test for differences between the datasets. The analysis shows that adverbs, modal and auxiliary verbs, verbal infinitives, particles, pronouns, deictic elements, adversative conjunctions and named entities tend to appear with significantly lower frequencies in the studied KRCs. Certain general nouns such as *вид* ("type", "kind") and *совокупность* ("the whole") for Russian or *Begriff* ("concept"), for German, were found significantly more often in the gold standard, whereas qualifying and quantifying adjectives (*новый*, "new", *gut*, "good", *wenige*, "few") and sentential adverbs (*даже*, "even", *очень*, "very", *nur*, "only") appear with a significantly lower frequency in the gold standard. Generalising quantifiers such as *alle* ("all") or *jeder (*"every"*)* are light indicators for KRCs. Taken together, this amounts to a tendency towards a rather unpersonal and generalising style. On the other hand, typical lexical elements of definitions or general statements (such as frequency adverbs) as described in the literature could not be found in high quantity, neither did we find linguistic evidence for hedges as discussed, for example, in Marshman (2008). In general, Cramer's (2011) claim that lexical properties of defining statements can be observed rather frequently by type, but quite infrequently by value, seems to hold for our data, too. Obviously, much larger datasets are necessary for an in-depth study of the lexical properties of KRCs.

Another feature that we studied were *semantic classes of verbs* (simple predicates) that may function as building blocks for "knowledge patterns" (Barrière 2004), that is, we investigated which kinds of verbs (besides interpunctional and syntactic features such as apposition) typically introduce conceptual information and, therefore, are candidates for extraction patterns. To this end, we manually identified candidates for such verbal patterns in the German and Russian development corpora in order to compare them to the predicates in equal amounts of non-KRCs from the gold standard corpora. For these verbs, we then extracted semantic classes from GermaNet (Hamp/Feldweg 1997), a wordnet-like lexical resource for German, and a Russian thesaurus (Baranov 1995). For German, more than 50 % of the identified verbs in the gold standard pertain to the GermaNet classes of "general" (*Allgemein*) verbs or verbs of communication (*Kommunikation*) which are both significantly overrepresented in the gold standard. For Russian it was found that the classes "relation" (*отношение*), "composition" (*состав*), "foundation" (*основа*), and "characteristic" (*характеристика*) cover the majority of verbal pattern candidates in the gold standard while being rather untypical for the reference data which contains higher percentages of action and process verbs. A comparison of the verb lemmas that were identified in both the gold and the reference data shows that both datasets share a core of predicates, but that the reference data exhibits many more rather infrequent verbs and a higher degree of variation in the use of predicates. Other analyses dealt with *grammatical and structural properties* of KRCs such as verb morphology or the syntactic function of the definiendum, passive use and nominalisations.

References

Baranov, O. S. 1995. *Ideografičeskij slovar' russkogo jazyka.* Moskva: ÈTS.

Barrière, C. 2004. "Knowledge-rich Contexts Discovery". In Tawfik, A.Y. and Goodwin, S.D. (eds.) *Advances in Artificial Intelligence*. (Lecture Notes in Computer Science 3060). Berlin/Heidelberg: Springer.

Bierwisch, M. and Kiefer, F.. 1969. "Remarks on Definitions in Natural Language". In Kiefer, F. (ed) *Studies in Syntax and Semantics.* (Foundations of Language 10). Dordrecht: Reidel.

Cramer, I. M. 2011. *Definitionen in Wörterbuch und Text: Zur manuellen Annotation, korpusgestützten Analyse und automatischen Extraktion definitorischer Textsegmente im Kontext der computergestützten Lexikographie.* Published PhD thesis. Technical University Dortmund. Available online at https://eldorado.tu-dortmund.de/bitstream/2003/27628/1/Dissertation.pdf.

Del Gaudio, R. and Branco, A. 2007. "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". In Neves, J., Santos, M.F. and Machado, J.M. (eds.) *Progress in Artificial Intelligence.* (Lecture Notes in Artificial Intelligence 4874). Berlin: Springer.

Drescher, M. 1992. *Verallgemeinerungen als Verfahren der Textkonstitution: Untersuchungen zu französischen Texten aus mündlicher und schriftlicher Kommunikation".* Stuttgart: Franz Steiner Verlag.

Fahmi, I. and Bouma, G. 2006. “Learning to Identify Definitions using Syntactic Features”. Workshop on Learning Structured Information in Natural Language Applications at EACL 2006, April 3, Trento, Italy: 64-71. Available online at http://ai-nlp.info.uniroma2.it/eacl2006-ws10/WS10-eacl2006-proceedings.pdf.

Fišer, D., Pollak, S. and Vintar, Š. 2010. “Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources”. LREC 2010, May 19-21, Valletta, Malta: 2932-2936. Available online at http://www.lrec-conf.org/proceedings/lrec2010/pdf/141_Paper.pdf.

Halliday, M. 1994. *An Introduction to Functional Grammar*. London: Arnold.

Hamp, B. and Feldweg, H. 1997. “GermaNet – a Lexical-Semantic Net for German”. Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications at ACL 1997, July 12, Madrid, Spain: 9-15. Available online at http://www.aclweb.org/anthology-new/W/W97/W97-0802.pdf.

International Organization for Standardization. 2000. International Standard ISO 1087-1: 2000 – Terminology Work – Vocabulary – Part 1: Theory and application. Geneva: ISO.

Malaisé, V., Zweigenbaum, P. and Bachimont, B. 2005. “Mining defining contexts to help structuring differential ontologies”. *Terminology 11 (1)*: 21-53.

Marshman, E. 2008. “Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts”. *Terminology 14 (1)*: 124-151.

Meyer, I. 2001. “Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework”. In D. Bourigault, C. Jacquemin and M.-C. L’Homme (eds) *Recent Advances in Computational Terminology*. (Natural Language Processing 2). Amsterdam/Philadelphia: John Benjamins.

Meyer, I., Skuce, D., Bowker, L. and Eck, K. 1992. “Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base”. COLING 1992, August 23-28, Nantes, France: 956-960. Available online at http://acl.ldc.upenn.edu/C/C92/C92-3146.pdf.

Muresan, S. and Klavans, J. 2002. “A Method for Automatically Building and Evaluating Dictionary Resources”. LREC 2002, May 29-31, Las Palmas, Spain. Available online at http://www1.cs.columbia.edu/~smara/publications/lrec2002a4.pdf.

Pearson, J. 1998. *Terms in Context*. (Studies in Corpus Linguistics 1). Amsterdam/Philadelphia: John Benjamins.

Przepiórkowski, A., Degórski, Ł., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V. and Wójtowicz, B. 2007. “Towards the automatic extraction of definitions in Slavic”. BSNLP workshop at ACL 2007, June 29, Prague, Czech Republic. Available online at http://www.hum.uu.nl/lt4el/extern/files/SlavicDefinitions.pdf.

Schumann, A.-K. 2011. “A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German”. LTC 2011, November 25-27, Poznan, Poland: 516-520.

Sierra, G., Alarcón, R., Aguilar, C. and C. Bach. 2008. “Definitional verbal patterns for semantic relation extraction”. *Terminology 14 (1)*: 74-98.

Storrer, A. and Wellinghoff, S. 2006. “Automated detection and annotation of term definitions in German text corpora”. LREC 2006, May 24-26, Genoa, Italy: 2373-2376. Available online at http://www.lrec-conf.org/proceedings/lrec2006/.

Walter, S. 2010. *Definitionsextraktion aus Urteilstexten*. Published PhD thesis. Saarland University. Available online at: http://www.coli.uni-saarland.de/~stwa/publications/DissertationStephanWalter.pdf.

Westerhout, E. 2009. “Definition Extraction Using Linguistic and Structural Features”. First Workshop on Definition Extraction, Borovets, Bulgaria: 61-67. Available online at http://www.aclweb.org/anthology-new/W/W09/W09-4410.pdf.

The discursive representation of animals

Alison Sealey
University of Birmingham
a.j.sealey@bham.ac.uk

1 Introduction

A new research project[1] will develop the use of corpus assisted analysis in a particular field of discourse. One strand of this project 'People, Products, Pets and Pets: the discursive representation of animals' is the collection of an extensive corpus of discourse about non-human creatures. Preparation for this new project has involved some piloting of the approach with two contrasting specialised corpora, as well as liaison with non-linguists whose work involves the production and/or interpretation of discourse about animals.

This presentation is about three aspects of this topic: (1) the theoretical context for researching the linguistic representation of animals; (2) findings from a study of texts written by members of the public about the role of animals in their lives; and (3) findings from a corpus analysis of the commentaries accompanying a wildlife documentary television series.

2 The theoretical context

Corpus assisted discourse analysis has been used extensively to investigate: the ways in which people who are perceived as belonging to particular social categories are named; how their attributes and actions are denoted in positive or negative ways; and how labelling patterns and their connotations serve to influence perceptions – of women, migrants or the disabled, for example (Aaron 2010; Caldas-Coulthard and Moon 2010; Gabrielatos and Baker 2008). Meanwhile, as the effects of human activity on the planet become increasingly obvious and alarming, a concentration on humanity alone has for some come to seem rather parochial, and a human-centred view of the world is being challenged from many directions. In the social sciences, Actor Network Theory has encouraged a perception of actors other than human beings (including, for example, sheep and scallops: e.g. Law and Mol 2008; Callon 1986) as making contributions to the social world, while 'post-humanist' and 'post-anthropocentric' theories generally (e.g. Haraway 2008) suggest that the boundaries between humans and other animals have become blurred and uncertain. At the same time, there is pressure to recognise the distinctive characteristics of our own and other species, and to resist anthropomorphising the latter. For corpus linguists with perspectives such as the 'Discourse, Politics and Society' strand at this conference, an obvious response is to analyse corpora comprising texts about the environment, ecology and so on (e.g. Alexander 2009; Goatly 2001, 2002). Some corpus-based analyses have also looked specifically at how animals are denoted and connoted in heterogeneous collections of texts (e.g. Gilquin and Jacobs 2006; Gupta 2006; Stibbe 2012). Corpus assisted analysis highlights parallels between choices about the language used to categorise humans (Sealey 2010, 2012; Sealey & Carter 2004) and that used to denote animals, with the latter raising some new challenges – ontological, epistemological and ethical – which will be explored in this presentation.

3 The Mass Observation Project directive on 'Animals and humans'

The MOP, based at Sussex University in the UK, sends out two or three directives a year, asking correspondents to write in response to a series of questions and prompts. The 'Animals and Humans' directive (August 2009) began with the question, 'what do animals mean to you?' followed by questions about respondents' experience of animals in childhood and in their current daily routines, whether they ate meat or wore clothing made from animal products, how their use of animal products related to the way they felt about animals, and their views on animal welfare, on sport involving animals and on media representations of animals. 244 written responses were received, but the analysis presented here focuses on the 103 which were submitted electronically, as these are more readily analyzable with corpus software. This section of the presentation is concerned with three main questions:

What kinds of animal are referred to across the responses and in what proportions?
Which nouns do these writers use to denote both specific and generic animal kinds, and what criteria underpin the categories they use?
Which naming terms connote evaluative or attitudinal stances towards which kinds of animal?

Various functions within *WordSmith Tools* (Scott 2008) helped to reveal aspects of people's values,

[1] Funded by the Leverhulme Trust for three years from September 2013 at King's College London and the University of Birmingham; investigators Guy Cook and Alison Sealey

attitudes and assumptions as they reported on the role of animals in their lives. Some of the responses included metalinguistic observations (e.g. one writer's definition of a pet was that 'It's named'), while the analysis also highlighted some problems with identifying words in the corpus to be classified as names for animals (e.g. ambiguous terms denoting both animals and their 'products' such as 'lamb', 'chicken', 'fish'). The presentation of this corpus-assisted method of analysing these texts[1] has led to wider interest in the approach both at the Mass Observation Project and among sociologists and political scientists.

4 Linguistic patterns in the wildlife documentary series *Life*

The data for the study reported in this section comprises transcripts of each of the ten episodes in the highly successful wildlife documentary series *Life*, narrated by Sir David Attenborough. The ten episodes are each organized around a particular theme or animal group, and each programme constitutes approximately 45-50 minutes of film. Once transcribed, the texts were converted to text-only format and collated into a digital corpus for analysis with *WordSmith Tools*, providing a small corpus of just under 30,000 words. One approach to the analysis conducted (with co-author Lee Oakley)[2], built on studies of the frequency and discoursal effects of pronoun choice such as that by Gupta (2006: 107), who sees the selection of *who* rather than *which* as a marker of 'a high level of sentience in a nonhuman animal' (see also Gilquin and Jacobs 2006). Use of the gendered pronouns usually confers greater individuality on to living creatures than the neuter *it*, and commentators have observed how such language choices in wildlife documentaries may invoke ideological aspects of human gender roles and sexuality (Chris 2006; Crowther 1999; Crowther and Leith 1995). This part of the presentation discusses some grammatical dimensions of anthropomorphism, including the discursive effect of switches between gendered and neuter pronouns in the corpus, as well as the range of pragmatic functions of *you* in this kind of discourse. The findings of this pilot project have been shared with staff in the BBC's Natural History Unit, who are assisting us in the new research project mentioned above.

[1] Sealey, A. & Charles, N. (in press) 'What do animals mean to you?': naming and relating to non-human animals. *Anthrozoos*.

[2] Sealey, A. & Oakley, L. (in press) Anthropomorphic grammar? Some linguistic patterns in the wildlife documentary series *Life. Text and Talk*.

References

Aaron, J.E. (2010) An awkward companion: disability and the semantic landscape of English *lame*. *Journal of English Linguistics* 38 (1), 25-55.

Alexander, R.J. (2009) *Framing Discourse on the Environment: a critical discourse approach*. New York and London: Routledge.

Caldas-Coulthard, C.R. and Moon, R. (2010) 'Curvy, hunky, kinky': using corpora as tools for critical analysis. *Discourse & Society* 21, 99 – 133.

Callon, M. (1986) Some elements of a sociology of translation: domestication of the scallops and the fishermen of Saint Brieuc Bay. In J. Law (ed.) *Power, Action and Belief: a new sociology of knowledge? (Sociological Review Monograph 32)*. London: Routledge and Kegan Paul.

Chris, C. (2006) *Watching Wildlife*. Minneapolis: University of Minnesota Press.

Crowther, B. (1999) The birds and the bees: narratives of sexuality in television natural history programs. In D. Epstein and J.T. Sears (eds) *A Dangerous Knowing: sexuality, pedagogy and popular culture*: Continuum.

Crowther, B. and Leith, D. (1995) Feminism, language and the rhetoric of television wildlife programmes. In S. Mills (ed.) *Language and Gender: interdisciplinary perspectives*: Longman.

Gabrielatos, C. and Baker, P. (2008) Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English Linguistics* 36 (5), 5 – 38.

Gilquin, G. and Jacobs, G.M. (2006) Elephants who marry mice are very unusual: the use of the relative pronoun who with nonhuman animals. *Society & Animals* 14 (1), 79 – 105.

Goatly, A. (2002) The representation of nature on the BBC World Service. *Text* 22 (1), 1-27.

Gupta, A.F. (2006) Foxes, hounds, and horses: who or which? *Society & Animals* 14 (1), 107 – 128.

Haraway, D. (2008) *When Species Meet*. Minneapolis: University of Minnesota Press.

Law, J. and Mol, A. (2008) The actor-enacted: Cumbrian sheep in 2001. In L. Malafouris and C. Knappett (eds) *Material Agency*: Springer US.

Scott, M. (2008) *WordSmith Tools*. Liverpool: Lexical Analysis Software

Sealey, A. (2010) Probabilities and surprises: a realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics* 31 (2), 215-235.

Sealey, A. (2012) 'I just couldn't do it': representations of constraint in an oral history corpus. *Critical Discourse Studies* 9 (3), 195 – 210.

Sealey, A. and Carter, B. (2001) Social categories and sociolinguistics: applying a realist approach. *International Journal of the Sociology of Language* 152, 1 – 19.

Stibbe, A. (2012) *Animals Erased: discourse, ecology, and reconnection with the natural world*. Middletown, CT: Wesleyan University Press.

Building a corpus of evaluative sentences in multiple domains

Jana Šindlerová
Charles University
sindlerova@ufal.mff.cuni.cz

Kateřina Veselovská
Charles University
veselovska@ufal.mff.cuni.cz

1 Introduction

The area of sentiment analysis (see Liu 2003), or in other words, the automatic extraction of subjective information from a given text, has gained much attention lately, both in the commercial sphere (search for information on customer/consumer attitudes) and in public media (opinion polls vs. web sentiment analyses in politics, cf. recent presidential elections in Czech Republic). As a consequence of an increasing number of user-generated data, we witness a growing need to classify it with respect to the opinion expressed in it. However, to perform this task automatically, we first need to create and annotate a representative corpus of evaluative data and explore it from the linguistic point of view.

In this talk we describe our work on building and annotating corpora intended for the task of sentiment analysis in Czech, and developing classifiers based on this data to prove its credibility. Besides, we present the newly created Czech subjectivity lexicon (see section 3).

2 Building and evaluating the corpora

In our project we have built two plain-text corpora covering two different domains. First, a corpus of news articles was acquired, containing 560,000 words in 1661 articles. A part of this corpus (about 450 articles) has been labeled for being subjective or not. About 410 segments (mostly sentences, but also headlines and subtitles) from the subjective section, i.e. 6,686 words, have been chosen for manual annotation. On this subcorpus, the annotation scheme, based on Wiebe (2002), has been established. Manual annotation has been performed by two annotators. Interannotator agreement has been measured and the analysis of places of disagreement has been made.

During the manual annotation of the news subcorpus, several issues have arisen to be solved in order to increase the interannotator agreement. In accordance with Balahur et al. (2010), we decided not to follow reader's perspective in our future work, but instead to focus on the sentiment content of the text. Moreover, our annotating experiment resulted in a strong need for capturing

additional features, such as keeping a separate category for good/bad news, elusive expressions (subjective, but non-polarized), or false polarity expressions (only seemingly subjective due to a metaphorical transfer of meaning).

The annotation scheme was then enhanced accordingly, and transferred to another acquired corpus, containing amateur movie reviews. To compare the results, we again chose 405 segments, let the same two people annotate them, measured agreement, and finally made another annotator disagreement analysis. For the sake of comparison, we also created a third corpus, containing data from a retail server with explicit prior item evaluation done by internet users (authors of the evaluative commentaries) themselves.

To verify the reliability of the corpora, we decided to train a standard unigram-based Naïve Bayes classifier together with a lexicon-based classifier, and compare their performance with the state-of-the-art results. We have also built a classifier based on the data annotated by the retail server customers to see the difference. The best performance measured by f-score was 0,89, i.e a value close to the state of the art (see Cui et al. 2006).

3 Subjectivity lexicon

In order to improve the task of sentiment analysis in Czech further, we have built a Czech subjectivity lexicon, SubLex 1.0 (Veselovská and Bojar 2012). A subjectivity lexicon is a list of domain-independent evaluative items bearing an inherent positive or negative value (see Wiebe 2004). These expressions can be used as key words in polarity detection task. The Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon [1], using Czech-English parallel corpus CzEng (Bojar and Žabokrtský 2006). After manual refinement it contains 4950 unique lemmas. The evaluative items collected in SubLex1.0 will be compared with the expressions obtained from the annotated data and used for an advanced classification of evaluative corpora.

4 Analysis of the data

In the course of the annotation process, we have realized that since we are dealing with a subjective text, it is necessary to give quite precise instructions about what should be annotated and what should not. For this reason, we have built a database of basic morphosyntactic patterns often used in evaluative sentences. Together with the newly established subjective lexicon SubLex, these two resources build a concise model of evaluative language in Czech sentiment discourse.

5 Conclusion

On the basis of our research, we argue that the number and type of annotated features is dependent on the domain of the annotated text. Moreover, the dependency on text domain can even be seen in the ways the evaluation is expressed, both lexically and structurally. Either way, the presented data have laid foundations for further research into sentiment analysis in Czech. To conclude, using our manually annotated corpora, we helped to improve the task of automatic sentence-level polarity detection.

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B. and Belyaeva, J. 2010. *Sentiment Analysis in the News*, In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Cui, H., Mittal, V. and Datar, M. 2006. *Comparative experiments on sentiment classification for online product reviews*, In proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence.

Liu, B. 2009. "Sentiment Analysis and Subjectivity". Invited Chapter for the *Handbook of Natural Language Processing*, Second Edition. Marcel Dekker, Inc: New York.

Veselovská, K. and Bojar, O. 2012. *SubLex, the Czech subjectivity lexicon*, version 1.0.

Wiebe, J. 2002. *Instructions for Annotating Opinions in Newspaper Articles*, Department of Computer Science Technical Report TR-02-101 , University of Pittsburgh, Pittsburgh, PA.

Wiebe, J., T. Wilson, R. Bruce, M. Bell and Martin, M. 2004. Learning subjective language. *Computational Linguistics, 30*, 3.

[1] http://www.cs.pitt.edu/mpqa/subj_lexicon.html

Lexical, corpus-methodological and lexicographic approaches to paronyms

Petra Storjohann
Institut für Deutsche Sprache Mannheim
storjohann@ids-mannheim.de

1 Introduction

German lexical items with related morphological roots and similar semantic potential are easily confused and misused both by native speakers and learners, respectively. Examples of so-called paronyms include *effektiv – effizient – effektvoll*, *scheinbar – anscheinend*, *formell – formal*. These are generally not regarded as synonyms. However, first empirical studies suggest that in some cases items of a paronym set have undergone meaning change and developed synonymous notions. In other cases, they remain similar in meaning, but can show subtle differences in definition and restrictions of usage.

Whereas the treatment of synonyms has received attention from corpus-linguists (cf. Partington 1998, Taylor 2003), the subject of paronyms has unfortunately not been revisited with empirical, data-driven methods neither in terms of semantic theory nor in terms of practical lexicography. Lexicographically, some German paronyms have been documented in an outdated printed dictionary (Müller 1973). However, there is no corpus-guided reference guide describing paronym sets empirically enabling readers to find the correct usage of such lexical items. Overall, paronymy needs to be addressed from new perspectives. First, the phenomenon has not been accounted for comprehensively in linguistic theory. Secondly, from a corpus linguistic view, we need to search for suitable corpus methods for detailed semantic investigation. Finally, solutions to some lexicographic challenges are required.

2 Linguistic treatment of paronyms

As Hausmann (1990) points out the subject of paronymy has mainly been approached linguistically from typological, language contrastive perspectives, particularly in the field of translation studies. However, we lack an empirical treatment and theoretical account of paronymy as a lexical phenomenon in general (cf. Làzàrescu 1995). Hence, there are no widely tested methods that proved suitable for semantic analysis of such words.

To be able to derive conclusion and to develop hypotheses, it is suggested to work with corpus-driven corpus procedures to examine paronyms closer. Empirically, corpus-driven investigations of paronyms can provide valuable insights into principles of language change in semantically related lexical items.

3 Corpus-linguistic approaches to paronyms

The examination of paronym sets necessarily incorporates contrastive meaning analyses. Methodologically, it is advantageous to use corpus tools that are not only able to analyse patterns by exploring co-occurrences. They should also be capable of measuring semantic similarity or distance by contrasting collocation profiles pairwise to systematically detect differences in terms of contextual behaviour.

One possibility could be the visual representation of topographic profiles of the involved lexical items and the comparison of those with self-organising feature maps (cf. Kohonen 1990; Keibel and Belica 2007) in order to contrast paronyms. Topographic profiles break down unstructured collocation patterns and hence complex semantic properties (see Figure 1).

Figure 4. Topographic profile of German *effektiv*

Furthermore, self-organisation maps can be used to contrast patterns of usage between two lexical items by comparing them with words which exhibit collocation profiles that are most similar to the two items in question (see Figure 2). This procedure referred to as CNS-model (Contrasting Near-Synonyms) has been developed and implemented in a German linguistic work bench (CCDB: co-occurrence database) by Belica (2001 ff).

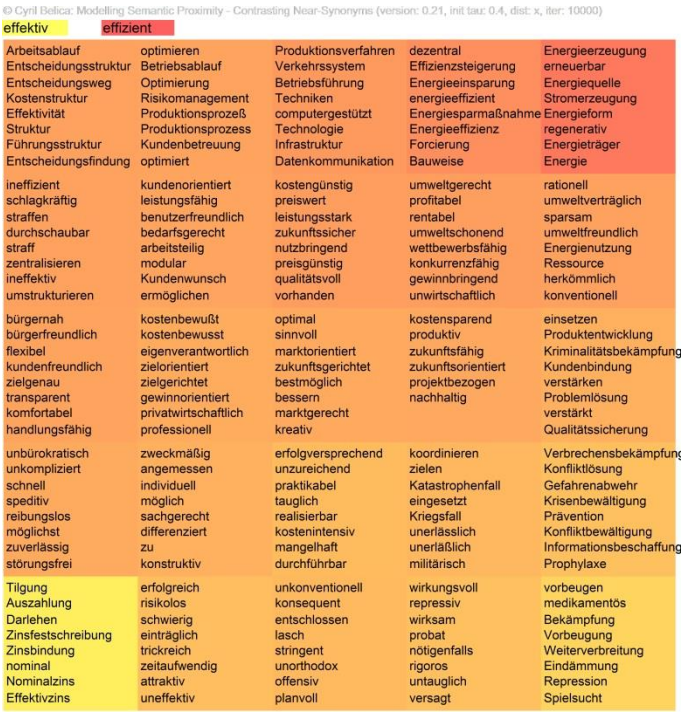


Figure 5. Contrasting German *effektiv –effizient* with SOM

The CCDB is used "for the study, development, and evaluation of methods for the data-driven exploration and modelling of language use" (Keibel and Belica 2007). SOMs arrange lexical items in two-dimensional lattices such that proximity on the grid reflects semantic similarity between collocation profiles. As suggested by Vachková and Belica (2009), this approach to collocational patterning might be applicable for lexicographic investigations of synonyms. The semantic properties of near-synonyms are contrasted with each other. Markova (2012), for example, puts forward examples of synonyms set which she investigated with the CNS-model successfully.

Consultations and interpretations of self-organisation feature maps might be a suitable approach to the analysis and semantic description of paronyms sets. It is argued that it might also be a practical corpus procedure for the examination of paronyms where usage aspects that are shared and not shared between easily confused words are detected.

4 Lexicographic challenges

From a lexicographic point of view, a number of challenges are encountered when documenting usage-based findings in a dictionary of paronyms which users generally expect to be rather prescriptive and where they demand definite answers for doubtful language situations. One central problem regards the interpretation and documentation of language change and normative restrictions. This is particularly relevant for pairs that are recorded as semantically distinct lexical items in traditional reference works and that have assimilated semantically over time due to common, allegedly "false" use. This assimilation process will have developed to different degrees among different paronym pairs/sets. In some cases, corpus analyses signal tendencies that paronyms might or might not possibly turn into synonyms.

Therefore, one of the major challenges of a corpus-based paronym dictionary is the lexicographic interpretation of ambiguous data, especially paronym usage with a similar proportion between contexts with clear semantic difference between the terms and contexts exhibiting synonymous use. The lexicographic interpretation of such data requires a certain sensibility, as a specific conflict is expected to be encountered with corpus data. One the one hand, false language use caused by confusing paronyms needs prescriptive correction. On the other hand, gradual language change caused by frequent misuse of a certain lexical item needs descriptive documentation of contemporary language use.

References

Belica, C. 1995. *Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode*. Institut für Deutsche Sprache: Mannheim.

Belica, C. 2001ff. *Kookkurrenzdatenbank CCDB – V3.3. Eine korpuslinguistische Denk- und Experimentierplattform*. Institut für Deutsche Sprache: Mannheim. Available online at http://corpora.ids-mannheim.de/ccdb/

Hausmann, F.J. 1990. "Das Wörterbuch der Homonyme, Homophone und Paronyme". In F.J. Hausmann, O. Reichmann and H.E. Wiegand (eds.) *Wörterbücher. Dictionaries. Dictionnaires*. vol. 2. Berlin/New York: de Gruyter, pp. 1120-1125.

Keibel, H. and Belica, C. 2007. "CCDB. A Corpus-Linguistic Research and Development Workbench". *Proceedings of the 4th Corpus Linguistics Conference* (CL 2007), Birmingham.

Kohonen, T. 1990. *The Self-Organizing Map. New Concepts in Computer Science*. Proc. Symp. in Honour of Jean-Claude Simon, Paris. AFCET, p.181-190.

Làzàrescu, I. 1995. "Deutsche Paronyme". *Grazer Linguistische Studien* 43, S. 85-93.

Müller, W. 1973. *Leicht verwechselbare Wörter. Duden Taschenwörterbücher Vol. 17*. Mannheim: Bibliographisches Institut.

Partington, A. S. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Vachková, M. and Belica, C. 2009. "Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography". *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* 13/2: 223-260.

Taylor, J. 2003. "Near synonyms as coextensive categories:'Tall' and 'high' revisited". *Languages Science*, 25: 263-284.

Verbs with a sentential subject: A corpus-based study of German and Polish verbs

Janusz Taborek
Adam Mickiewicz University
taborek@amu.edu.pl

The purpose of this paper is to present a classification of verbs with a sentential subject in German and Polish. The study is grounded on a distributional and syntactic investigation of the electronic corpus data for German (DeReKo and DWDS) and Polish (IPI PAN and PWN). Based on corpus data, this work assures authentic examples of language use. The theoretical basis is the concept of a lexicon-grammar developed for sentential subjects of English applied by Gross (1994) and Salkoff (2002) and the idea of verb classification depending on its complementation clauses like Biber et al. (1999).

Verbs whose subject can be represented by a sentential clause allow one type of subject clause and do not allow another. For example, the German verb *wundern* 'to surprise' occurs with clauses introduced by *dass* 'that', *w*-word 'wh', *wenn* 'if/when', *wie* 'how' and infinitival clause, but on the other hand it does not allow, say, clauses introduced by *ob* 'whether'.

(1) Mich **wundert**, **dass** er Klavier spielt.
'It surprises me that he plays piano.'

(2) *Mich **wundert**, **ob** er Klavier spielt.
'It surprises me whether he plays piano.'

On the other hand, the verb *freuen* 'to make happy' only allows a *dass*-clause, infinitive, *wenn*-clause and *wie*-clause as a subject, hence the conclusion: the verbs *wundern* and *freuen* represent different (distributional) classes of verbs. According to Biber et al. (1999: 577) verbs are grouped into distributional classes depending on types of subject clauses. The analysis contains 205 German and 170 Polish verbs. Analogically to some previous studies such as those of IdS-Grammatik (Zifonun et al., 1997) for German or Biber et al. (1999) for English, this classification follows the possibility of taking one of the main types of subject clauses, that is: (1) clauses introduced by *dass* 'that' in German or by *że* 'that' (also: *aby*, *ażeby*, *by*, *iżby*) in Polish, (2) indirect questions with clauses introduced by *ob* 'whether' in German and *czy* 'whether' in Polish, (3) infinitive clauses (this group contains also Polish infinitive clauses introduced by *by*). These main classes are than divided into subclasses and the criterion for this is the appearance with other

subject clauses, e. g. clauses introduced by *wenn/kiedy* ‘if/when’, *wie/jak* ‘how’, *bis/że* ‘until’, *als/jakoby* ‘as’ and also finite clauses without introducing and with the finite verb coming second, the so-called “Verbzweitsatz”.

There are seven potential main classes of verbs established by its selection of main subject clause types and several subclasses, if we consider all types of subject clauses. Tables 1 and 2 present the main classes and subclasses of German verbs with sentential subjects. We classified Polish verbs in the same way.

Cl.	Examples	*dass*	*w*	Inf
1	*interessieren* ‘to interest‘	+	+	+
2	*auffalen* ‘to be noticed’	+	+	–
3	*gefallen* ‘to be pleased’	+	–	+
4	*geschehen* ‘to happen’	+	–	–
5	*gelingen* ‘to succeed’	–	–	+
6	*sich fragen* ‘to be a question‘	–	+	–
7	*freistehen* ‘to be free‘	±	+	+

Table 1. Main classes of German verbs with sentential subject (Taborek, 2008)

The next aim of this paper is to check whether and how verbs with sentential subjects are represented in two German-Polish and Polish-German dictionaries. It is to be expected, that the most frequent types of subject clause will be included as examples in the dictionary entry.

The first analyzed verb *gefallen* ‘to be pleased’ selects the following clauses as a sentential subject: *dass*-clause, infinitival clause, *wenn*-clause and *wie*-clause and it belongs to class 3 of our classification.

Investigation using Dereko/Cosmas II shows that 50.3% of subject clauses selected by *gefallen* are *dass*-clauses, 20.0% embedded infinitival clauses, 17.0% *wie*-clauses and 12.6% *wenn*-clauses (cf. table 3).

However, in monolingual dictionaries of German (DUW and WDW) only infinitival clauses are noted and no examples using the most frequent dass-clause are given. An example of a sentential subject for gefallen with dass-clause is given in the PONS German-Polish dictionary (3) but not in the PWN German-Polish dictionary.

(3) **gefallen** […] **es gefällt mir gar nicht, dass**... nie podoba mi się, że (PONS)

On the other hand, the PWN dictionary presents an example of use of the Polish verb *podobać się* ‘to be pleased’ with a subject clause (4) whereas there is no example of this in the PONS dictionary.

(4) **podobać się** […] **~ło jej się, że chcieli pomagać starszym** ihr gefiel es, dass sie älteren Menschen helfen wollten (PWN)

The next verb is the event verb *geschehen/zdarzyć się* ‘to happen, to come about’, which represents class 4a of verbs with sentential subjects and is used as in (5).

(5) It **happened that** the father of Publius lay sick with fever and dysentery. (Acts 28:8)

Es **geschah** aber, **daß** der Vater des Publius am Fieber und an der Ruhr lag. (Luther 1912)

I **stało się, że** ojciec onego Publijusza, mając gorączkę i biegunkę, leżał. (Biblia Gdanska)

The event verb *geschehen* can be used only with a subject clause introduced by *dass* ‘that’ and this possibility is represented in monolingual dictionaries of German but is not represented in either of the new German-Polish dictionaries (PONS and PWN). Furthermore, the Polish verb *zdarzyć się* ‘to happen’ selects the assertive clause *że* ‘that’ and also *by* ‘that’, which is used for modals and negation. The Polish verb *zdarzyć się* can also select an infinitival clause if it is used with an experiencer (like in English *it happened to me that*).

The third example is the evaluative verb *gelin*

Class	Examples	*dass*	*w*	*ob*	Inf	*wenn*	*wie*	V2	*als*
1a	*interessieren* ‘to interest‘	+	+	+	+	+	+	–	–
1b	*feststehen ‘to be fixed’*	+	+	+	+	–	+	–	–
1c	*wundern* ‘to be surprized‘	+	+	–	+	–	+	–	–
2a	*entfallen* ‘to escape’	+	+	+	–	–	+	±	–
2b	*einleuchten* ‘to be clear’	+	+	–	–	–	+	+	–
2c	*auffalen* ‘to be noticed’	+	+	+	–	+	+	±	–
3a	...								

Table 2. Classes 1 and 2 of German verbs with sentential subject in detail (Taborek 2008)

Class	Examples	*dass*	*w*	*ob*	Inf	*wenn*	*wie*	V2	*als*
3a	*gefallen* ‘to be pleased’	50.3	–	–	20.0	12.6	17.0	–	–
4a	*geschehen* ‘to happen’	100.0	–	–	–	–	–	–	–
5a	*gelingen* ‘to succeed’	1.9	–	–	98.1	–	–	–	–

Table 3. Types of subject clauses of German verbs *gefallen*, *geschehen* and *gelingen*.

gen/udawać się ‘to succeed’. It prefers to be used with an infinitive clause (98.1%) and only occasionally with a that-clause (*dass*, 1.9%). The Polish verb *udawać się* ‘to succeed’ selects an infinitive

clause as noted in (6) and (7) and occasionally the modal *by* ‘that’.

(6) **es gelang mir nicht, sie davon abzuhalten** nie zdołałem jej od tego a. przed tym powstrzymać (PWN)

(7) **jdm gelingt es etw zu tun** komuś udaje się coś zrobić (PONS)

In conclusion, our results demonstrate the variety and complexity of sentential subject necessary to explain the presentation of verbs in mono- and bilingual dictionaries. Furthermore, the analysis shows how grammatical and lexical information extracted from corpora can enrich lexicographical knowledge and can form the first step to a planned corpus-based contrastive German-Polish grammar (cf. Schmied 2009 for German-English).

That the use of corpora can enrich the lexicographical work is shown by the Polish-English dictionary (PWN/Oxford), which contains examples of subject clauses of all the verbs analysed in this paper – *podobać się* ‘to be pleased’ (8), *zdarzyć się* ‘to happen’ (9) and *udać się* ‘to succeed’(10).

(8) **[podoba]ło jej się, że chcieli pomagać starszym** she was pleased that they wanted to help the elderly

(9) **może się [zdarzyć], że...** it may happen that...

(10) **[uda]ło mi się znaleźć pracę** I managed to find a job; **[uda]ło mu się wszystkich przekonać** he succeeded in convincing everybody

References

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Edinburgh: Longman.

Gross, M. 1994. “Constructing Lexicon-Grammars”. In B. T. S. Atkins and A. Zampolli, Antonio (eds.), *Computational Approaches to the Lexicon*, Oxford: OUP, 213-263

Hunston, S. and Francis, G. 2000. *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Salkoff, M. 2002. Verbs with a sentential subject. A lexical examination of a subset of psych verb. *Lingvisticæ Investigationes* 25(1): 97-147.

Schmied, J. 2009. “Contrastive corpus studies”. In A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Vol. 2. Berlin and New York: Mouton de Gruyter.

Taborek, J. 2008. *Subjektsätze im Deutschen und im Polnischen. Syntaktisches Lexikon und Subklassifizierung der Verben*. Frankfurt a.M.: Peter Lang Verlag.

Taborek, J. 2011. “Überlegungen zur korpusbasierten kontrastiven Grammatik. Dargestellt am Beispiel der Distribution der Subjektsätze mit dass im Deutschen und ihrer polnischen Entsprechungen”. In M. Kotin and E. Kotorova (eds.) *Geschichte und Typologie der Sprachsysteme*. Heidelberg: Winter Verlag, 317-325.

Zifonun, G., Hoffmann, L. and Strecker, B. 1997. *Grammatik der deutschen Sprache*. Berlin and New York: Walter de Gruyter Verlag.

Dictionaries

DUW = Drosdowski, G. (ed.) 2006. *Duden. Deutsches Universalwörterbuch*. Mannheim: Bibliographisches Institut.

PONS = Dargacz, A. (ed.) 2008. *PONS. Wielki słownik niemiecko-polski*. Poznań: Lektor-Klett.

PWN = Wiktorowicz, J. and Frączek, A. (eds.) 2008. *Wielki słownik polsko-niemiecki*. Warszawa: Wydawnictwo Naukowe PWN.

WSW = Wahrig-Burfeind, R. (ed.) 2008. *Wahrig. Deutsches Wörterbuch*. Gütersloh and München: Wissen Media Verlag.

"Criterial feature" extraction from CEFR-based corpora: Methods and techniques

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

1 Introduction

In the field of corpus applications in foreign language teaching and second language acquisition research, there is a growing interest in using corpora for identifying "criterial features" for L2 target language. The notion of "criterial features" came from a research team involved in the English Profile programme (http://www.englishprofile.org). Their primary purpose is to provide a set of language inventory which serves as criteria for L2 proficiency. The need for this kind of inventory stems from the fact that the Common European Framework of Reference for Languages (CEFR), a language-independent reference framework for foreign language proficiency, needs to have a set of specifications of lexis and grammar when it is applied to a particular language. There are a series of reports from the English Profile team in the form of the *English Profile Journal* and the *English Profile Series* (e.g. Hawkins & Filipović 2012; Green 2012), in which lists of English grammar items and vocabulary are provided.

Tono (2012a) argued that the vocabulary profile based on the Cambridge Learner Corpus might cause the resulting wordlist to be skewed toward productive knowledge of vocabulary and proposed the wordlist should have a distinction of receptive vs. productive vocabulary. Tono (2012b) also suggested that since the data used for the English Profile only contains the texts classified beyond A1 level, there is a lack of evidence for properly identifying A1 level features.

In this paper, I will report on the on-going project on systematic extraction of criterial features from multiple source CEFR-based corpora. First, a brief description of the project and the design of several different corpora newly compiled for the project will be given, followed by methodological issues regarding how to extract criterial features from CEFR-based corpora using machine learning techniques.

2 The CEFR-J and Reference Level Descriptions

The project aims to support the implementation of the CEFR-J, an adaptation of the CEFR into English language teaching in Japan (Tono & Negishi 2012). After the release of Version 1 of the CEFR-J in March, 2012, we launched a new government-funded project called the "CEFR-J Reference Level Description (CEFR-J RLD)" Project. RLD is a term used for the CEFR to prepare an inventory of language (lexis and grammar) for each individual language for the purpose of level specification.

Table 1 shows the list of corpora to be used for the project:

Type of Corpora	Name	Features
Input corpus	ELT materials corpus (to be completed)	ELT coursebooks Major textbooks that claim to be CEFR-based
Interaction corpus	Classroom observation data	30 hours secondary school ELT classes
Output corpus	JEFLL Corpus (0.7 million)	Written, secondary school, CEFR level
	NICT JLE Corpus (2 million)	Spoken, interview test scripts, 1,280 participants, CEFR level
	ICCI (0.6 million)	Written, primary & secondary school, 9000 samples, CEFR level
	GTECfS Corpus (to be comleted)	Written, exam scripts, 30,000 samples, CEFR level
	MEXT Corpus (S: 8,000 words) (W:3,0000 words)	S/W 2000 students randomly selected from all over Japan

Table1: Corpora used for the project

Three types of corpora have been either newly compiled or re-organised: input, interaction, and output corpora. For input corpora, major ELT publishers' CEFR-based course materials have been scanned and processed by OCR. For output corpora, major learner corpora for Japanese EFL learners, the JEFLL Corpus and the NICT JLE Corpus, have been selected, but for our project, the essays originally classified according to the school grades or oral proficiency test scores, have been re-classified according to the estimated CEFR levels assigned by trained raters based on their holistic scorings. Two additional corpora have been made available. One is an exam-based

corpus called the GTEC for STUDENTS Writing Corpus, provided by the Benesse Corporation. It consists of more than 30,000 students essay data with approximately 5,000 samples aligned with correction data. The other is the data collected by Ministry of Education (MEXT), in which more than 2,000 students were randomly selected from all over Japan. They were given written and oral proficiency exams in English. This data shows the average performance of EFL learners in Japan, after the three year instructions in secondary school.

Finally, a corpus of classroom interaction between teachers and students has been added to the resource. This is an on-going project and the size is relatively small, but I hope that it will shed light on the understanding of what is happening in the classroom.

Our aim is to identify criterial features by looking at input and output corpora across CEFR levels. The language presented in the input corpora may not be produced in the output corpora. By examining both input and output, descriptions of criterial features will become more systematic. The interaction corpus also helps better understand the learning/acquisition process in the classroom. Input from textbooks as well as input and interactions in the actual classroom will play an important role in learning a target language. The major goal is to find out criterial features for the levels specified in the CEFR-J and complete the inventory of grammar and vocabulary for teaching and assessment, with a special reference to teaching and learning contexts in Japan.

3 Issues of criterial feature extraction

In the past few years, various linguistic criteria have been proposed as "criterial", but they need to be validated against a particular learner group like Japanese EFL learners because the data used in Europe are very different from our learner group. Also each proposed criterial feature should be evaluated and weighed in terms of usefulness as CEFR-level "classifiers". Then a bundle of criterial features have to be tested and validated to find out which combinations of criterial features work best to predict the CEFR-levels. In a way, for assessment purposes, it is sufficient to identify the most salient criterial feature that can distinguish all the levels clearly. For teaching purposes, however, all the learning items need to be somehow evaluated against their 'criteriality.'

There are various ways of extracting criterial features from the data. Machine learning techniques such as *random forest* seem to be very promising for this purpose. For instance, random forest is very useful in that it gives estimates of what variables are important in the classification. Table 2 shows the results of variable importance measure by Gini impurity criterion. Basically, the higher the score is, the more important the variable is. By using this kind of information, one can profile which linguistic feature will be most effective in classifying texts into CEFR levels. The major aim of the project is to decide on which machine learning algorithms to take, and evaluate a range of criterial features for its effectiveness as assessment and teaching points.

Linguistic features	Mean Decrease of Gini
Total n. of words	440.3
Total n. of sentences	134.8
N. of VPs	277.2
N. of clauses	182.4
N. of T-units	121.3
N. of dependent clauses	102.6
N. of complex T-units	114.6
N. of complex nominals	210.2

Table2: Variable importance measured by Mean Decrease of Gini

In this paper, I will report on the performance of different machine learning techniques, including random forest, support vector machine, decision tree (C4.5), and naïve Bayes over CEFR-level classified texts and compare which programs produce the best result and useful additional information to evaluate the importance of criterial features. Also I will propose a methodological framework in which extraction of criterial features is to be done strategically, (1) by using different types of corpora, e.g. input and output corpora, for different aspects of feature extraction, and (2) by using different types of statistical tools for different purposes (data mining or classification tasks).

References

Hawkins, J.A. & Filipović, L. (2012). *Criterial Features in L2 English*. Cambridge: Cambridge University Press.

Tono, Y. 2012a. Developing corpus-based word lists for English language learning and teaching: A critical appraisal of the English Vocabulary Profile. In J. Thomas & A. Boulton (eds). *Input, Process and Product: Developments in Teaching and Language Corpora* (pp.314-328). Brno: Masaryk University Press.

Tono, Y. 2012b. International Corpus of Crosslinguistic Interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In Y. Tono, Y. Kawaguchi & M. Minegishi (eds.) *Developmental and Crosslinguistic*

Perspectives in Learner Corpus Research (pp.27-46). Amsterdam: John Benjamins.

Tono, Y. & Negishi, M. 2012. The CEFR-J: Adapting the CEFR for English language teaching in Japan. *JALT Framework & Language Portfolio SIG Newsletter* No.8 (September, 2012), pp. 5-12.

Reflexivity of high explicitness metatext in L1 and FL research articles from the Soft and Hard Sciences: A corpus-based study

Naouel Toumi
University of Monastir
touminanou@yahoo.fr

1 Introduction

Nowadays, most scholars looking for promotion and other academic rewards must publish in impacted journals that use English. However, writing in this language is difficult for Non Native English academics who use other languages in their national disciplinary context. In such contexts, Non Native English researchers may benefit from comparative studies of the local and the native English writing conventions. Results from these studies can help Non Native English academics write more felicitous research texts in English.

The present study comparatively analyses the use of Reflexivity of High Explicitness Metatext (RHEM) – as a subcategory of Reflexive Metadiscourse – in research articles (RAs) written in English by natives and Tunisian researchers. Here, *Reflexive Metadiscourse* is defined as the cover term for the self-reflexive expressions used by writers to negotiate meaning in their texts. It is the writers' explicit commentaries on their own ongoing texts. Reflexive Metadiscourse is divided into three broad categories: RHEM, Reflexivity of Low Explicitness Metatext and Writer/Reader-Oriented Metadiscourse. RHEM (the focus of this paper) includes expressions which explicitly refer to the text, its writing, its organisation and language. These come in three subcategories: (a) REFERENCE TO THE TEXT (such as reference to the whole text, reference to part of the text, adverbs *here* and *now*), (b) DISCOURSE LABELS (such as: illocutionary verbs and nouns, illocutionary verbs + adverbs, text reflexive adverbs of manner), and (c) PHORIC MARKERS (such as: topic shift markers, previews and reviews).

Although metadiscourse has been studied in different languages in addition to English, in different spoken and written genres, and in different academic disciplines, there is no previous systematic research on the textual distribution of metadiscourse, and particularly in the different sections of the RA.

The non-integrative model of metadiscourse

(such as the one adopted by this study) has been little used in the analysis of RAs. Although Mauranen (1993) used a non-integrative approach in the analysis of metatext in economics RAs, her model is less comprehensive than the one proposed in this study. Pérez-Llantada (2010) also employs a non-integrative model in biomedical RAs but with a focus on the discourse functions of metadiscourse, rather than its manifestations. The latter study is also limited to two RA sections: Introductions and Discussions.

To my knowledge, no previous studies of reflexive metadiscourse have focussed on the academic writing of Tunisian writers as representatives of a hybrid linguistic situation. In fact, academics in Tunisia speak Tunisian, a non-coded variety, as L1. They use Modern Standard Arabic at early school stages, and for further studies in the Humanities. They also have to master French as a medium of instruction in the hard and most of the soft sciences. This language is also needed for local publication, and for publication in and academic communication with Francophone countries. English, on the other hand, is necessary for international publication and academic communication.

In this regard, the present study intends to help fill in some of these research gaps by comparatively analysing RHEM's use in Native English Anglo-American (AA) and Tunisian (TU) RAs. The study hopes to answer three research questions:

RQ1: Is RHEM a marker of disciplinary communities (here, soft and hard Sciences)? If yes, do TU authors respect these tendencies?
RQ2: Do certain sections of RAs use different amounts of RHEM? If yes, do TU authors respect these tendencies?
RQ3: Do empirical and theoretical research articles use RHEM differently, in terms of frequency and type? If yes, do TU writers respect this tendency?

2 Methods

To answer the research questions, the present analysis investigates a corpus of 100 RAs (524,526 words) from the hard and soft sciences, with 50 RAs from each cultural group. The focus of this work is on Economics, Business and Management RAs as samples of the soft sciences (25 RAs from each cultural group) and Earth and Planetary Sciences RAs as samples of the hard sciences (25 RAs from each cultural group).

This work uses an analytical model which draws its main components from Mauranen (1993) and Adel (2006) with some adjustments. These modifications are made in order to render the model more applicable to the research article genre. This framework allows to investigate an under-researched aspect of metadiscourse, namely *Reflexive Metadiscourse*. The model also overcomes one of the major weaknesses of the earlier frameworks by providing clear criteria for the identification of reflexive metadiscourse.

This study employs corpus-based approaches to language description. It uses both manual and computer-assisted (three software packages) methods to identify and count the metadiscursive instances. The analytical procedure follows different steps: text annotation, concordancing and extraction. *Note Tab* is used in the manual coding process which consists in reading the corpus in order to identify and tag the RHEM units in their contexts. *AntConc* is used to serve both qualitative and quantitative purposes. Using the tags as search terms, *AntConc* makes it possible to calculate and compare the occurrences of each type of RHEM in the two cultural groups, in the soft and in the hard sciences RAs. In another step, *XTractor* (version 0.9), an extraction software package, is used for the comparison of RHEM units across the RAs sections.

In order to obtain a more comprehensive analysis, the corpus analysis is supplemented with ethnographic tools, i.e., questionnaires and interviews. A set of 60 questionnaires are sent by email to writers whose addresses appeared in the analysed RAs. As for interviews, three specialists from each of the fields investigated are interviewed –as community representatives (Hyland 2001)– about the use of the linguistic aspect in focus.

3 Results and implications

The analysis reveals that the AA writers use more markers of total RHEM than their TU counterparts. It is also found that the AA writers employ more RHEM markers in the soft than in the hard sciences contrary to the TU writers whose use of RHEM is higher in the hard sciences. These differences apply not only to frequency but also to type. The TU authors employ more REFERENCE TO PART OF THE TEXT (here mostly tables and figures) + REFLEXIVE VERB than REFERENCE TO THE WHOLE TEXT, ILLOCUTIONARY VERBS and ILLOCUTIONARY NOUNS which are more frequent in the AA texts.

RHEM is most frequently encountered in the argumentative sections of the AA RAs followed by the opening and closing sections, however it is much less frequent in the expository sections. While the TU writers show agreement with their

AA counterparts in this distribution, they record much smaller frequencies than those encountered in the same sections of the AA RAs.

This work also proves that the use of RHEM in theoretical RAs outnumbers that in empirical IMRD RAs, and thus it is more needed in the former than in the latter RAs. In the two cultural groups, the highest frequencies are encountered in the soft science theoretical RAs. It is also found that the AA writers use more RHEM in their theoretical RAs than their Tunisian counterparts in the equivalent RAs.

The present work contributes to EAP teaching by providing more accurate guidance to Tunisian EAP course designers and/or instructors about the use of RHEM in English RAs. It also suggests a corpus-based teaching method of RHEM to tertiary level EAP students.

References

Adel, A. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamin publishing company.

Hyland, K. 2001. "Humble servants of the discipline? Self-mention in research articles". *English for Specific Purposes* 20 (1): 207-226.

Mauranen, A. 1993. *Cultural differences in academic rhetoric: A text-linguistic study.* Frankfurt am Mein: Peter lang.

Pérez Llantada, C. 2010. "The discourse functions of metadiscourse in published academic writing: Issues of culture and language". *Nordic Journal of English Studies* 9 (2): 41-68.

Instrumental and integrative approaches to language in Canada: A cross-linguistic corpus-assisted discourse study of Canadian language ideologies

Rachelle Vessey
Queen Mary, University of London
r.vessey@qmul.ac.uk

1 Introduction

Although Canada is an officially bilingual country, the idea that it consists of "two solitudes" (MacLennan 1945), according to which the two dominant (English and French) linguistic groups live in separate worlds with little interaction or communication, has received attention in sociolinguistic circles (e.g. Heller 1999). This paper examines Canada's "two solitudes" by comparing language ideologies in English and French Canadian newspapers using cross-linguistic corpus-assisted discourse studies. More specifically, this paper compares how the English and French languages are represented differently and serve different purposes in English- and French-speaking society. Findings from this study suggest that while English newspapers tend to represent languages in instrumental terms, French newspapers predominantly evoke the integrative nature of the French language. A case study of representations of language education is used to illustrate this argument.

2 Data and methods

The data consist of corpora of 7.5 million words in English and 3.5 million words in French. The data are drawn from the most widely-circulated French and English newspapers across Canada and are analysed using cross-linguistic corpus-assisted discourse studies. These are methods combining multilingual approaches to corpus linguistics and discourse analysis. The objective of this combination is to explore, compare, and contrast discursive representations in large corpora of different languages (Vessey forthcoming). The French and English corpora are analysed using frequency, concordance, and keyword procedures to establish the salience and significance of trends; the concordance lines are also analysed using critical discourse analysis. In addition, entire articles are sampled from the corpora for in-depth critical discourse analysis; sampling was done by establishing which

individual articles contained high and low concentrations of core query terms (i.e. concentrations of words per million of ENGLISH/ ANGLAIS, FRENCH/ FRANÇAIS, LANGUAGE/ LANGUE). The theoretical framework of language ideologies (e.g. Woolard 1998; Milani and Johnson 2008) is adopted to explain how language, identity, nationhood, and the state become interconnected in the social imaginary and represented in discourse.

3 Context and findings

This study's findings suggest the predominantly *instrumental* role that languages play in English-speaking Canada and the predominantly *integrative* role that the French language has in French-speaking Canada. A case study focusing on language education serves as an illustration of these different approaches to languages. For example, not only does the English corpus contain far more education-related keywords (e.g. SCHOOL, STUDENTS, EDUCATION, CLASSES) than the French corpus, but also top-ranked non-education-related keywords such as LANGUAGE, FRENCH, and ENGLISH tend to have high numbers of education-related collocates (e.g. LANGUAGE [N=671] collocates with SCHOOL [N=20], LEARNING [N=11], TEACHING [N=6], among many others). Furthermore, of the four English articles that were downsampled for in-depth critical discourse analysis (see above), two are focused specifically on language education. Such is not the case in the French corpus, where few keywords pertain to education, few top-ranked keywords collocate with education-related words, and none of the four downsampled French articles focus on education. More importantly, languages are discussed more frequently in the French corpus than the English corpus (16% of the French articles contain references to language as compared with 8% of English articles), and the French language is more topical than the English language in the English corpus (as compared with French being the topical language in the French corpus); these additional contextual factors reinforce the pertinence of discussions of (foreign – French) language education in the English corpus.

The focus on education in the English corpus fits within broader corpus evidence of languages being represented as commodities. For English speakers, languages – and French in particular – have only attained real commodity value since the country was made officially bilingual in the 1960s; language policies of that era made bilinguals the Canadian social elite. Accordingly, education has become a means for all Canadians to have equal access to the commodity value of language. Frequency, concordance lines, and downsampled articles from the English corpus all demonstrate that education continues to be perceived as a highly valued democratic process through which all Canadians should be able to obtain coveted social resources, especially fluency in the other official language. In contrast, the comparative non-focus on education in the French corpus fits within the broader corpus evidence of languages *not* being represented as commodities. Concordance lines and downsampled articles from the French corpus strongly indicate that the French language – which is the most frequently discussed language – is represented as having integrative value as an important national identity marker. This is perhaps because the aforementioned Canadian language policies did not serve to encourage French speakers to learn the other official language; rather, policies empowered French speakers by enabling them reinforce their own linguistic community and to avoid assimilation into the dominant English-speaking country. Accordingly, the French newspapers do not represent access to foreign (i.e., English) language education as central to equality in democratic society, and the low frequency of references to French language education suggests that it is largely taken for granted – except in the case of immigrants, who are strongly asserted to require fluency in French in order to integrate into society. As a result, instrumental and integrative approaches to language manifest themselves differently in newspapers, which can be seen in the different representations of language education.

4 Implications

These findings with regard to education suggest that although Canada's pioneering efforts in language education (e.g. in the form of French immersion) are often considered a hallmark of positive efforts to bridge the "two solitudes", divergent approaches to language education may be a diagnostic of the different understandings of language in English- and French-speaking Canada that are at the very foundation of the "solitudes". Specifically, French language education has been central to English Canadian engagement with bilingualism; however, this kind of focus continues to re-assert that languages have instrumental rather than integrative roles in society. Such an instrumental approach misses out on the more fundamentally important integrative role that the French language plays for French speakers in Canada. Indeed, it has been the integrative (and not instrumental) role of the

French language that has been the driving force for French-speaking Canadians to preserve their language and culture over the past four centuries. It is thus argued that the Canadian bilingual model may be better served by improved cross-cultural rather than language education.

In addition, the use of corpus linguistics in the form of cross-linguistic corpus-assisted discourse studies is also found to be an indispensable new component in the study of language ideology.

5 Conclusion

In sum, corpus findings pertaining to the case study of language education indicate that French and English serve predominantly different purposes in Canada. While in English-speaking Canada they serve as commodities for social mobility, in French-speaking Canada, the French language tends to be used as an identity marker. The different approaches to languages, and education in these languages, help us to understand why Canada continues to be a country consisting of "two solitudes".

References

Heller, M. (1999). Heated language in a cold climate. In J. Blommaert (Ed.), *Language ideological debates* (pp. 143-170). Berlin: Mouton de Gruyter.

MacLennan, H. (1945). *Two solitudes.* Kingston/Montreal: McGill-Queen's University Press.

Milani, T. M. and Johnson, S. (2008). CDA and language ideology: Towards a reflexive approach to discourse data. In: Ingo H. Warnke and Juergen Spitzmuller (eds), *Methoden der Diskurslinguistik Sprachwissenschaftliche Zugaenge zur transtextuellen Ebene*. Berlin: Mouton de Gruyter, pp. 361-384.

Vessey, R. (forthcoming 2013). Challenges in cross-linguistic corpus-assisted discourse studies. *Corpora, 8* (1).

Woolard, K. A. (1998). Introduction: Language ideology as a field of inquiry. In B. B. Schieffelin, K. A. Woolard and P. V. Kroskrity (Eds), *Language ideologies. Practice and Theory* (pp. 3-50). Oxford: Oxford University Press.

V wh semantic sequences: the communicating function

Benet Vincent
University of Birmingham
bdv700@bham.ac.uk

1 Introduction

Semantic sequences have been proposed as an approach to identifying "regularity in text" in terms of "sequences of meaning elements rather than... formal sequences" (Hunston 2008: 271). Taking such an approach is advocated because "language is full of regularities of meaning that may have little or no surface realisation of the kind that ... a collocational or a lexical bundle analysis might identify" (Groom 2007:50). The power of semantic sequence analysis thus rests in its ability to identify phraseological units which an automatic analysis would miss and to find "what is commonly said" (Hunston 2008) in a particular discourse or situation. The method typically involves concordance analysis focused on closed-class words or grammar patterns (Hunston and Francis 1999) identified as important to the discourse type since they are seen as 'meaning classifiers' (Hunston 2008). Hitherto, semantic sequence analysis has generally been confined to studies of specific discourses (e.g. Gledhill 2000; Groom 2007). However, while evidence presented in Hunston (2008) has suggested that semantic sequence analysis carried out on a general corpus can produce interesting results, this proposal has not been tested on a large scale.

One promising candidate for semantic sequence analysis is the **V wh** pattern, that is verbs which licence subordinate interrogative clauses; the embedded questions thus produced have a range of functions and a wide range of verbs occur in the pattern. Francis et al. (1996) and Huddleston and Pullum (2002) have already shown that **V wh** verbs can be divided into semantically related groups, although these studies do not explain how these groups were arrived at. The proposal is, therefore, that a semantic sequence analysis of this pattern using a general corpus may provide a more motivated phraseological classification of verbs that commonly occur in the pattern and a test of the utility of semantic sequences as applied to general language. The paper reports some of the results of this analysis, restricting itself to one group of sequences with a common function related to communicating.

2 Method

The first stage of this study was to identify those verbs that most commonly occur in the **V wh** pattern in the British National corpus (BNC), accessed using the BNCweb interface (Hoffman and Evert 2008). This involved searching for all forms of candidate verbs which were followed by common *wh*-words and filtering out those which did not pass an arbitrary threshold of 100 instances of the infinitive form of the verb followed by a *wh*-word. Where more than 300 instances of **V wh** were found, a random sample of 300 lines was taken; in other cases, all occurrences were analysed. It was first necessary to separate 'true' instances of interrogative clauses from hits which did not constitute instances of the **V wh** pattern, for example, free relative clauses. Once this was done, a semantic parse was carried out for each verb, which led to the formulation of sequences involving multiple verbs. This semantic parsing takes the perspective that the meaning of the sequence results from a complex interaction between the meanings of the verbs, the subjects of these clauses and the typical meanings expressed by *wh*-clauses. An analysis of example (1), an instance of the 'communicating' sequence with the highest frequency, SOURCE + DESCRIBE + *wh*, can show how this works.

(1) *Wilson-Barnett (1988) indicates how information-giving seems to be gravitating towards 'teaching' individuals how to cope.* [BNC file B14]

In (1), the subject, *Wilson-Barnett*, is parsed as the 'source' of the information presented in the *wh*-clause on the basis that the verb *indicates* here construes *Wilson-Barnett* as an animate agent readily communicating this information to an audience. This analysis is not of course based on one example, but on numerous similar instances across a range of verbs referring to communicative acts (see Results section). The complex interaction between subject, verb and *wh*-clause can be demonstrated by a change of subject type. In example (2), the *work* referred to is not read as *intentionally* communicating the information provided in the *wh*-clause, but is instead an evidence for it, leading to the parse EVIDENCE + SHOW + *wh*, one of the 'proving' sequences (see Results section).

(2) *The work of the Community Education Project indicates how a new pattern of education could emerge in sparsely populated areas.* [BNC file ALE]

3 Results

The methodology described above has yielded five main groups of sequences, divided by general meaning: 'communicating', 'thinking', 'knowing', 'discovering', and 'determining/proving'. This paper focuses on the 'communicating' group, whose sequences can be broadly divided into three main communicative functions:

- information exchange from knowledgeable source to less knowledgeable audience (SOURCE + DESCRIBE + *wh*)
- a request for information from a less knowledgeable to a more knowledgeable person (INQUIRER + ASK + *wh*)
- exchange of views in a group (GROUP + DISCUSS + *wh*)

The first of these sequences, SOURCE + DESCRIBE + *wh* is used to express the willingness, intention, ability, obligation, failure or refusal of a source to reveal or explain the information implied in the *wh*-clause. The main verbs in sequence include DEMONSTRATE, DESCRIBE, DISCUSS, EXPLAIN, INDICATE, REVEAL, SPECIFY and TELL (capitals denote the lemma). With a change in the attitude of the source, there is also a change in the typical meaning encoded in the *wh*-clause. For willingness, for example, we see a predominance of *how* clauses, many of which are arguably not really embedded questions since no information is missing. In contrast, where a source is presented as having failed or refused to reveal information, this information remains unknown, and is hence encoded in a 'true' embedded question, which is much less likely to be introduced by *how*.

The second main function identified for sequences included in the communicating group relates to enquiries or questions and mainly involves the verbs ASK, QUESTION and WONDER. In fact, two functionally distinct sequence types are identified here, the first involving a genuine enquiry, and the second typically expressing criticism or doubt.

Where the inquirer is presented simply as expressing ignorance regarding the answer, the most common verbs are ASK and WONDER. This sequence involves (usually past) reports of questions (e.g. *he asked if she'd had time to think about it*) as well as (usually present) indirect questions and requests (e.g. *I wonder if Ian's heard yet?*).

A second sequence related to questioning, CRITIC + QUESTION + *wh* involves a 'questioner' expressing doubt or criticism about a situation or proposal and typically includes the verb

QUESTION. This sequence has three main characteristics: the questioning is commonly presented as an (inevitable, logical) result; the sequence is also part of a 'Concession–Counter assertion' clause relation, marked by words such as *but*, *although* and *yet*; the *wh*-clause is of the type Huddleston and Pullum (2002) refer to as a 'biased question', that is, one that expects a particular answer.

The third sequence type related to communicating is GROUP + DISCUSS + *wh*. In this sequence, two or more people are presented as discussing an open question, often with a view to coming to a decision about it. In formal discussions, for example conferences or committees, some kind of obligation is likely to be expressed and direction questions are also more likely than in discussions presented as taking place between friends.

4 Discussion

This study shows that a semantic sequence analysis of instances of the **V wh** pattern can reveal a surprising degree of regularity in terms of sequences of meaning elements. The findings can arguably claim to reveal what this pattern is commonly used to say and hence contribute to what Francis (1993) terms a 'grammar of meaning'. Moreover, the phraseological approach taken, which attempts to account for the interaction between characteristics of subject, verb and *wh*-clause type, arguably allows for a more motivated semantic classification of the *wh*-verbs than has been seen up to now.

At the same time, it must be acknowledged that this approach has some drawbacks, perhaps the most important of which is that it is very time-consuming. As Groom (2007) points out, we are a long way from a position in which computers can be programmed to carry out semantic sequence parsing. Another drawback of this approach is the sampling method, which systematically under-represents very frequent verbs such as KNOW. This can be addressed to some extent by using the findings of common patterns to search in corpora for the most common verbs.

References

Francis, G., Hunston, S. and Manning, E. *Grammar Patterns I: Verbs*. London: Collins.

Gledhill, C. 2000. *Collocations in Science Writing.* Tubingen: Narr.

Groom, N. 2007. *Phraseology and epistemology in humanities writing*. Unpublished PhD thesis, University of Birmingham.

Hoffman, S. & Evert, S. (2008) BNCweb (CQP-Edition). Online resource. Available at [http://bncweb.lancs.ac.uk/] (Accessed 5/10/2010)

Huddleston, R. and Pullum, G. 2002. *The Cambridge Grammar of English.* Cambridge: CUP.

Hunston, S. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *IJCL*, 13(3): 271-295

Hunston, S. and Francis, G. 1999. *Pattern Grammar*. Amsterdam/Philadelphia: John Benjamins.

The role of corpus linguistics in social constructionist discourse analysis

Fang Wang
University of Birmingham
w_wangfang@hotmail.com

This paper firstly gives a brief introduction to several approaches to social constructionist discourse analysis, including Ernesto Laclau and Chantal Mouffe's (1990) discourse theory, critical discourse analysis, and discursive psychology. Then, the concept of discourse defined in my recent studies and how corpus linguistics has played an important role in such studies are explained.

Two case studies have been introduced in detail to illustrate my point: in the first study I investigate the 'global warming' discourses in three newspapers, the *Guardian*, the *Washington Post* and the *People's Daily*. A diachronic corpus which includes the articles in which the lexical item *global warming* occurs at least once in these three newspapers during the last 20 years is built and this corpus is further divided into different time periods based on the frequency explosions of the articles concerned. Thus, comparisons of the constructions of the meaning of global warming between three newspapers over time are displayed. It is found that in the *Guardian* discourse, global warming is represented as an accepted fact, while the *Washington Post* discourse remains sceptical. The *People's Daily* also constructs global warming as an irrefutable fact, and it highlights China's contribution to this problem.

In the second study, I examine how the discourse object 'mental depression' has been represented in the British and Chinese national newspapers in the last 25 years, aiming at delivering a contribution to people's understanding of the link between the discourse and the social reality of depression. Two large diachronic corpora have been built to suit this purpose: the English Corpus of Depression (ECD) and the Chinese Corpus of Depression (CCD). Furthermore, to draw a diachronic comparison of the representation of the meaning of depression and 抑郁症 (*yiyuzheng*, 'depression'), these two corpora have been divided into five different time Phase[1] based on obvious frequency explosions of the news articles under investigation. By applying several essential corpus methods such as frequency analysis, collocation analysis, key words analysis and concordance analysis to different Phases of the ECD and the CCD, this study pays special attention to those top frequent words, collocates and key words that enter each Phase and the words that phase out, and never come back. In this way, the features of the development of the meaning of depression or 抑郁症 (*yiyuzheng*, 'depression') can be captured. Lastly, meaning paraphrases of important top frequent words, top collocates and top key words extracted from the ECD and the CCD have been provided. The function of such paraphrases is to give new definitions, to replace what was said before about the meaning of these words, by a new and therefore, we must assume, better way to talk about these concepts. This will help us to complement the previous analysis by spotlighting the instances of meaning negotiations. The main findings of this study can be summarized as: in the beginning Phases of the ECD, depression is, in most cases, constructed as a psychological illness caused by major life events, and the main form of treatment has been represented as psychotherapy. While in the following Phases, depression is more constructed as a biological disease and thus can be treated by antidepressants. The final Phase of the ECD constructs depression as a rather complex problem that needs more scientific research, and more integrated forms of treatment are recommended. By contrast, in the CCD, it is found that depression has been always considered as being caused by external factors, such as problems of human relationships and so on. Therefore, to cure depression, both psychological treatment and the repair of human relations or other external problems have been constructed as crucial in Chinese context. Medication, on the other hand, has been marginalized and represented as a last choice, though in the last Phase of the CCD, the role of medication started to be addressed.

The findings revealed by these two studies show that the traditional synchronic perspective in corpus linguistics needs to be complemented by a diachronic dimension. It is also explained that the corpus findings generated by scientific research methods can only become valid when complemented by interpretation. The combination of a traditional corpus research and innovative paraphrase analyses will give us a clearer picture of how a certain discourse object was understood at a given time, which is an extension of corpus linguistics through the inclusion of the diachronic dimension of discourse.

A new concept of discourse is proposed:

[1] The capitalized word Phase is specially used to refer to the different Phases of a corpus.

discourse is all that has been said about one discourse object (Teubert, 2010). A new way of social constructionist discourse analysis by applying the methodology of corpus linguistics is introduced. It is argued that the task of corpus linguists is not only to look at scientific evidence generated by corpus research tools, but to interpret the findings in the light of social theories such as social constructionism and social epistemology. With this in mind we might envisage a new focus for corpus linguistics in the future, where a new research methodology of the humanities and social sciences is formed, incorporating both scientific and interpretative methods, in order to get a fuller understanding of language.

References

Laclau, E. and Mouffe, C. 1990. 'Post-Marxism without apologies', in E. Laclau, *New Reflections on the Revolution of Our Time*. London: Verso.

Teubert, W. 2010. *Meaning, discourse and society*. Cambridge: Cambridge University Press.

Using life-logging to re-imagine representativeness in corpus design

Stephen Wattam
Lancaster University
s.wattam@lancs.ac.uk

Paul Rayson
Lancaster University
p.rayson@lancs.ac.uk

Damon Berridge
Lancaster University
d.berridge@lancs.ac.uk

1 Introduction

Conventional corpus building efforts have focused primarily on sampling language as a persistent entity, distinct from its origins as a social event. Though the sampling methods first used in the Brown/LOB corpus family have yielded many valuable results – and will doubtless continue to do so – this strategy has led to a number of pragmatic and scientific issues that continue to limit corpus linguistics.

Many of these issues have been framed in terms of improved description of language variation, most notably by Biber (1993) who assesses the variability within various language categories (e.g. preposition and relative clause usage) in order to determine the extent to which we may, probabilistically speaking, extract further useful meaning by sampling further. He goes on to conclude that we cannot know, a priori, the relative proportions of register usage within a language, and that we must go about sampling in an iterative manner, learning from our findings in order to improve further efforts.

Biber's call for multi-phase sampling stems from a number of unknowns regarding language: we simply do not know what language is used on a daily basis, by whom, or for what purpose, with sufficient resolution to construct a demographically- and linguistically-balanced sample of language use with any certainty. Some of this uncertainty is, in part, down to the pragmatic decision to sample primarily in terms of language as a separate entity, making attempts to balance demographics by using proxy variables such as relative popularity of a text, formality and other socio-cultural external descriptors. This means we are sampling yet one step further from Evert's (2006) definition of the ideal, intensional, properties we wish to measure.

As discussed in Leech (2006), a prime example of this obfuscation is selection of language according to production or consumption, something only vaguely addressed by many

corpora yet particularly important to the findings of many studies. Leech calls for a re-inspection of proportionality, saying "the representation of texts should be proportional not only to their initiators, but also to their receivers". Hoey (2005) also laments this effect, stating "A corpus… represents no-one's experience of the language."

Since corpus studies are necessarily concerned with the social demographics underlying language (either directly or by assumption), there is a need to improve the quality of social and demographic information in corpora, and its relationship to real population distributions.

One way to examine these contextual and demographic markers is to construct a personal corpus – a record of all language input and output by a single person. This sample would represent a trade-off: accurately sampling text type proportions whilst sacrificing power in describing a wider population.

Sampling in this manner makes explicit many of the assumptions made during conventional corpus sampling, such as those involving variables with difficult-to-enumerate populations:

- Temporal information, such as the age of a document when used;
- The relative properties of language produced and consumed, and frequency of re-read material;
- Detailed demographic information surrounding those using the language;
- Proportionality of language categories used;
- Representative sampling of ephemeral and short texts (e.g. greetings, flyers, advertisements).

Such a sample may be seen as a first effort towards a model of language proportionality that is not based upon proxy variables such as setting, formality or plurality.

In this study, we describe the process of designing and building a personal corpus using techniques derived from life-logging. Ultimately, we aim to:

- Develop a process of recording, digitising and coding multi-modal linguistic data in-the-field;
- Identify proportions of language used for a single demographic, and compare these to existing corpora for an initial indication of possible improvements;
- Develop methods for resampling corpora to align them in terms of a personal corpus, including relation of existing corpus genres to those commonly seen;
- Develop a methodology that may be used to reweight corpora with less intrusion upon a given subject, through digital methods (such as logging) and conventional methods borrowed from the social sciences.

It is hoped we will be able to work towards an answer to some of the more persistent questions of corpus design: sample size, variability in real-world contexts, and stratum proportions, and, in so doing, assist those attempting to infer information from existing corpus resources by augmenting their meta-data.

2 Literature

One of the major challenges in capturing language use comprehensively is sampling small texts (such as greetings or brochures). These features are only sufficiently represented by verbatim recording methods, such as those used in the life-logging community. Though life-logging is commonly accomplished with video recording systems, making the data hard to summarise and digest without manual review, there have been some efforts to use textual and multi-modal input.

Many of these systems have been designed to augment human memories: Microsoft's MyLifeBits project (Gemmell et al. 2002) uses, amongst other things, their SenseCam device (Hodges, S. et al. 2006) to incorporate documents and other ephemera, using their metadata to inform memories and records and integrating this with calendar software.

Combining the life-logging concept with that of learner corpora, Roy's work on the 'Human Speechome' project (2006) follows the linguistic development of his own son over a three year period. His methods centre around continuous video recording and automated processing.

Efforts in low-impact personal archives have produced methods for recording 'high-level' details about spoken interaction (Lee and Ellis, 2006): the number of speakers, duration, length of interaction, etc. These methods, combined with those designed for other 'real world' tasks such as automotive voice recognition and mobile phone noise cancellation, may be used to reduce the load of processing long recordings.

3 Capture

One other primary consideration when sampling

language is how to retain comparability with other corpora. This is especially challenging in the case where data sampled may fall beyond the bounds of existing taxonomies. Since the utility of a personal corpus is damaged by any disruption caused by its sampling methods, any metadata stored on texts must therefore be succinct, easily related to other corpora, and 'evocative' (such that a researcher may interpolate their data at the end of a day).

With this in mind, we have decided to sample genres, contexts and purposes using semi-free coding – the researcher will be aware of existing genre distinctions, but will insert others where they most accurately describe the use of the text.

Sampling methods have been selected to cover most common forms of interaction with language, with a focus on capturing those that may have 'slipped through the cracks' of other corpora. These include:

- Manual annotation, to record long or inconvenient texts (such as books) and metadata on the following;
- Continuous audio recording, to capture speech, radio and short interactions;
- Photography, for billboards and posters;
- Digital logging of web use, keyboard input, and online chat;
- Web storage for digital documents encountered elsewhere;
- Physical record-keeping.

Each of these presents a number of ethical and practical issues, which require resolution before such methods are practical for general use.

4 Output and utility

Though the statistical ideal would be to have many personal corpora, and combine them into a linguistically-balanced single sample, a single sample may yield useful information when combined with existing corpora. This extreme form of personal corpus may be used to investigate complex theories of language acquisition and use, such as those covered in Hoey's work on lexical priming, in a rigorously generalisable manner.

Even without 'full coverage', we may use demographic information to locate the subject (or subjects) within the distribution of language users covered by a conventional corpus, allowing us to relate the language proportions observed to a wider context with a high degree of accuracy. This is especially useful for establishing what text types are missing altogether.

Reversing this process allows us to re-sample existing corpora using the identified personal proportions of text types, in order to produce an augmented corpus that is balanced for a given demographic, yet large enough to train existing NLP systems. This process is also the key to easing the ethical concerns surrounding such invasive recording, as it makes possible the elicitation of language proportions through the use of questionnaires or irreversibly-hashed recording techniques, as proposed in (Ellis and Lee, 2004). This, in turn, would allow for the application of current general-purpose corpus data to targeted research questions.

5 Conclusion

We believe that the concept of a personal corpus forms a complementary sampling strategy to that of conventional corpora. The rise in digital language consumption, along with recent advances in speech recognition and portable device power, allows us to start the process of re-examining and improving our language resources with more targeted sampling strategies. This will ultimately allow us to better align corpora both to the ground truth and to the aims of linguistic studies, making them both easier to use and more fruitful.

References

Biber, D 1993 "Representativeness in corpus design" *Literary and linguistic computing* (8), p243-257

Ellis, D.P.W. And Lee, K. 2004 "Minimal-impact audio-based personal archives" *Proceedings of the 1st ACM workshop on Continuous archival and retrieval of personal experiences* p39-47

Evert, S. 2006 "How random is a corpus? The library metaphor" *Zeitschrift für Anglistik und Amerikanistik* (54) p177-190

Gemmell, J. and Bell, G. and Lueder, R. and Drucker, S. and Wong, C. 2002 "MyLifeBits: fulfilling the Memex vision" *Proceedings of the tenth ACM international conference on Multimedia* p235- 238

Hodges, S. and Williams, L. and Berry, E. and Izadi, S. and Srinivasan, J. and Butler, A. and Smyth, G. and Kapur, N. and Wood, K. 2006 "SenseCam: A retrospective memory aid" *UbiComp 2006: Ubiquitous Computing* p177-193

Hoey, M. 2005 *Lexical priming: A new theory of words and language* p14 Routledge

Lee, K. and Ellis, D.P.W. 2006 "Voice activity detection in personal audio recordings using autocorrelogram compensation" *INTERSPEECH 2006: ICSLP: Proceedings of the Ninth International Conference on Spoken Language*

Processing: September 17-21, 2006, Pittsburgh, Pennsylvania, USA p1970-1973

Leech, G. 2006 "New resources, or just better old ones? The Holy Grail of representativeness" *Language and Computers* (59) p133-149

Roy, D. and Patel, R. and DeCamp, P and Kubat, R. Fleischman, M and Roy, B. and Mavridis, N. and Tellex, S. and Salata, A. and Guinness, J. and Levit, M and Gorniak, P. 2006 "The Human Speechome Project." *Proceedings of the 28th Annual Cognitive Science Conference.*

Code-mixing: exploring indigenous words in ICE-HK

May L-Y Wong
University of Hong Kong
wlymay@gmail.com

1 Introduction

This paper attempts to demonstrate the usefulness of a corpus to study the phenomenon of code-mixing on the basis of language data gathered from the Hong Kong component of the International Corpus of English (ICE-HK), which was made publicly available in March 2006 (Nelson 2006) as opposed to introspective examples as used in previous literature.

2 Sociolinguistic background and code-mixing

Hong Kong is basically a monoethnic society with over 95 percent of its total population being Chinese. In this regard, Chinese is considered in this paper as the dominant language and English the non-dominant language. However, English still retains a very strong influence in the territory. A detailed account of language situation in Hong Kong has been given in Bolton (2000; 2003: 93-99) and Wong (2012). Given the available linguistic resources and emerging community norms towards different functions of the English and Chinese language in Hong Kong, an interesting question is what governs the choice of language in interactions, and to what extent previous models of code choice are applicable.

This paper focusses on the Hong Kong community in order to provide a case study of code-mixing in the Cantonese world. There are many examples of English elements being mixed into Cantonese and this has been reported extensively in the literature (e.g. Gibbons 1983, 1987; Li 1996; Chan 1998, 2003, 2007; Bauer 2006; Wong et al. 2007). However, there is much less available information on code-mixing of indigenous Cantonese words into English. The use of a single indigenous expression (e.g. *chah chaan teng* 'a type of Hong Kong-style restaurant serving a mixture of Chinese and western food' as in example (1)) in largely English-language discourse could provide some hints about the motivation of code-mixing in relation to ethnic (i.e. Hong Kong Chinese) identity (cf. Martin 2005: 120). This paper, then, seeks to offer a small contribution in this context.

(1) B: It's just like a normal fast food not fast food uhm <indig> *chah chaan teng* </indig> <&> a Cantonese bistro serving Chinese and Chinese-Western food </&> kind of restaurant. (S1A-100#86)

3 ICE-HK and indigenous words

The ICE-HK project was initiated in the early 1990s (Bolt and Bolton 1996). The ICE-HK corpus follows the common design of other ICE corpora worldwide, containing approximately one million words and including both spoken and written data. Cantonese indigenous words occur 5.54 times per 10,000 tokens in the spoken section of the ICE-HK corpus and 3.70 times per 10,000 tokens in the written section. Roughly speaking, the ratio of indigenous expressions in speech to writing is 1.5:1. The written samples in ICE-HK seem to be akin to the spoken texts in the use of local terms. This finding lends some support to what Yau (1993) has found with some locally printed publications such as newspapers and popular magazines that written material tends to be more conservative and less prone to code-mixing. However it should be borne in mind that her written samples might be too formal to allow for any mixed code, which usually carries the stigma of a decline in language standards at the time of her writing the paper (Joseph 1996). In contrast, the ICE-HK corpus contains both formal and informal genres and thus serves as a much better basis for a comparison between speech and writing with respect to the occurrence of indigenous words. As expected, private dialogues such as face-to-face conversation (S1A) as well as unprepared speeches (S2A) have shown a high incidence of indigenous expressions, which make up 17.6% and 14.7% respectively of the indigenous words in the corpus, as do student writing (W1A) and creative writing (W2F) which together account for nearly half of the local terms.

Type of Indigenous Words	Freq.	%
A: colloquial formulaic sequences	226	31.9%
B: Chinese / Hong Kong customs	218	30.7%
C: local food and cooking	32	4.5%
D: kinship terms	13	1.8%
E: proper nouns (person)	43	6.1%
F: proper nouns (place)	35	4.9%
G: proper nouns (organisation)	22	3.1%
H: miscellaneous English vocabulary	120	16.9%
Total:	709	100.0%

Table 1: classification of indigenous words in ICE-HK

Table 1 illustrates the wide variety of Cantonese indigenous words used in Hong Kong English.

4 Major corpus findings

About a third (226 out of 709) of the total amount of Cantonese code mixed in English discourse is comprised of formulaic sequences. The Cantonese formulaic sequences are often used when people feel the need to express themselves explicitly, as shown in examples (2) and (3). Besides conveying linguistic meaning, languages also carry with them extralinguistic meaning. In the context of Hong Kong, English carries a status of power, education and wealth whereas the use of Chinese is a symbol of ethnic solidarity. Extensive research on the mixing of English vocabulary into Cantonese has been conducted with anecdotal evidence, speech samples, newspapers and magazines. Less common, however, has been the use of code-mixing the other way round – that is, Hong Kong people regularly use indigenous Cantonese words while speaking in English. This study has shown that most of the factors and motivations of code-mixing of this kind can be situated in existing frameworks. In particular, solidarity has been proved to be a key factor in determining the occurrence of colloquial formulaic sequences and Cantonese kinship terms in English spoken and written texts. Negotiating a local identity through indigenous vocabulary is another motivation in code-mixing. This is illustrated in local food items, some of which (e.g. *dim sum* and *chau mihn*) have attained currency in world English. Ethnic identity has also made the mixing of Cantonese terminology into English discourse a popular choice when talking about Chinese or Hong Kong customs and culture e.g. *feng shui*, *lai see*, *Cantopop* and *gwai lo*. In the case of proper nouns, mixed code performs a referential and expressive function for Hong Kong people, when they lack knowledge of the English equivalents of some Chinese words. Another reason for using mixed code is the spontaneous nature of conversation, making it impossible for a person to instantly think of the English terms which are too technical or too uncommon in everyday use. There are several cases of occasional code switching in full Cantonese utterances reported in this study. These rare code switching cases as well as Cantonese substitutions for basic-level English vocabulary are simply an act of accommodation, associating speakers with other in-group members and enhancing social solidarity.

(2) Actually I am a member of lacrosse team also but I

never go to practice because I I guess is <indig> *aai* </indig> <&> an interjection of regret </&> you know I'm lazy to uh take so long for travelling and (S1A-005#231)

(3) A: I think especially in Hong Kong it's a very busy city
A: And and the and the language is changing so fast
A: Even you know uh if a person who live in Hong Kong you live for one or two years and when you come back maybe you cannot catch up with the language
A: Because there are always some kind of language, newly invented, changed
Z: Yeah
Z: Changed
C: <indig> *Hou gik a* </indig> <&> Very extreme/awesome </&>
A: Yeah (S1A-009#196 – 203)

References

Bauer, R. 2006. The stratification of English loanwords in Cantonese. *Journal of Chinese Linguistics* 34 (2): 172-191.

Bolton, K. 2000. Researching Hong Kong English: bibliographical resources. *World Englishes* 19 (3): 445-452.

Bolton, K. 2003. *Chinese Englishes: A sociolinguistic history*. Cambridge: Cambridge University Press.

Chan, B. H.-S. 1998. How does Cantonese-English code-mixing work? In M. Pennington (ed.) *Language in Hong Kong at century's end* 191-216. Hong Kong: Hong Kong University Press.

Chan, B. H.-S. 2003. *Aspects of the syntax, the pragmatics and the production of code-switching: Cantonese and English*. New York: Peter Lang.

Chan, B. H.-S. 2007. Hybrid language and hybrid identity: the case of Cantonese-English code-switching in Hong Kong. In Chan K.-B., J. Walls and D. Hayward *East-west identities: globalisation, localisation and hybridization* 189-202. Leiden and Boston: Brill Academic Press.

Gibbons, J. 1983. Attitudes towards languages and code-mixing in Hong Kong. *Journal of Multilingual and Multicultural Development* 4 (2-3): 129-147.

Gibbons, J. 1987. *Code-mixing and code choice: a Hong Kong case study*. Clevedon and Philadelphia: Multilingual Matters.

Joseph, J. 1996. English in Hong Kong: emergence and decline. *Current Issues in Language and Society* 3 (2): 166-179.

Li, D. C. S. 1996. *Issues in bilingualism and biculturalism: a Hong Kong case study*. New York: Peter Lang.

Nelson, G. 2006. *The ICE Hong Kong Corpus: user manual*. London: University College London.

Wong, C., Bauer, R. and Lam, Z. 2007. The integration of English loanwords in Hong Kong Cantonese. Paper presented at the 17th annual meeting of the Southeast Asian Linguistics Society (SEALSXVII), 31 August – 2 September 2007. University of Maryland, USA.

Wong, M. L-Y. 2012. Hong Kong English. *The Mouton World Atlas of Variation in English*, ed. Bernd Kortmann and Kerstin Lunkenheimer. Berlin and New York: Mouton de Gruyter.

Yau, M-S. 1993. Functions of two codes in Hong Kong Chinese. *World Englishes* 12 (1): 25-33.

Using corpora in forensic authorship analysis: Investigating idiolect in Enron emails

David Wright
University of Leeds
en07dw@leeds.ac.uk

1 Forensic linguistics, idiolect and corpora

The forensic linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of their language, their own *idiolect* (Coulthard 2004:431). However, Kredens (2002:405) highlights that 'this claim has not thus far been supported by empirical research', and Barlow (2010:1) notes the remarkable 'gap between the familiarity of the concept and lack of empirical data on the phenomenon'. As a result of this abstract nature of idiolect, over recent years there has been discussion into how the theory can be best conceptualised for use in forensic authorship analysis (Grant 2010; Turell 2010). Thus, although forensic linguistics is the field which arguably relies most heavily on the concept of idiolect, the discussion remains largely a theoretical one. Meanwhile, empirical research into the existence and accessibility of idiolect is emerging from corpus linguistics (e.g. Mollin 2009; Barlow 2010; Mollet et al. 2010), and has shown that it is possible to use corpora effectively in making observations regarding linguistic individuality and uniqueness. In turn, this paper draws on such corpus linguistic methods in investigating idiolect, and thus responds to calls for the use of corpora in forensic linguistic contexts (Cotterill 2010; Kredens and Coulthard 2012).

Using email data from the former American energy company Enron, this paper empirically investigates the ways in which corpora can be applied in identifying author-distinctive linguistic variation. The approach here draws on the established relationship between genre and language use (Hymes 1974), as forensic linguists, in both casework and research, have productively made use of the ways in which individuals' linguistic choices vary within text-type conventions (Turell 2010; MacLeod and Grant 2012). Therefore, the main research aim of this paper is to identify how distinctive an author's linguistic choices within two structural conventions of the email genre – greetings and farewells – can be when tested against a relevant population of writers.

2 The Enron email corpora

In the early 2000s, as part of the legal investigation into Enron's illegal accounting practices, the email data of 150 Enron employees, containing approximately half a million emails, was made publically available online. The source of the data for this study is that provided by Carnegie Mellon University (CMU) (Cohen 2009). The data was extracted, prepared and designed for the purposes of this research by Woolls (2012) using specialised data extraction software. The extraction process created two specific sub-corpora from the CMU dataset: the 'Trader Sent Corpus' (TSC) comprising 2,622 emails and 86,902 words extracted from the 'sent' folders of four Enron traders, and the much larger 'Enron Sent Email Author Reference Corpus' (ESEARC) containing a further 40,236 emails and almost 2 million words sent by an additional 126 Enron employees.

3 Methodology: Comparing corpora and calculating likelihood ratios

The emails for each of the four authors in the TSC were coded, using both qualitative and computational methods, for their use of greeting and farewell variants. For greetings, a specially-designed computer program (Woolls 2012) was used to extract all of the first words of all of the emails in TSC. The resulting list of first words for each of the four authors was sorted computationally and coded manually. Ultimately, all of the greeting forms which occurred more than once could be categorised as belonging to one of three main types: *naming greetings* (i.e. those which include the recipient's name), *hi/hey/hello greetings*, and *no greeting*. In turn, email farewells were coded entirely manually. All of the farewells which occurred more than once in the corpus belonged to one of four main categories: *naming farewells* (i.e. those which include the sender's name), *thanks/thank you farewells* (i.e. any form which included *thanks* or *thank you*), *regards/love/later farewells* and *no farewell*. In the coding of both greetings and farewells, differences in punctuation, line spacing and capitalisation were taken into account. That is, if two greeting forms, for example, differ only in the capitalisation of their first letter, then they are considered different variants. The motivation for this is that it is often such linguistic choices that are least likely to be consciously controlled

by the author (Johnson 2012).

The comparison of corpora in identifying distinctive and significant lexico-grammatical patterns of texts, genres and language varieties is a commonly employed corpus linguistic method (e.g. Biber 1988; Baker 2010; Conway 2010; Mollet et al. 2010; Kwary 2011). Drawing on this approach, with a focus on the linguistic individual, those greeting and farewell forms which were identified as being individuating or distinctive of one of the four authors in the TSC were then tested against the ESEARC to measure their relative frequency, rarity and distinctiveness against a reference set of Enron writers. Using a specialised reference corpus in this way allows for the measurement of the 'population-level distinctiveness' (Grant 2010:515) of these greeting and farewell forms, i.e. how distinctive, unusual or unique a writer's linguistic choice is within a relevant population of writers. In this study, this involved using simple likelihood ratios (McGee 2002), the calculation of which hinges on comparing the number of times a trader in question in the TSC uses a particular variant as a proportion of all of their emails, with the number of times that this variant was found in the 40,236 emails of the ESEARC. The result produced represents the likelihood that an email would include a particular variant if the trader in question had or had not written it. In turn, such observations can offer evidence of author-distinctive idiolectal linguistic choices.

4 Results: identifying author-distinctive patterns

The main categories of greetings (no greeting, naming greetings and hi/hey/hello greetings) and farewells (no farewell, naming farewells, thanks farewells, and regards/love/later farewells) identified in TSC all align with much of the previous research on email, which has reported these types of greetings and farewells as recurrent in a range of email corpora (e.g. Waldvogel 2007; Biber and Conrad 2009). More specifically, within these categories, the analysis of the TSC identified 19 greeting and 12 farewell variants which can be used to distinguish between the emails of the four traders, with each trader having a range of forms distinctive of their email style. Interestingly, many of these variants which were distinctive of particular traders were low in frequency, sometimes being used in less than 1% of their emails.

The population-level distinctiveness of these 31 variants was then tested by identifying their frequency in the ESEARC. The likelihood ratio results showed that four of the variants lost their distinctiveness altogether; that is, they were as likely, or more likely, to occur in an email written by a member of a relevant population in ESEARC than to occur in an email written by the trader of which they were initially distinctive. Overall, however, especially for greetings, the results were encouraging, with some variants being 60, 230 and 500 times more likely to appear in an email written by the trader in question than another writer in the relevant population. In addition, likelihood ratios for greetings and farewells were multiplied to measure the author-distinctiveness of the co-selection of forms, showing that, although a greeting form and a farewell form may independently be rather common, their co-occurrence in an email can be much rarer. Further, evidence suggests that greeting and farewell variants are pre-determined and pre-conditioned given the recipient and function of the email (request, order, etc), and the ways in which these linguistic and situational features relate and interact can reveal author-distinctive patterns. The results of these analyses highlight the importance of low-frequency features, as it was often the less frequently occurring variants which were the most distinctive at population level. This has important implications for forensic linguistics, particularly at a time in which the increased attention paid to ethics and standards of best practice in forensic linguistics is raising questions regarding the significance and validity of low-frequency features in linguistic evidence (Butters 2012).

5 Conclusion: corpus linguistics advancing forensic authorship analysis (and vice-versa)

This research offers methodological and theoretical contributions to both corpus linguistics and forensic authorship analysis. Methodologically, the approach used here demonstrates the invaluable ways in which carefully designed corpora can be utilised in the identification of author-distinctive patterns of use in email. In turn, the results produced highlight just how distinctive particular linguistic choices can be of an author when tested against relevant population data, even within as few as two structural conventions of the email genre. It may be that for forensic linguists, the best way to conceptualise idiolect, and the most accessible way of analysing it, is by identifying an individual's distinctive linguistic behaviour within the conventions of specific genres, or their individuating *genre-lects*. Overall, this paper emphasises the central role corpora must hold in the empirical investigation of idiolect, with

significant potential for application in forensic linguistic research and casework.

References

Baker, P. 2010. Representations of Islam in British broadsheet and tabloid newspapers 1999–2005. *Journal of Language and Politics* 9 (2): 310-338

Barlow, M. 2010. Individual Usage: A corpus-based study of idiolects. *34th International LAUD Symposium*, Landau, Germany.

Biber, D. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press

Biber, D. and Conrad, S. 2009. *Register, Genre and Style.* Cambridge: Cambridge University Press.

Butters, R. 2012. Retiring President's closing address: ethics, best practices, and standards. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*, Aston University, Birmingham, 351-361. From: www.forensiclinguistics.net

Cohen, W.W. 2009. *Enron Email Dataset*. Retrieved October 2010 from: http://www.cs.cmu.edu/~enron/.

Conway, M. 2010. Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing* 25 (1): 23–35.

Cotterill, J. 2010. How to use corpus linguistics in forensic linguistics. In A. O'Keefe and M.McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 578-90.

Coulthard, M. 2004. Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics* 24 (4): 431-447.

Grant, T. 2010. Txt 4n6: Idiolect free authorship analysis?. In M. Coulthard and A. Johnson (eds.) *The Routledge Handbook of Forensic Linguistics.* London: Routledge, 508-522.

Hymes, D. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. London: Tavistock.

Johnson, A. 2012. Applying forensic linguistics in professional settings: Implications for research. Paper presented at the 1st Inter-university PhD Seminar on Forensic Linguistics (University of Leeds & IULA/Universitat Pompeu Fabra), Universitat Pompeu Fabra, Barcelona. 30 March 2012.

Kredens, K. 2002. Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In B. Lewandowska-Tomaszczyk (ed.) *PALC'01: Practical Applications in Language Corpora.* Peter Lang: Frankfurt am Mein, 405-437.

Kredens, K. and Coulthard, M. 2012. Corpus Linguistics in authorship identification. In P. Tiersma and L. Solan (eds.). *The Oxford Handbook of Language and Law.* Oxford: Oxford University Press, 504-516.

Kwary, D.A. 2011. A hybrid method for determining technical vocabulary. *System* 39 (2): 175-185.

MacLeod, N. and Grant, T. 2012. Whose Tweet? Authorship analysis of micro-blogs and other short form messages. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*, Aston University, Birmingham, 210-224. From: www.forensiclinguistics.net

McGee, S. 2002. Simplifying likelihood ratios. *Journal of General Internal Medicine* 17 (8): 647-650.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N.R. and Wright, M.J. 2010. Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics* 15 (4): 429-473.

Mollin, S. 2009. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14 (3): 367-392.

Turell, M. T. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law* 17 (2): 211-250.

Waldvogel, J. 2007. Greetings and closings in workplace email. *Journal of Computer-Mediated Communication* 12 (2): 456–477.

Woolls, D. 2012. *Description of CFL extraction routines for CMU Enron Sent email database.* Retrieved March 2012 from: http://www.cflsoftware.com/CFL_CMU_Enron_Sent_email_Extraction.mht

A multidimensional contrastive move analysis of native and nonnative English abstracts

Richard Xiao
Lancaster University
z.xiao@lancaster.ac.uk

Yan Cao
Ludong University
yan_cao@163.com

1 Introduction

This article aims to explore, from a contrastive perspective, textual variations in discourse moves in English abstracts by native and nonnative writers, adopting the multidimensional analysis (MDA) approach in conjunction with the more traditional move analysis in EAP research, on the basis of a sizeable corpus composed of two balanced matching components representing respectively English abstracts written by native English and Chinese writers from twelve academic disciplines.

2 Review of approaches to RA abstracts

As a critical part of the research article (RA) and a specific genre in its own right (Hyland 2004), abstracts have become a focus of research that has attracted growing attention in recent years. A number of studies have focused on the general organisation of RA abstracts across different disciplines (e.g. Samraj 2005; Kanoksilapatham 2007). More specifically, rhetorical functions have been investigated along with linguistic features, revealing a close relationship between the rhetorical functions of abstracts and the linguistic features they use. Cross-linguistic variations have also been an important concern for researchers (e.g. Van Bonn and Swales 2007), with the associated studies either seeking to uncover the underlying causes of cross-linguistic variations or identifying writing problems and difficulties for writers of particular mother tongues.

In spite of the range of topics addressed in previous studies of abstracts, they have focused primarily on the macro structure, limited in both subject areas covered and the number of linguistic features investigated. To our knowledge, Kanoksilapatham (2007) is perhaps the only published study which combines the MDA approach and move analysis. In this study, a corpus of 60 articles sampled from five journals of biochemistry, all published in 2000, is annotated with fifteen move types in four different sections (Introduction, Methods, Results, Discussion) as well as linguistic features used in the factor analysis that establishes a 7-dimension MDA model of RAs. The combined approach demonstrates its strength in characterizing the textual variation at both macro and micro levels. The present research will shift the focus from RAs to abstracts instead, complementing previous research of abstracts by adopting the MDA approach in a contrastive analysis of English abstracts written by native and nonnative English writers.

The great number of abstracts sampled, the wide range of disciplines covered, the large number of linguistic features investigated, the rigorous statistic measures taken, and the depth of analysis made in the present study have enabled us to provide a more realistic and accurate account of Chinese writers' English RA abstracts in relation to native English norms. Methodologically, this study has innovatively expanded Biber (1988) and Xiao's (2009) MDA models by integrating colligation patterns into the multidimensional analytical framework, which in combination with grammatical and semantic features, have helped the MDA approach to offer even stronger interpretations of the discourse functions of dimensions based on closer integration of form and meaning in language variation research. In addition, the research presented in this article has also helped to extend the MDA approach from vastly different genres to those of similar nature.

3 Methodology

We designed and developed two matching balanced corpora that are composed respectively of English abstracts written by native English (NS) and native Chinese (NNS) writers from twelve academic disciplines, with the NS and NNS subcorpora amounting to approximately 890,000 and 1,050,000 words respectively. Each abstract in the corpus was properly marked up and annotated with word class and semantic information using Wmatrix (Rayson 2008). Since the manual annotation of discourse moves was necessarily time and labour consuming and could only be applied on a small scale, move analysis was only undertaken on the Biology subcorpus, which comprises 600 NS abstracts and 600 NNS abstracts, amounting respectively to 129,340 and 133,355 words. Santos' (1996) five-move model was used as a framework for analysing and annotating the structure of abstracts: background (B), introduction (I), methodology (M), result (R), and conclusion-discussion (C-D).

The present study takes Biber's (1988) MDA approach and follows the steps outlined in Biber

(1988) and Xiao (2009), but it further integrates into the new MDA model a selection of colligation patterns, which are the most frequently used 3-grams chosen from each group on the basis of the word classes of the colligative cores, namely nouns, verbs, adjectives, adverbs and prepositions. Our feature list includes a total of 163 linguistic features, which can be divided into three categories, i.e. grammatical features, semantic features, and colligation patterns. The factor analysis based on the normalised frequencies (per 100 words) of these linguistic features reduced our initial selection of 163 linguistic features to 47 with significant loadings on seven underlying dimensions. The new MDA model has provided a theoretical framework for our further move analysis of NS and NNS abstracts in the Biology discipline.

4 Results and discussion

The contrastive move analysis of the Biology subcorpus reveals that both NS and NNS abstracts have a high percentage of moves I (introduction), M (method) and R (result), suggesting that both NS and NNS writers consider these moves as obligatory parts of RA abstracts. However, NS writers tend to use move B (background) more frequently than NNS writers. While NNS writers tend to use a separate move M, NS writers are more likely to combine move M with move I (introduction) or move R (result), which is probably influenced by economy of style of English abstract writing.

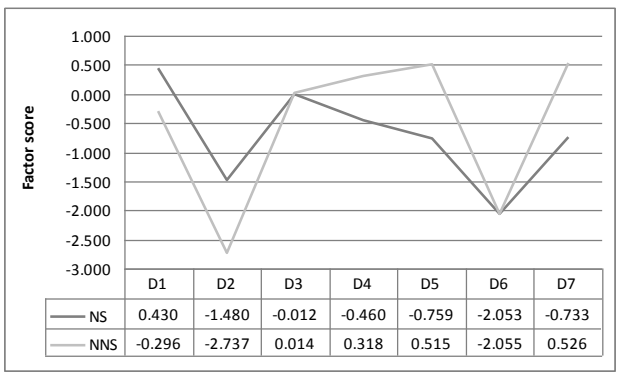
Our new MDA model of RA abstracts comprises seven dimensions as indicated in Table 1, which also gives, in the order of their weights, significant linguistic features loaded on each dimension. A more detailed discussion of the dimensions established in the MDA model, together with their associated linguistic features, can be found in Cao and Xiao (2013).

Following the three formulae for computing factor scores developed in McEnery et al. (2006), we obtained the factor score of each individual abstract and then of each discipline and move, in both NS and NNS abstracts, and finally the dimension scores of NS and NNS abstracts. Figure 1 compares the overall dimensional variations in NS versus NNS abstracts.

A more fine-grained analysis of discourse moves in Biology abstracts along the seven dimensions shows some interesting differences and similarities between NS and NNS abstracts in terms of the communicative functions of specific discourse moves. NS abstracts display a significantly higher score in Dimension 1 in all move types where there is a significant difference (I, M, R), suggesting that in relation to native English writers, Chinese writers are less focused and less confident in using intensifying devices to emphasise the points they make, using a less confident writing style.

Dimension	Linguistic features
D1. Focusing and intensification	other adverbs; adv: degree; adv: measurement; adv+V; adv: boosters; V+adv; general emphatics; split auxiliaries
D2. Active involvement and interaction	pronouns; zero relative; 1st person pronouns; *we*+V+N/adj/adv/Art; *that*-clauses as verb complements; *we*+V+*that*; private verbs; demonstratives; V+N/adj/adv/Art+N; pronoun *it*
D3. Explicit conceptualisa-tion of methodology	nouns: mental object; verbs: using; verbs: getting, giving, possession; science and technology in general; nouns: evaluation; verbs: speech acts; prep, *be* as main verb
D4. Conceptual elaboration	attributive adjectives; adj: measurement; adj+adj+N; adj: physical attributes; adj+N+N; prep+art+adj+N; adj: importance; adj: evaluation
D5. Formal, abstract and impersonal style	agentless passives; verbs: being; *be*+past participle+prep; *by*-passives; V+prep
D6. Textual cohesion	prep; grammatical bin; *the*+N+*of*; prep+art+N+*of*; V+prep; (prep+art+adj+N)
D7. Informational density	nominalisation; mean word length; nouns: affect; N+coordinating conj+N; phrasal coordination

Table 1. MDA model of RA abstracts



	D1	D2	D3	D4	D5	D6	D7
NS	0.430	-1.480	-0.012	-0.460	-0.759	-2.053	-0.733
NNS	-0.296	-2.737	0.014	0.318	0.515	-2.055	0.526

Figure 1. NS and NNS along seven dimensions

Similarly, along Dimension 2, NNS writers also demonstrate a less active involvement and interaction in all move types.

In addition NS abstracts are more likely to express the means and method explicitly in moves I, R and CD, demonstrating a higher propensity for explicit conceptualisation of methodology than NNS abstracts in Dimension 3.

In contrast, Chinese writers are found, in Dimensions 4 and 5, to show a stronger preference for conceptual elaboration and for a formal, abstract and impersonal style, for example, through a more frequent use of pre-modification of nouns and passives in most discourse moves.

On the other hand, NS and NNS abstracts demonstrate very little significant difference along Dimension 6, which concerns the authors' ability to achieve textual cohesion. This is true in all move types other than R (result), where results are presented more coherently in NS abstracts.

Finally, NS and NNS differ in terms of informational density (Dimension 7), with a greater score for NNS abstract. However, informational load varies across disciplines, with a significant difference in all disciplines other than biology covered in our corpus. The move analysis of the biology discipline also shows that the NS versus NNS difference is only significant in move I, with a greater score for NNS abstracts. The results appear to suggest that wherever there is a significant difference in Dimension 7, NNS abstracts have a heavier informational load.

On the basis of the contrastive analysis of NS and NNS abstracts along seven dimensions, and for Biology abstracts also across discourse moves, a variety of possible reasons are put forward in this article to account for the observed divergences in NNS abstracts from NS abstracts including, for example, the transfer of native language writing, improper classroom instructions, different conventions widely accepted by academic communities, and conflicting requirements of individual journals.

5 Conclusion

In conclusion we would like to suggest, on the basis of our research findings, that there is a need to make Chinese writers fully aware of specific linguistic features associated with discourse moves in RA abstracts, and the preferred writing style of international journals so as to make their writing more direct and readable. It is hoped the original aspects of the present research will become welcome new developments of the MDA approach.

References

Biber D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Cao, Y. and Xiao, R. 2013. A multidimensional contrastive study of English abstracts by native and nonnative writers. *Corpora* 8(2). [In press]

Hyland, K. 2004. *Disciplinary discourses: Social interaction in academic genres*. Ann Arbor: The University of Michigan Press.

Kanoksilapatham, B. 2007. "Rhetorical moves in biochemistry research articles". In D. Biber, U. Connor and T.A. Upton (eds.) *Discourse on the move: Using corpus analysis to describe discourse structure*, 73-119. Amsterdam: John Benjamins.

McEnery, T., Xiao, R. and Tono, Y. 2006. *Corpus-based language studies: An advanced resource book*. New York: Routledge.

Rayson, P. 2008. "From key words to key semantic domains". *International Journal of Corpus Linguistics* 13 (4): 519-549.

Samraj, B. 2005. "An exploration of a genre set: Research article abstracts and introductions in two disciplines". *English for Specific Purposes* 24: 141-156.

Santos, M.B.D. 1996. "The textual organization of research paper abstracts in applied linguistics". *Text* 16: 481-499.

Van Bonn, S. and Swales, J.M. 2007. "English and French journal abstracts in the language sciences: Three exploratory studies". *Journal of English for Academic Purpose* 6 (2): 93-108.

Xiao, R. 2009. "Multidimensional analysis and the study of world Englishes". *World English* 28 (4): 421-450.

The metaphoricity of *fish*: implications for part-of-speech and metaphor[1]

Xu Huanrong
Hangzhou Dianzi University
xhrhsy@hotmail.com

Hou Fuli
Communication University of China
1931466627@qq.com

1 Introduction

Conceptual/cognitive Mapping Theory (hereafter CMT) scholars mainly took linguistic metaphors as evidences for cognitive metaphors; however, their main resort to artificially created examples rather than naturally occurring linguistic data has been criticized by many (Semino 2008; Cameron 2003; Deignan 2005; Stefanowstisch 2006) whose efforts have greatly pushed forward the corpus-based study of CMT. Stefanowstisch (2006) declared that corpora made it possible for the study of the relationship between cognitive metaphors and linguistic metaphors. Deignan (2005, 2006, 2008) is such a major scholar who has explored the grammar of linguistic metaphors to inform CMT. This paper is related to part of her study, i.e. the grammar of linguistic metaphors in terms of parts of speech.

2 Research questions

This paper has three interrelated research questions :

> **RQ1**: Whether there is a correspondence between contents of metaphoric mappings and their language realizations in parts of speech;
>
> **RQ2**: Whether the metaphor rate is always high when a concrete noun shifts its part of speech to verbs and adjectives;
>
> **RQ3**: Whether the metaphoric meaning(s) can be carried over between different parts of speech of a word.

According to CMT, attributes, entities/slots and/or relations (including knowledge) are mapped (Lakoff and Turner 1989) from the source domain to cognitive the target domain, and the mapping is governed by the Invariance Principle (Lakoff et al. 1980, 1990, 1993). Since attributes are normally realized by adjectives, entities/slots by nouns, actions by verbs, etc., Deignan (2005, 2006, 2008) maintained that the Principle does not always work given the fact that some verbal and adjectival metaphors do not have literal counterparts in the source domain, like the verb *fox* having no literal verbal counterpart whose meaning is more basic to be the origin of the metaphoric, hence no correspondence or invariance therein. Although she still includes such words into the scope of the metaphorically used, her view justifies the exclusion of them according to the metaphor identification procedures proposed by the PRAGGLEJAZ Group (2007) and Steen (2010). This paper attempted to prove that language correspondence may not be there, but conceptual correspondence is always there, and lack of language correspondence cannot be the touchstone for metaphoricity. This is what RQ1 is for, and the rest two RQS are further studies.

3 Methodology

The British National Corpus (hearafter BNC) at Lancaster University (bncweb.lancs.ac.uk) was employed to get the data based on which manual analysis was carried out. Since the ways in which BNC was used varied in different cases, the details would be given in actual studies.

4 Case study

We chose the word *fish*, concordanced **fish** and looked at its inflections, finding its noun and verb forms identical, and its adjective form being *fishy* which is derived by suffixation and can be changed to *fishiness* similarly. There are other inflections, such as fish-like, fish-tail, etc., all of which were ignored due to being subsumed in the study.

***Fish* as noun** (see table 1): We concordanced {fish/N} in BNC, getting 10,399 hits containing the two types *fish* and *fishes*. The huge number rendered impossible an exhaustive manual analysis , so we checked the collocations at left 3 and right 3 positions with the least frequency of 5. The search returned 11,170 different types, with the first 50 highest values totalling 9925, occupying 88.85% among the whole, so all the other collocations beyond (except *as)* were dismissed. Because most words are in the same semantic domains with fish/fishes: some are in the food domain, like *chip(s), meat, shop, (fish) fingers, eating*, some as a species of fish, like *species, tropical, tuna, bony, shoals, squid*, and some as part of the living environment of fish either in nature or in captivity, like *tank, freshwater, marine, aquarium, catch, feeding, eggs, birds, pond, predatory, sea, stocks*, etc. When they go together with *fish/fishes*, they are

[1] This paper is part of the research project *Functions and Distributions of Metaphor in Discourse* (11JCWY14YB) funded by Zhejiang Provincial Planning Office of Philosophy and Social Sciences, P.R. China.

highly literal except in some fish-based structural metaphors or analogies which are highly retrievable through some metaphor markers, such as *like* and *as*. Considering these, we picked 6 types which are the most likely to signal a metaphor: *small* and *big* (evaluative, likely to be the ground of a metaphor), *like* and *as* (signalling metaphor or analogy; *as* can also help find other metaphor-signalling expressions like *as if*, *as it were*), *kettle* and *water* (highly proverbial).

Collocations	*No. of metaphors*	*No. of tokens*	*Metaphor ratio*	*Content of Mappings*	*Metaphoric meanings*
Small	5	192	2.60%	Attribute: Noun phrase: (a) **small** fish(es)	Attribute: Noun phrase: Somebody/something of**no/little importance**
Big	37	121	30.58%	Attribute: Noun phrase: (a) **big** fish(es)	Attribute: Noun phrase: Somebody of **much importance**
Like	112	252	44.44%	...(including all mapping types, many being one-off inventions)	...(Situation-dependent)
As (incl. *as if* and *as it were*)	22	413	0.48%	Ditto	Ditto
Kettle	40	44	90.91%	Entity: Noun phrase: (Different/fine) kettle of fish	Entity: Noun phrase: a dilemma/ difficult situation
Water	16	99	16.16%	Relation: Nominative absolute structure : Fish out of water, fish swimming in water, fish upheading from deep water, etc.	Relation: Adjective, noun, nominative phrase: Helpless/despaired; ease; something hidden/suspicious

Table 1: *Fish* as noun

Type	*No. of tokens*	*Number of metaphors*	*Metaphor ratio*
Fishes	10	4	40%
Fishing	181	19	10%
Fish	523 (143 nouns are mistagged as verbs)	42	12% [42/(523-143]
Fished	326	122	37%
Major metaphoric patterns and meanings: **Fish for**: search so to find (esp. followed by abstract nouns like compliments); **Fish out**: produce after search			

Table 2: *Fish* as verb

Type	*Collocations*	*Metaphoric meanings*
Fishy	Something	Questionable, doubtable
	Smell	Suggestive of the smell of fish, of an unpleasant smell
	About	Questionable, doubtable, tricky
	There	Often used in *there {be/V} something fishy about something or somebody.* See *something.*
Fishiness	Food-related lexis	Suggestive of *fishy* smell/taste

Table 3: *Fishy* and *fishiness*

Although this method cannot exhaust all fish-based metaphors, and some examples overlap, all different types of mappings should be there, and the result of the study will not be highly affected.

***Fish* as verb** (see table 2): we concordanced {fish/V}.

***Fishy* as adjective** (see table 3): We searched for *fishy* in BNC and got 108 hits among which 72 are metaphoric, leading to a 66.67% metaphoric ratio. Its inflectional noun *fishiness* returned 2 hits, both metaphoric.

5 Discussion

Concerning RQ1: From Table 1, it can be seen that all the mappings are conceptually correspondent; however, some mappings are linguistically correspondent, some are not. This is explainable: words are only signs leading to meanings, surface linguistic mis-matchings may have conceptual correspondences in depth. This can be further proved by *fishy* and *fishiness* in Table 3, both referring back to the slot *smell* in the source domain despite their different word classes. An attribute can be realized by adjectives, and an action by a noun. The realization of relations is the most complicated: noun phrases, verbs and adjectives are all possible. So correspondences between contents of metaphoric mappings and parts of speech are not always there. Consequently, we do not need there being a word for some content to be mapped; we can invent one. This is how we get such words as *fox* (to fool) and *foxy* (sly). So lack of correspondences at the language level does not question the Invariance Principle of metaphoric mappings.

Concerning RQ2: According to Deignan (2005, 2006, 2008), when a noun shifts its word class to a verb or adjective, the resultant words are highly metaphoric. She explored some animal words. It is true that *to dog* and *to fox*, *dogged* and *foxy* are about 100% metaphoric; however, only 20.66% of the verb *fish (much higher than the overall rate) and 66.67% fishy* are metaphoric. If we look at other words, like *flower* and *stone* as a verb, the percentages can be even lower.

Concerning RQ3: It is true that many verbs and adjectives derived from their cognate nouns carry some meanings over, like between the nominal *fox* and the adjectival foxy; however, it is not the case at least when *fish* is a verb. The nominal *fish* denoting an animal becomes the patient of the action *fish* which means *to catch fish*. So NO metaphoric carry-over exists between the nominal *fish* and the verbal *fish*. While *fishiness* only carries the *smell* part of *fishy*.

There are other interesting cases. The noun *lamb* means *to give birth to a lamb* when shifted to a verb, non-metaphoric. The noun *stone* becomes the instrument in *stone somebody/something*.

6 Conclusion

The study in the paper says NO to absolute correspondences between the attributes, relations and entities metaphorically mapped from the source domain and their normal linguistic realizations in parts of speech; it does not provide evidences to question the Invariance Principle, either, since metaphoric mapping involves a conceptual correspondence or building of such a correspondence. In addition, when (concrete) nouns shift their part of speech, the resultant verbs or adjectives can be metaphoric to a different degree, and it is the same to the carrying-over of metaphoric meanings between different parts of speech. Everything is word-specific.

References

Cameron, L. 2003. *Metaphor in Educational Discourse*. London: Continuum.

Deignan, A. 2005. *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Deignan, A. 2006. "The Grammar of Linguistic Metaphors". In Stefanowitsch, A. and Stephen, G. (eds.). *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter.

Deignan, A. 2008. "Corpus Linguistics and Metaphor". In Gibbs, Raymond, Jr. (ed.). *The Cambridge Handbook of Metaphor and Thought*. Cambridge: Cambridge University Press.

Lakoff, G. 1980. *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lakoff, G & M. Turner. 1989. *More than Cool Reason*. Chicago and London: The University of Chicago Press.

Lakoff, G. 1990. "The Invariance Hypothesis: Is abstract reasoning based on image-schemas?" *Cognitive linguistics* 1: 39-74.

Lakoff, G. 1993. "The contemporary theory of metaphor". In A. Ortony (ed.), *Metaphor and Thought :* 202-251. Cambridge: Cambridge University Press. [Second edition].

Pragglejaz Group. 2007. "MIP: A method for identifying metaphorically used words in discourse." *Metaphor and Symbol* 22(1): 1-39.

Semino, E. 2008. *Metaphor in Discourse*. Cambridge: Cambridge University Press.

Steen, G. et al. 2010. *A Method for Linguistic Metaphor Identification*. Amsterdam: John Benjamins.

Stefanowitsch, A. 2006. "Corpus-based approaches to metaphor and metonymy". In Stefanowitsch, A. and Stephen, G. (eds.). *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter.

The structural and semantic analysis of the English translation of Chinese light verb constructions: A parallel corpus-based study

Jiajin Xu
Beijing Foreign Studies University
xujiajin@bfsu.edu.cn

Lu Lu
Beijing Foreign Studies University
luludew@bfsu.edu.cn

1 Light verbs and light verb constructions

Light verb (LV hereafter) (Jespersen 1965; Cattell 1984) constructions, such as 'have a try' and 'make a change' in English, are semi-productive structures in which a verb and its complement form a single semantic unit. One of the distinctive properties of light verb constructions (LVCs hereafter) is that the main predicational meaning is conveyed by the complement. In Chinese, typical LVs, such as 进行/*jin4xing2*, 'to carry out', and 予以/*yu3yi3*, 'to give' have been examined in the literature (Yin 1980; Zhu 1982, 1985).

Based on previous LV literature, this study focuses on 16 most frequently discussed Chinese LVs, namely 进行/*jin4xing2*, 做/*zuo4*, 作/*zuo4*, 干/*gan4*, 搞/*gao3*, 弄/*nong4*, 整/*zheng3*, 打/*da3*, 给以/*gei3yi3*, 给予/*ji3yu3*, 加以/*jia1yi3*, 予以/*yu2yi3*, 加/*jia1*, 予/*yu3*, 寄予（与）/*ji4yu3*, and 致力/*zhi4li4*.

Group	LVs	Valid cases
DO	搞/*gao3*	704
	进行/*jin4xing2*	597
	做/*zuo4*	224
	弄/*nong4*	86
	作/*zuo4*	77
	干/*gan4*	58
	打/*da3*	50
	整/*zheng3*	2
GIVE	加以/*jia1yi3*	170
	给予/*ji3yu3*	74
	予以/*yu2yi3*	50
	给以/*gei3yi3*	27
	致力/*zhi4li4*	22
	予/*yu3*	7
	寄予/*ji4yu3*	4
	加/*jia1*	3

Table 1. The use of 16 LVs in the five million word/character parallel corpus

The 16 Chinese LVs are classified into two groups according to their semantic content: DO group and GIVE group. Four LVs, that is, 搞/*gao*3 and 进行/*jin4xing2* of DO group, and 加以/jia1yi3 and 给予/*ji3yu3* of GIVE group are analysed in the current study. 645 valid sentence pairs were retrieved for our structural and semantic analyses.

2 Previous works on LVCs

Recent studies on LVCs have largely focused on the interface between syntax and semantics (e.g. Diao 2004; Feng 2005; Kuo and Jen 2006; Zhu 2011), among which some are within the framework of generative grammar, others from a cognitive linguistic perspective, and so forth. More often than not, LVC studies look into the semantic configuration, thematic role assignment for example, in relation to morpho-syntactic realizations. For example, the LVC 予以/*yu3yi3* in Chinese must be accompanied by a Theme, for instance, 对 好人好事 予以 奖励/*dui4 hao3ren2hao3shi4 yu3yi3 jiang3li4* where 好人好事/*hao3ren2hao3shi4* is the Theme of 予 以 /*yu4yi3* construction; yet 进 行 /*jin4xing2* constructions are freer, namely, it can take or not take a Theme. The form-meaning composite view has been the linguistic 'mindset' for an LVC study. Amongst various approaches to LVCs, the analysis on argument structure is something that no one could bypass.

However, very few studies have been done on the grammatical and semantic patterns of Chinese LVCs from a contrastive or translational perspective. This paper thus attempts to study the lexicogrammatical and semantic (non-)correspondences of the English translation of Chinese LVCs based on a large BFSU Chinese-English parallel corpus.

3 Corpus and methods

The corpus used in this study is the Beijing Foreign Studies University Chinese/English Parallel Corpus (CEPC) (Wang 2004), a balanced sentence-aligned database of 5 million characters/words. It consists of fiction (55%, covering novels, essays, and dramas.) and non-fiction texts (45%, covering humanities, social sciences, and natural sciences). The two broad text categories are divided into four sub-genres, namely, fiction (60,622 sentence pairs), humanities (22,031 sentence pairs), social sciences (37,568 sentence pairs) and natural sciences (2,011 sentence pairs).

The translational (non-)correspondences of the Chinese LVCs are addressed from two perspectives: the morpho-syntax and verb semantics of the complement of an LV. All the structural makeup and semantic pattern of LVCs between Chinese and their English translations were manually annotated and thoroughly checked.

4 Results and preliminary findings

This section briefly summarises the structural and semantic patterns of the English translation of the four cohorts of Chinese LVCs, containing such LVs as 给予/ji3yu3, 加以/jia1yi3, 搞/gao3 and 进行/jin4xing2.

给予/*ji3yu3*	37	加以/*jia1yi3*	191
Cmpl_v.[1]	11	Cmpl_v.	148
GIVE	8	Conc_v.	14
Conc_v.	7	Cmpl_n.	10
LVC	6	GIVE	5
Cmpl_n.	4	LVC	5
Prep_P	1	Gnrl_v.	2
		Cmpl_adj.	1
		Prep_P	1

Table 2. Structural patterns of English translations of GIVE group LVCs

搞/*gao3*	50	进行/*jin4xing2*	367
Cmpl_n.	21	Cmpl_v.	122
Cmpl_v.	10	Gnrl_v.	83
Gnrl_v.	10	Cmpl_n.	82
Conc_v.	5	Conc_v.	29
LVC	2	LVC	27
		Prep_P	12
		Cmpl_adj.	1

Table 3. Structural patterns of English translations of DO group LVCs

On the whole, both GIVE and DO group LVCs tend to be translated from their complements, yet 搞/*gao3* is inclined to translate into the nominal form of the complement, and 进行/*jin4xing2* into the verbal form. Both 搞/*gao3* and 进行/*jin4xing2* is less likely to be translated into concrete verbs or English LVCs, the case of which is different from GIVE group.

The majority of GIVE group has a semantic content of DISPOSITION, while DO group is often followed by an ACTIVITY. This could be explained by the inherent meaning of the Chinese LVs.

The non-correspondences of grammatical features and argument structures are strongly correlated. For example, English passivisation would trigger the addition of object, especially the object that has already been mentioned in the previous clause.

With regard to the interaction within argument structure, this study demonstrates the combined addition of Agent and Theme. The parataxis of Chinese and hypotaxis of English could account for this translational behavior. Besides, the changes of Theme would trigger the alteration of the theta-grid, possibly because of the individual features of Theme in the realisation of theta-grid.

Also of interest are the cases that the change of lexico-grammatical features would bring about the change in argument structure. The addition of object would cause the addition of Theme, which could be resulted from the great extent of correspondence between the Theme and the object in English. Besides, passivisation would cause the addition of Theme.

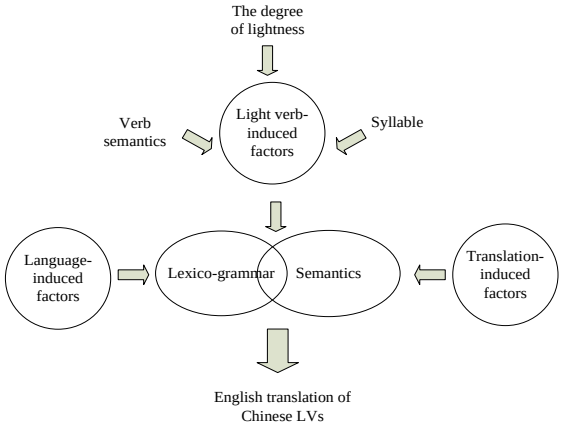


Figure 1. The analysis on the English translation of Chinese light verbs

Apart from the LV-related factors, the typological differences between English and Chinese have their role in the morpho-syntax and semantics of English translation. Chinese word order, for instance, allows the object to be introduced by 对于/*dui4yu2* or 将/*jiang1* and placed before the verb, but this could not be the case in English. Therefore, word order change is not rare in the English translations of Chinese LVCs.

The process of translation also exerts its influence on the diversity of the translated LVCs. The universal features of translation points out that the translated texts become explicit, such as the addition of connectives or other cohesive devices. This study further states that the explicitation could be extended to cover the semantic explicitness which reveals the addition of thematic roles.

In summary, the English translation of Chinese LVCs are conditioned by a number of potential factors: the verb sense of the LVs, differences between English and Chinese morph-syntax, and translation induced factors, and on top of all that, the interplay between lexicogrammatical and semantic features.

[1] Cmpl_v., Cmpl_n., and Conc_adj. refer to English translations as the verbal, nominal and adjectival forms of the complements in the original Chinese LVCs. Gnrl_v. refers to general verbs without concrete meaning, such as *conduct* and *carry on*.

References

Cattell, R. 1984. *Syntax and semantics*. Newcastle, N.S.W.: Academic Press.

Diao, Y. 2004. *Xiandai hanyu xuyi dongci yanjiu* [The study of weak verbs in modern Chinese]. Dalian: Liaoning Normal University Press.

Feng, S. 2005. *Qingdongci yiwei yu gujin hanyu de dongbin guanxi* [Light verb movement in modern and classical Chinese].*Yuyan kexue* [*Linguistic Sciences*] 4 (1). 3-16.

Jespersen, O. 1965. *A modern English grammar on historical principles (volume VI: morphology)*. London: George Allen and Unwin.

Kuo, P. -J. & Jen, T. 2007. "Light verb, heavy verb and verbal noun in Mandarin Chinese". In Doo-Won Lee (ed.) *The proceedings of the 9th Seoul international conference on Generative Grammar (SICOGG9): locality and minimalism.* Seoul: Hankuk Publishing.

Wang, K. 2004. *Shuangyu duiying yuliaoku yanzhi yu yingyong* [The creation and application of bilingual parallel corpus]. Beijing: Foreign Language Teaching and Research Press.

Yin, S. 1980. *Tan jinxing lei dongci weiyuju* [On the group of predicates like *jinxing*]. In S. Yin (ed.) *Hanyu yufa xiuci lunji* [The selected papers of grammatical rhetoric of Chinese]. Beijing: China Social Sciences Press.

Zhu, D. 1982. *Yufa jiangyi* [The lectures on grammar]. Beijing: The Commercial Press.

Zhu, D. 1985. *Xiandai shumian hanyu li de xuhua dongci he dongmingci* [The weak verbs and nominalized verbs in modern Chinese written texts]. *Journal of Peking University (Philosophy and Social Sciences)* (5). 10-16.

Zhu, L. 2011. *A contrastive analysis of light verbs in English and Chinese: a minimalist approach*. Unpublished PhD thesis. Beijing Foreign Studies University.

The search for units of meaning in terms of corpus linguistics: The case of collocational framework "the * of"

Suxiang Yang
Henan Polytechnic University
ysx@hpu.edu.cn

1 Introduction

This paper is grounded in Wittgenstein's (1958) philosophical perspective on "meaning as use" and Firth's (1957) "contextual theory of meaning". It employs Sinclair's (1996) "lexical grammar" model to systematically investigate and construct units of meaning based on a collocational framework "the * of" in a general written English corpus.

As previous corpus researches on units of meaning were mostly based on continuous words or phrases, this research tries to extend the units of meaning study on a basis of discontinuous words, i.e., collocational frameworks, specifically "the * of", with the research questions: How does the collocational framework "the * of" construct units of meaning? What are the lexical grammatical profiles of units of meaning based on collocational framework "the * of"?

2 Research methodology

Enlightened by, but not strictly following, Sinclair's (1996) lexical grammar model and Hunston's (2008) "semantic sequence" approach, that is, we change Sinclair's core, collocation, colligation, semantic preference and prosody into core, collocation, colligational sequence, semantic sequence and functional sequence. For the units of meaning, the author regards "the * of" as a core of units of meaning and expands it as a seven-word string with the words on the positions of Left 2 and Left 1 (L2C and L1C respectively) before the core, the collocated words in the middle position of the core (MidC), and the words in Right 1 and Right 2 positions (R1C and R21C respectively) after the core, hence a span of units of meaning based on collocational framework "the * of" covering, "L2C + L1C + **the** + MidC + **of** + R1C + R2C".

The reason for focusing on the span of left 2 to right 2 is that the most frequent words dropped between Left 2 to Right 2 distances can indicate the collocational features of "the * of" according to our pilot statistical studies.

The corpus used in this study is the BNC-Written, which comprises the written part of the British National Corpus. The data presented in this study are entirely based on the second release of the

BNC, also known as the World Edition. The reason for choosing written text is that written texts are comparatively stable and show less variation, which can represent general English.

The tool used in this study is WordSmith Tools 5.0. In this study, we use WordSmith 5.0 to extract the middles collocates, left collocates and right collocates. To state it differently, to use the Concord to extract Left 2, Left 1, Middle, Right 1, Right 2 collocates of the collocational framework "the * of".

Following the principle of frequency-driven in corpus linguistics, the author first retrieved cases of concordance lines based on the core "the * of" from the corpus, specifically, 897,628 concordance lines in BNC-W. From that pool, the author examined the top 300 collocates in the five positions: Left 2, Left 1, Middle, Right 1, and Right 2. Then categorizing the collocates at three levels – word class categories, semantic categories and functional categories, the co-occurrence of the top two categories (accounting for nearly 50 percent of the total frequency, therefore, which is statistically significant) at each level are analyzed so as to form typical patterns – colligational sequences, semantic sequences and functional sequences, and finally units of meaning based on the collocational framework "the * of" are generalized.

For the classification of word class category, this study adopts Biber et al.'s (1999) criteria in classifying word classes. For the classification of semantic category, it combines Biber et al.'s (1999) classifications of semantic domain and Aristotle's categories of meaning with the reference to Diniz's (2007) classification of the middle collocates. For the classification of functional category, Halliday and Matthiessen's (2004) idea of categorizing metafunctions of language sequences based on relationships between classes and functions is employed, as turns out to be more operational in constructing units of meaning based on the collocational framework "the * of".

3 Results

The research results show that the collocational framework which only consists of two functional words can be built up as units of meaning in a huge amount of cases in meaningful context. The construction process initiated by collocation and ended with colligational sequence patterns, semantic sequence patterns and functional sequence patterns reflects the nature of collocational framework, which consists of form, meaning and function. The colligational sequences are the co-selections of grammar, semantic sequences are the co-selections of meaning, and the functional sequences are the co-selections of function.

Specifically, in the BNC-W, the typical colligational sequence pattern is, "Verbs / Nouns + Prepositions / Verbs + **the** + Nouns + **of** + Determiners / Nouns + Nouns / Prepositions" – which indicates a structure of "predicate with object or a prepositional phrase".

Structurally, this sequence is the expense of Sinclair's (1991) the N1 + of + N2 structure in syntagmatic axis. However, the difference is that nouns N1 must follow the definite article "the" in paradigmatic axis. Moreover, the present study is on the seven-word string.

The typical semantic sequence pattern is, Existences / Activities + Existences / Activities + **the** + Essences / Actions + **of** + Social Lives / Persons + Social Lives / Existences (/Natural phenomena), this sequence can be simplified as "Existences + **the** + Essences / Actions + **of** + Social Lives / Natural phenomena" – which expresses an idea of "existing some action or essence in either social or natural world".

Semantically, the collocates around "of" are nearly in line with Groom's (2010) findings, i.e. the first two frequent semantic sequences are: PROPERTY + of + PHENOMENON, PROCESS + of + OBJECT. Here PROPERTY is roughly equally to "Essences" in the present study.

The functional sequence pattern is, (Relational) Processes / (Action) Participants + (Main verb) extensions / (Relational) Processes + **the** + Essence participants / Action Participants + **of** + Deictics / (Social life) participants + (Social life) participants / Circumstances, which can also be simplified as "(Relational) Processes + **the** + (Action) participants + **of** + (Social life / Natural phenomenon) participants" – which plays an ideational metafunction of language.

As to the boundary exploring of units of meaning based on collocational framework "the * of", it is found that both the five-word span (i.e., "L1C + **the** + MidC + **of** + R1C") and seven-word span (i.e., "L2C + L1C + **the** + MidC + **of** + R1C + R2C") can generate units of meaning, only with the latter as an extension of the former.

4 Conclusion

This study investigates the units of meaning based on collocational framework in general written English. It could be claimed that collocational framework can be the framework of meaningful units, whose meanings can be realized by colligational sequences, semantic sequences and functional sequences in terms of the collocates in Left 2, Left 1, Middle, Right 1 and Right 2 positions. That is, the units of meaning based on "the * of" reflect the social/cultural events and natural

phenomena in meaning, and relation process and participants in function.

The present study proposes to search for the units of meaning based on collocational frameworks. Previous studies lack the viewpoint of constructing units of meaning with the collocational framework as cores, this study yield some implications to language theories and practices, more importantly, to language teaching.

The significance of this study lies in that it extends the previous research of units of meaning to a less targeted phenomenon – collocational frameworks, thus giving another more specific and direct interpretation of Wittgenstein's (1958) "meaning as use" and Firth's (1957) "contextual theory of meaning", because the corpus we used (BNC-W) is the language in use and it provide context of the units of meaning of "the * of". More importantly, we expand Sinclair's (1996) "lexical grammar model" to a more concrete and operable one.

References

Biber, D., Conrad, S., Johansson, S. & G. Leech. 1999. *Longman Grammar of Spoken and Written English*. London: Pearson Education Limited.

Diniz, L. 2007. *Highly frequent function words in the light of the idiom principle: the case of "the"*. Unpublished dissertation, Georgia State University, Atlanta.

Firth, J. 1957/1968. *A Synopsis of Linguistic Theory, 1930-55*. Studies in linguistic analysis (Special Volume of the Philological Society), 1957. Reprinted in F. Palmer. Selected Papers of J. R. Firth 1952-59. London and Harlow: Longmans, Green and Co., Ltd., 168-205.

Groom, N. 2010. Closed-class keywords and corpus-driven discourse analysis. In M. Bondi and M. Scott (eds.) *Keyness in Texts*. Amsterdam and Philadelphia: Benjamins.

Halliday, M. & C. Matthiessen. 2004. *An Introduction to Functional Grammar* (3rd Edition). London: Arnold.

Hunston, S. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of corpus Linguistics*, 13(3): 271-295.

Owen, C. 2007. Notes on the ofness of 'of' – Sinclair and grammar. *International Journal of Corpus Linguistics* 12(2):

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1996. The search for units of meaning. *Textus* IX: 75-106.

Wittgenstein, L. 1958. *Philosophical Investigations*, (3rd edn.), trans. G. E. M. Anscombe, New York: Macmillan Publishing Company.

Posters

New methods of annotation: The 'humour' element of Engineering lectures

Siân Alsop
Coventry University
alsops@uni.coventry.ac.uk

The questions which drive the creation of specialised corpora can necessitate the use of very specific systems of annotation.

This is certainly the case with the Engineering Lecture Corpus (ELC), a small (but growing) specialised corpus of lectures from three countries: the UK, Malaysia, and New Zealand. The ELC is designed to investigate the discourse of undergraduate lectures, and answer the following question: do discourse differences exist in lectures delivered in different parts of the world, but in the same language medium (English) from the same discipline (Engineering) at the same level of study (undergraduate)?

The ELC is annotated for 'pragmatic' features (c.f. Simpson-Vlach and Leicher, 2006). This means that stretches of text within the transcriptions have been identified as 'elements' representing six functions: defining, housekeeping, humour, prayer, storytelling, and summarising. Subcategories have been attributed to some of these elements; eight 'types' have been attributed to humour, for example.

The elements are currently identified by inline XML annotation, which looks like this:

```
<humour type="black">in Japan they call it karōshi that mean death attributed to uh stress at the work place so just like me come here and teach and collapse and pass away</humour>
```

Identifying humour inline allows a middle ground between video data and raw transcription to be provided. Irony, for example, may not be recoverable from a transcript without an added layer of annotation to index its occurrence in the spoken data.

The pragmatic annotation of categories such as humour in the ELC allows the analysis of difference in speech function across cultures and provides examples of authentic data through which this difference can be illustrated

The particular problem that humour presents is that it is both a communicatively important and functionally specific feature of academic speech (e.g. Reershemius 2012) and at the same time laborious to identify systematically. As has been noted of MICASE: "instances of […] humor were simply far too numerous for us to code all instances, so we excluded those categories from our coding" (Maynard and Leicher 2007: 112).

Studies of the main corpora of academic speech have relied heavily on structural markup for laughter to identify the occurrence of humour (e.g. Lee 2006 on MICASE and Nesi 2012 on BASE). However, although related, laughter and humour are by no means coextensive (Attardo 2003: 1288); laughter is not a reliable indicator of humour (Ross 1998; Swales 2006).

The annotation of humour in the ELC data shows that: 1. not all humour types commonly elicit laughter; and 2. laughter is not necessarily a reaction to humour (it can be prompted by, for example, anxiety or relief). Depending on type, the instances of humour identified elicited laughter (by lecturer or audience) in between 11-64% of cases. In total, almost 9% of instances of laughter recorded during the lectures was unrelated to deliberate linguistic *humour-creating manoeuvres* (cf. Fillmore 1995), and only 31% of such deliberate manoeuvres elicited a laughter response. The annotation of humour, then, allowed the identification of over twice as many instances of humour as recovery based on the record of laughter alone.

In response to the research question, identifying the boundaries of specific speech functions inline allowed comparison of their duration and dispersion (where in the lecture, and for how long they occur) across the cultural components of the corpus.

To calculate this, a script was used to loop through all the corpus files and count: 1. the total number of tokens (not including markup and annotation), and 2. the start point and end point of each pragmatic chunk. As well as providing the raw quantitative data for analysis purposes, this process provided the information needed to plot normalised dispersion and duration in visual form, allowing it to be simply rendered in any statistical software.

Preliminary findings suggest that significant differences occur in lectures cross-culturally.

The normalised figures show that, as an umbrella category, speech that performed the humour function was most commonly employed in the UK lectures; over twice as many instances were identified in comparison to the lectures from Malaysia and New Zealand. The average token length of each instance showed significant variation between humour types, but not between the cultural sub-corpora.

Perhaps most significantly, specific differences in the occurrence of type also emerged (see Fig. 1). 'Irony', for example, was significantly more

Figure 6: normalised distribution of humour types across the ELC

common in the UK component than in the Malaysian component, as were instances of 'black', 'bawdy', 'mock threatening' and 'disparaging' humour. 'Teasing', however, was marginally more common in Malaysian lectures, and the most equally weighted function in terms of occurrence across the components. 'Self-deprecation' occurred most often in the New Zealand lectures, and was significantly more common than in the Malaysian lectures.

Particular humour types perform specific, and often very different, communicative functions such as enabling rapport-building, constructing in-group cohesion, mitigating conflict and modelling identities (see e.g. Lee 2006; Kotthoff 2007; Nesi 2012; Norrick and Spitz, 2008; Partington 2006; Reershemius 2012; Stebbins 1980). By identifying where and to what extent these functions occur, we can begin to better understand the dynamic of the academic lecture theatre across cultural settings.

We know that humour does not travel well across cultures. It can cause particular problems of miscommunication for the lecturers delivering and students receiving it in unfamiliar cultural contexts (e.g. Wang 2012; Zhang 2005).

The results of this study will be of interest to ESP practitioners. They will help to promote greater awareness of the potential mismatch between intention and reception in the delivery of humour to students in unfamiliar cultural contexts. Students working across cultural contexts in the field of Engineering may benefit from exposure to examples of culture-specific humour types. Lecturers may benefit from increased understanding of their function.

References

Attardo, S. 2003. "Introduction: The Pragmatics of Humor". *Journal of Pragmatics* 35: 1287–1294.

Fillmore, C. 1994. "Humour in Academic Discourse". In Grimshaw, A., and Burke, P., J. (eds.) *What's Going on Here? Complementary Studies of Professional Talk*. Norwood, NJ: Ablex.

Lee, D. 2006. "Humour in Spoken Academic Discourse". *NUCB JLCC* 8(3): 49-68.

Maynard, C. and Leicher, S. 2007. "Pragmatic Annotation of an Academic Spoken Corpus for Pedagogical Purposes". In Fitzpatrick, E. (ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodophi: 107-116 .

Nesi, H. 2012. "Laughter in University Lectures". *Journal of English for Academic Purposes* 11(2): 79-89.

Norrick, N. R., and Spitz, A. 2008. "Humor as a resource for mitigating conflict". *Journal of Pragmatics* 40: 1661-1686.

Partington, A. 2006. The Linguistics of laughter: A corpus-assisted study of laughter-talk. London: Routledge.

Reershemius, G. 2012. "Research cultures and the pragmatic functions of humor in academic research presentations: A corpus-assisted analysis". *Journal of Pragmatics* 44: 863–875.

Ross, A. 1998. *The Language of Humour*. London: Routledge

Simpson, R. C., Lee, D. Y. W., and Leicher, S. 2007. *MICASE Manual: The Michigan Corpus of Academic Spoken English* [online] Version 3 edn. Michigan: The English Language Institute, University of Michigan. Available at <http://www.lsa.umich.edu/eli/micase/index.htm>

Stebbins, R., A. 1980. "The role of humour in teaching: Strategy and self-expression". In Woods, P. (ed.) *Teacher Strategies (RLE Edu L): Explorations in the Sociology of the School*. New York: Routledge, 62-84.

Swales, J. M. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Wang, T. 2012. A study of humour in British academic lectures and Chinese students' perceptions of this. Unpublished PhD thesis. The Open University.

Zhang, Q. 2005. "Immediacy, humor, power distance and classroom communication apprehension in Chinese college classrooms". *Communication Quarterly* 53(1): 109-124.

Oxford Children's Corpus: a corpus of children's writing, reading, and education

Nilanjana Banerji
Oxford University Press, Education Division
nilanjana.banerji@oup.com

Vineeta Gupta
Oxford University Press, Education Division
vineeta.gupta@oup.com

Adam Kilgarriff
Lexical Computing Ltd.
adam@lexmasterclass.com

David Tugwell
Lexical Computing Ltd.
dtugwell@gmail.com

1 Introduction

The Oxford Children's Corpus (OCC), as it was in 2011, is described in Wild et al (2011, 2012). This was a corpus of writing for children. Since then OUP has developed a 'children's writing' component of the corpus, primarily with data from the BBC Radio 2 '500 Words' short story writing competition. This is a competition that runs in the spring every year with children aged 4-13 submitting entries up to 500 words long, with winners announced at the Hay Literary Festival. All shortlisted items can be read online.[1]

Lexical Computing Ltd is working with the Children's Dictionary and Language team at Oxford University Press to analyse the language that the children use. The 74,000 entries received in 2012 (called Beebox below) form a large part of OCC-W, the Children's Writing component of the OCC. The OCC as it was when last reported on forms the hub of the Reading component (OCC-R) and we have also gathered curriculum materials to form the Education component (OCC-E).

Here we focus on Beebox, describing the data and presenting some first results from the analysis of the 2012 data. In April this will be joined by the 2013 data, and any conference presentation in July 2013 will talk about the new data too.

2 The Beebox data

There are a total of 73,875 stories, with distribution by age and gender as in Figure 1.

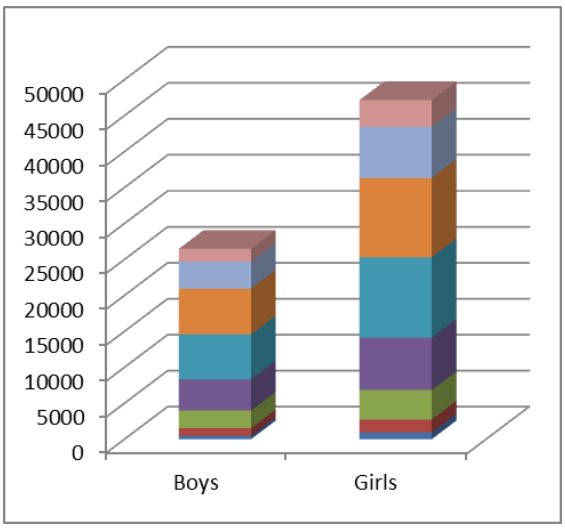


Figure 1. Age bands are, from bottom: up to 6, 7, 8, 9, 10, 11, 12, 13.

Most stories are close to 500 words. The total corpus size is 32.4 million words. From a statistical point of view this is a dream: very large numbers of same-size samples.

We also have the BBC region (in most cases, this is the same as the county) for each story. There are 54 of these regions, and for all but two, there are over 100 stories. For 37 of 54 regions there are over 1000 stories.

The stories have been delivered online, with no editing or correction by the BBC or OUP, so are complete with grammar, spellings and punctuation as provided. There are 48,000 hits for *friend* – and 311 for *freind.*

The data has all been lemmatised and part-of-speech-tagged, and then loaded into the Sketch Engine (Kilgarriff et al 2004).

3 Analyses

We have looked at contrasts with writing for children (OCC-R) and variation by age, gender and region.

Contrast with writing *for* children: We looked at the 200 keywords of Beebox in contrast to the 9 million words of 21st century fiction written for children that we had within OCC-R. These were examined by one of the authors and classified. [2] At the most general level, the classification was between writing problems, and themes. The writing problems included uncapitalised names, missing apostrophes (*cant, wont),* hyphens (*hearted, haired, headed)* and inter-word spaces (*anymore, aswell, infront*) as well as spellings (*whent, thay, solder* for *soldier,*

[1] http://www.bbc.co.uk/radio2/500words/2012/

[2] The keyword lists included only lowercase lemmas of at least three characters, with a simplemaths parameter of 100: for details of the statistic and method see Kilgarriff (2012).

minuet for *minute*, *cheater* for *cheetah*).

More interesting were the themes that children wrote about notably more than adults writing for children:

- **Scary stories**
  - *creepy creaky croaky dreaded foggy ghost gloomy graveyard haunted mansion misty mysterious petrify scared scary spooky undead vampire zombie*
- **Traditional**
  - *pixie elf genie goblin leprechaun gnome*
  - *prank potion robber*
- **People**
  - *mum mummy mom dad daddy auntie grandpa grandma*
- **Space/war**
  - *alien asteroid astronaut galaxy portal rocket spaceship teleporter*
  - *ammo ninja sniper spaceship teleportal*
  - *airport*
- **Animals**
  - *cheetah dolphin hippo kitten ladybird panda penguin squirrel zebra*
  - *unicorn*
  - *bunny teddy*
  - *woof meow tweet* (what birds do)
  - *vet zoo*
- **Food**
  - *candy cupcake coke marshmallow*
- **Jewels**
  - *diamond emerald gem locket necklace*
- **Other nouns**
  - *clown diary bully snowman*
  - *gymnastics karate sleepover medal*
  - *foster orphanage*

These (but for the scary ones) were largely nouns. There were also:

- **Adjectives**
  - *adorable adventurous bouncy comfy fluffy ginormous horrific horrifying humongous magical sparkly stormy super wrinkly yummy*
- **Adverbs**
  - *extremely happily luckily speedily unfortunately worriedly*
- **Verbs**
  - *cuddle investigate sprint stroll stutter unpack wake*
- **Other:**
  - *(*ding*) dong phew*
  - *bye okay soo*

Gender: The gender analysis is somewhat painful.

Girls in contrast to boys:

- **Romance**
  - *blush boyfriend cheek cuddle darling hug kiss snuggle sweetheart sweetie wedding xxx*
- **Horses**
  - *canter chestnut groom mane neigh pony riding stable unicorn*
- **Nature**
  - *butterfly cherry daisy flower kitten lilac lily petal poppy rainbow rose*
- **Dance**
  - *ballet chorus dance*
- **Adjectives**
  - *adorable beautiful cute dainty delicate flowery fluffy glittery gorgeous hazel pink silky sparkly rosy*
- **Traditional**
  - *diary fairy locket maid pixie mermaid*
- **Hard stuff**
  - *cancer comfort cope fault foster upset*
- **Textures/clothes**
  - *cardigan stroke (v) velvet ribbon silk silky skirt*
- **People**
  - *daddy daughter lady princess sibling sister twin*
- **Food**
  - *candyfloss bun(1)*
- **Hair and beauty**
  - *bun(2) glossy wavy blonde curly plait makeup necklace*
- **Pronouns**
  - *her hers herself she*
- **Other**
  - *doll giggle girl girlie pink soo sparkle sprinkle teddy skip sleepover shyly*

Boys in contrast to girls:

- **Fighting**
  - *aim ambush ammo armed armor armored army arrow assassin assassinate assault attack base battle blast bullet bunker cannon captain chopper civilian cockpit combat commander defend defender defense destroy device engine explosion explosive*

fighter fireball fuel general grenade guard gun gunfire helicopter helmet himself knight laser launch launcher leader league machine military missile mission nuclear opponent parachute patrol pilot pistol radar rifle robot scout sergeant shield shoot shot shotgun smash sniper soldier spear survivor sword system tank target teleport temple terrorist troop warrior weapon

- **Sport**
  - *football footballer goal goalkeeper penalty player ref referee score squad stadium striker tackle team training*
- **Other**
  - *himself galaxy teleport zombi*e

Age: We divided the authors into three bands: up to eight, nines and tens, and 11+, and found the keywords of each age group in contrast to the remainder.

Up to 8:

- **Fairy stories**
  - *once upon magic end happily castle fairy adventure magical king princess spell wand queen palace*
- **Other adjectives**
  - *naughty sunny sad sparkly lovely excited shiny friendly*
- **Food**
  - *cake party chocolate eat yummy tea*
- **Pirates**
  - *pirate cave dragon treasure*
- **Other**
  - *dinosaur swim play pet lot*

9 and 10:

- **Reporting verbs**
  - *mumble moan yell stammer shout agree exclaim sneak boom*
- **-ly adverbs**
  - *suddenly excitedly sadly loudly angrily extremely luckily*
- **Scary adjectives**
  - *dusty gloomy spooky*
- **Other adjectives**
  - *gigantic exciting famous ugly colossal annoying enormous cute bore sunny super lovely brilliant*
- **Nouns**
  - *alien cage robot rope potion lightning ginger breakfast adventure mansion lady mum hamster sword ship portal*
- **Other**
  - *meanwhile later bye once hello yes zoom*

11+:

- **Body parts**
  - *blood body cheek eye face fear hand heart shoulder throat*
- **Body/mind functions**
  - *breath feeling memory mind pain smile sweat tear thought*
- **Abstract nouns**
  - *darkness death echo force life murder silence soul word*
- **Atmospherics**
  - *alone cold dead pale silent slowly wind*
- C**onnectives**
  - *against almost since though within yet*
- **Verbs**
  - *die feel glance lie fill seem sense stand stare*
- **Romance**
  - *figure woman*
- **Pronouns**
  - *myself nothing*

The steps from childhood towards adolescence are vividly shown. The 11+ keywords (deeply indebted, we suspect, to the Twilight novels) scarcely need commentary, so loud do they sing of teenage concerns. The two pronouns which have made it into the list – *myself, nothing* – sum up all by themselves the agony of being a teenager.

Less obvious, and more intriguing, are the clusters of reporting verbs and –ly adverbs that the nines and tens use, and the adjectives, in the younger two age groups, switching to connecting words amongst the 11+s. They may relate to the National Curriculum, and story-telling techniques that children are taught at particular stages.

Region: The top keyword for Birmingham-and-the-Black-Country is *mom.* The top keyword for Tyne-and-Wear is *mam.* Children tend to write as they speak, and in the northeast the usual short name for a mother rhymes with 'Sam' and around Birmingham it rhymes with 'Tom'. For the rest of us it rhymes with 'plum'. The corpus is closer to a spoken data collection than most written corpora would be.

At a level of themes, the keywords for Norfolk have seven animals in the top twelve; the top three keywords for Wales are *sheep, bus, dragon;* for Scotland, *beside, wee, gran.*

Our explorations in this area are very

preliminary, but we suspect the corpus offers a great deal to dialectologists.

References

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell. D. 2004. "The Sketch Engine". *Proceedings of Euralex*, Lorient, France.

Kilgarriff, A. 2012. *Getting to know your corpus.* in: *Proc. Text, Speech, Dialogue (TSD 2012)*, Lecture Notes in Computer Science. Sojka, P., Horak, A., Kopecek, I., Pala, K. (eds). Springer.

Wild, K., Kilgarriff, A., Tugwell, D. 2011. "Oxford Children's Corpus: A corpus of writing for children". Poster at ICLIC (*International Corpus Linguistics Conference*), Birmingham, UK.

Wild, K., Kilgarriff, A., Tugwell, D. 2012. "The Oxford Children's Corpus: Using a Children's Corpus in Lexicography". *International Journal of Lexicography*, doi: 10.1093/ijl/ecs017. Oxford UniversityPress.

LinguisticsWeb.org: a web for learning and teaching corpus linguistic tools and methods

Sabine Bartsch
Technische Universität Darmstadt
bartsch@linglit.tu-darmstadt.de

1 Introduction

This abstract introduces linguisticsweb.org [1], a wiki-based website providing students and researchers with tutorials, how-tos, links, tools, corpus access and other types of information to help them learn corpus and computational linguistic methods, techniques and technologies for linguistic text analysis.

2 Learning corpus linguistics

Learning corpus and computational linguistics and other disciplines within the digital humanities requires students with different computational expertise to familiarize themselves with an array of methods, techniques and technologies and to stay abreast of new developments. The skills that have to be mastered range from basic data preparation such as preparing textual data from hybrid sources, via tokenization and segmentation to appropriately querying and analysing corpora by means of creating word frequency lists, concordances and suitable filters and views of the data.

Annotation tasks such as part of speech tagging and lemmatisation are indispensable prerequisites for many research tasks. However, many automatic annotation tools offer a vast array of options and possibilities, yet often lack the convenient graphical user interfaces that computer users today are familiar with and that make available different options at a click of the mouse. And even where tools are seemingly geared towards the less computer savvy user, the initial learning curve is often relatively steep, especially when the user is facing the decision of whether a given tool is the right choice for the task at hand.

These issues are aggravated when tasks become more complex such that different tools have to be used in combination, i.e. when processing pipelines of multiple tools are necessary in order to achieve ones goals.

Additionally, many advanced tools offered by the community – at least for a while, but often indefinitely – remain in the state of research

[1] http://www.linguisticsweb.org

prototypes and are thus often rather unwieldy and make it difficult even for the more experienced user to harness their full processing power. This is even more so in cases where documentation is sparse and not geared towards students.

Another obstacle facing students of corpus and computational linguistics is that many tools are presented detached from potential research scenarios such that it is often difficult to decide for the student whether a tool has the desired functionalities, whether it is suitable for the task at hand and how it is to be used in an authentic research setting. It is no secret that learners benefit from well-prepared examples, but those are often lacking in websites and documentation aimed at an expert community.

3 Issues

The issues arising in the process of learning and teaching methods and techniques in corpus and computational linguistics are thus found to be at least three-fold:

- besides having to learn the theory and methodology of their field of studies, learners have to select and employ the relevant methods and tools for a given research task,
- information about tools and methods is often not presented in a user-friendly and integrated way, such that especially less experienced users are struggling to identify and make use of relevant information and apply it to their research task, and
- information on methods and techniques is often geared towards expert users and rarely prepared with didactic aims in mind.

4 Linguisticsweb.org: motivations and goals

In order to help overcome these issues, we have set up linguisticsweb.org, a website offering different types of information for students of corpus and computational linguistics. Besides offering one-stop access to many widely used tools and how-tos explaining how they are used, the website offers example processing scenarios and sample analyses and guides the learner through their application. It also links in with other relevant information sources such as a glossary of linguistic terminology, online resources such as dictionaries, papers and other relevant websites, and further external information sources.

The goals motivating linguisticsweb.org are based on hands-on experience concerning the needs of students. Tutorials, how-tos, links, tools, corpus access and other types of information are prepared to help students master methods, techniques and technologies for corpus-based linguistic text processing and analysis. Its aim is to serve as a source of information for students of linguistics at different skill levels corpus linguistic methodology and with different levels of computational expertise.

Linguisticsweb.org is maintained for students and with students. A team of students is continuously helping to expand the knowledge base and ensure that the information offered is up-to-date and suitable for other students. Feedback is elicited from the users. The aim of the website is to alleviate the frustrations often entailed in hands-on data processing and to encourage students to use new tools on their own. It is open and accessible to the community and constantly being expanded and, we hope, improved.

References

Barstch, S. 2013. Linguisticsweb.org. URL: http://www.linguisticsweb.org; last accessed on: 15 January 2013.

TILCE – the Turin Italian Learner Corpus of English

Luisa Bozzo
University of Turin, Italy
luisa.bozzo@unito.it

TILCE – the Turin Italian Learner Corpus of English – is a corpus of academic written and computer-mediated English extracted from the output of Italian students of English participating in an experimental online workshop in English linguistics. The workshop runs parallel to one of the English Language courses for MA students of foreign languages held at the University of Torino, and is managed on the Moodle[1] platform of the university. The corpus is being compiled within the research project *English in Italy* coordinated by Professor Virginia Pulcini and financed by the Fondazione San Paolo, Italy; it is expected to reach the size of half a million tokens by its planned date of publication in 2014.

The design of TILCE is provided together with the illustration of its principles informing the collection of text data. Unlike "peripheral" learner corpora (Nesselhauf 2004:128, quoted in Gilquin 2012:5), whose texts tend to be elicited and collected for their own sake irrespective of the students' learning context, the texts in TILCE are purposeful on their own since they all belong to a complete learning programme which includes forum discussions, analysis essays, research tasks and reports, and peer feedback (Bozzo 2012a, 2012b, 2012c, 2013). The Moodle learning management system is the environment where the constructionist task-based activities take place, thus greatly facilitating the communicative and experiential processes of learners. Students participate in the workshop on a voluntary basis; their native language is Italian and their average level according to CEFR[2] is C1. Besides the needlessness to persuade students to engage in the project, the advantageous features of TILCE, as compared to most learner corpora, are manifold: motivated communication, authenticity of interaction, naturalness of discoursal functions, controlled semantic fields, focus on learners' needs in terms of language learning, textual variety, range of learners' errors unlimited by task narrowness.

The mode of collection of the texts is eased by the workshop's medium, since these are all produced electronically and stored online; therefore transcription is expected to present only minor difficulties.

The corpus is meant to be made available online on a Creative Commons[3] Attribution-NonCommercial-NoDerivs 3.0 Unported License. TILCE is planned to be released in three main formats: raw, POS-tagged, and annotated for lexical errors. The raw .txt format files will be accompanied by information about students (each identified with a code), year of attendance (2011, 2012, etc.), text-types (forum discussion, written paper, peer feedback), task-types (analysis, report, webquest etc.), so that they may be organized into subcorpora; normalization of frequently-used emoticons will be necessary. The POS-tagged XML format is meant to be used with concordancing tools like the SketchEngine[4]; the POS-tagging process is expected to be relatively error-free since the morphology and the syntax of the students' interlanguage are quite appropriate; however, a degree of manual control and correction will be performed. The annotated format will be based on an *ad hoc* cluster-tagging system to identify lexical errors and to classify them according to typology and most probable causes.

The purpose of the corpus is to provide a tool for investigation on Italian upper-intermediate and advanced students' written and computer-mediated interlanguage to linguists, language educators, materials designers, lexicographers and the selfsame students. In its annotated format, TILCE's aim is to help researchers retrieve and classify some of the most frequent lexical errors, thus developing better understanding of their typology and heightened awareness of their causes. Ultimately, TILCE may offer a model of data elicitation, collection and treatment for comparable corpora of the same kind.

References

Bozzo, L. 2012a. "Il blended learning all'Università: sperimentazione di un paradigma di apprendimento esperienziale costruttivista", in T. Roselli, A. Andronico, F. Berni, P. Di Bitonto, V. Rossano (eds.), *DIDAMATICA 2012 – Informatica per la Didattica. Taranto, 14-16 maggio*, *Mondo Digitale*, Anno XI, n.2, Giugno 2012. Available online at http://mondodigitale.aicanet.net/ultimo/index.xml

Bozzo, L. 2012b. "Student-Driven Moodle Courseware Design for Advanced English Language Teaching". In G. Fiorentino (ed.), *Atti del MoodleMoot Italia 2012*. Available online at

[1] moodle.org
[2] *Common European Framework of References for Languages*, available online at http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf
[3] creativecommons.org
[4] www.sketchengine.co.uk

http://www.moodlemoot.it/mod/data/view.php?d=24&advanced=0&paging=&page=0

Bozzo, L. 2012c. "Collaborative Construction of Glossaries in Language Learning: an integration of the lexical approach and constructionism through blended learning". *ICT for Language Learning Conference Proceedings*, Florence, 15-16 November 2012. Available online at http://www.pixel-online.net/ICT4LL2012/index.php

Bozzo, L. 2013. "Developing advanced language learners' autonomy in blended learning". In Menegale, M., Coonan, C.M., (eds.), *Autonomy in Language Learning: Getting learners actively involved*. Canterbury: IATEFL.

Gilquin, G., De Cock, S., Granger, S. 2010. *Louvain International Database of Spoken English Interlanguage. CD-Rom and Handbook*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Nesselhauf, N. 2004. "Learner corpora and their potential in language teaching". In Sinclair, J. (ed), *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamins, 125-152.

Uncovering second language learners' miscollocations using SketchEngine

Howard Hao-Jan Chen

National Taiwan Normal University

hjchen@ntnu.edu.tw

Collocation forms an important aspect of vocabulary learning. There have been more and more studies showing that collocation played a significant role in developing learners' mental lexicon, indicating that they influence learners' success of language acquisition (Ellis, 1996; Lewis, 2000, Lien, 2003). Nevertheless, collocations were found to be problematic for many second language learners. Many studies have consistently revealed that EFL learners had insufficient knowledge of English collocations (Channel, 1981; Bahns & Eldaw, 1993; Gitsaki, 1997; Liu,1999, Chen, 2008). Some researcher also found that lexical miscollocations were the most common errors made by EFL learners (Newman, 1988; Bahns & Eldaw, 1993). While verb-noun collocations "form the communicative core of utterances where the most important information is placed" (Altenberg, 1993), researchers in indicated that the most frequently occurred miscollocations were verb-noun miscolloations, and they were particularly difficult for second language learners (Wu, 1996; Wang, 2001; Liu, 2002; Li, 2005).

Empirical studies have offered clear evidence that EFL learners lack verb-noun collocational knowledge and made suggestions for explicit instructions on collocations. According to Woodlard (2000) and Lewis (2000), helping students observe and notice their own miscollocations would enhance students' awareness of acceptable collocations. Therefore, investigating learners' miscollocations can help teachers and researcher better understand learners' general pattern of collocational errors, thus shedding lights on what collocations to teach and how to help language learners.

In most existing studies on miscollocations, researchers in these studies collected L2 learners' compositions, assignments, or examination essays, and manually extracted the miscollocations. Nevertheless, even though these small numbers of manually-extracted miscollcoations provided some evidence for students' collocational errors, the miscollocations found were fairly restricted due to the limited data.

As learner corpora become more widely available, there have been several investigations into learners' miscollocations through learner corpora (Nesselhauf, 2003, 2005; Shih, 2000; Liu, 2002; Chang & Yang, 2009). One of the most comprehensive studies is the verb-noun collocation study carried out by Nesselhauf (2005). She investigated the use of verb-noun collocations produced by advanced German learners of English based on International Corpus of Learner English (ICLE). Nesselhauf manually extracted and analyzed the verb-noun combinations in the 318 essays selected from the sub-corpus (GeCLEE) which contained around 150,000 words. The results indicated that 2082 lexical verb-noun combinations were indentified, and 507 VN miscollocations were found.

It is evident that Nesselhauf spent great amount of time in searching through the learner corpus and provided useful information about learners' miscollocations. For other researchers, it seems to be a daunting task to engage in this type of miscollocation analysis. It is very labor-intensive and time-consuming to examine second/foreign language learners' miscollocations. Facing such a challenging task, researchers around the world might need a more robust corpus research tool to uncover miscollocations more efficiently.

In this paper, we will introduce a useful corpus research tool called Sketch–Diff, a tool included in SketchEngine (SKE) developed by Adam Kilgarriff and his associates. This powerful tool can help language researchers to compare various collocations used in a native corpus and a learner corpus. In our experimental study, we uploaded the following three corpora onto SKE: one native reference corpus – British Academic Written English – (6.5 million words) and two EFL learner corpora, a Taiwanese college learners' written corpus (2.2 million words) and a Chinese college learners' written corpus (3.8 million words). After we uploaded these three corpora into SKE, The SKE can automatically tag and analyze these corpora. After the pre-processing stage, the Sketch-Diff tool can then be used to compare all the collocations used by native speakers and non-native speakers. When the user input any target word, he/she can then find all collates of the target word used in the uploaded native corpus and non-native corpus. Various collocation patterns can be displayed and the significant differences between the usage of native speakers and non-native speakers can be revealed. The system also used different colors to show the users about the significant differences between native and nonnative usage. For instance, the following collocation errors are quickly identified, *study knowledge (acquire/gain knowledge); do *sport (do exercise).

Based on our empirical tests, the Sketch-Diff can be a very effective tool in comparing collocations used in English native and non-native corpora. With this new comparison tool, researchers no longer need to painstakingly examine each and every lines of students' essay and manually mark each and every collocation errors. This new tool provides a more convenient and thorough way of uncovering the possible differences between native and non-native's collocational competence. At the current stage, the corpus tool is surely not fully automatic in uncovering all the collocation errors made by L2 learners. Researchers still need to further verify the differences recommended by SKE. However, the tool has already made the contrastive interlanguage analysis more manageable and feasible.

References

Altenberg, Bengt (1993). Recurrent verb-complement construction in the London-Lund Corpus. In Jan Aarts, Pieter de Haan, & NellekeOostdijk (Eds.), English language corpora: Design, analysis, and exploration. Paper from the thirteenth International Conference in English Language Research on Computerized Corpora, Nijmegen 1992 (pp. 227-245). Amsterdam: Rodopi.

Bahns, J.,& Eldaw, M. (1993). Should we teach EFL students collocations? System, 21, 101-114.

Channell, J. (1981). Applying semantic theory into vocabulary teaching. ELT Journal, 35(1), 115-122.

Chen, M. H. (2008). A study on the English collocation competence of college students in Taiwan. Unpublished master's thesis, National Taiwan University of Science Technology, Taiwan, R.O.C.

Chen, P.C. (2002). A corpus-based study of the collocational errors in the writings of the EFL learners in Taiwan. Unpublished master's thesis, Natinal Taiwan Normal University, Taiwan, R.O.C.

Lewis, M. (1997). Implementing the lexical approach. London: Language Teaching Publications.

Li, C.C. (2005). A study of collocational error types in ESL/EFL college learners' writing. Unpublished master's thesis, Ming Chuan University, Taiwan, R.O.C.

Li, C. & Thompson, S.A. (2005). Mandarin Chinese: A functional reference grammar. Berkeley: University of California Press.

Lien, H. Y. (2003). The effects if collocations instruction on the reading comprehension of Taiwanese college students. Doctorial Dissertation, Indiana University of Pennsylvania.

Liu, C. P. (1999). A study of Chinese Culture University freshman's collocational competence: "Knowledge" as an example. Hwa Kang Journal of English language & Literature, 5, 81-99.

Liu, C. P. (1999b). An analysis of collocational errors in EFL writings. The proceedings of the Eighth International Symposium on English Teaching. Taipei: Crane

Liu, L. E. (2002). A corpus-based semantic investigation of verb-noun miscollocations in Taiwan learners' English. Unpublished master's thesis, Tamkang University, Taiwan, R.O.C.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. Applied Linguistics, 24, 223-242.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

Wang, C. J. (2001) A study of the English collocational competence of English majors in Taiwan. Unpublished master's thesis, Fu Jen Catholic University, Taipei.

Woolard, G. (2000). Collocation: Encouraging learner independence. In M. Lewis (Ed.), Teaching collocation: Further developments in the lexical approach (pp.28-46). London: Language Teaching Publication.

Wu, W. S. (1996). Lexical collocations: One way to make passive vocabulary active. Paper from the eleventh conference on English teaching and learning I nthe Republic of China (pp. 461-480). Taipei: Crane.

A Verbal Autopsy corpus annotated with cause of death

Samuel Danso, Eric Atwell, Owen Johnson
University of Leeds
scsod@leeds.ac.uk

1 Introduction

We report on a corpus of Verbal Autopsy documents, each annotated by expert clinicians with the Cause of Death. Verbal Autopsy is recommended by the World Health Organisation as a pragmatic substitute for a clinical autopsy to establish cause of death in regions where death may occur well away from clinical services (Soleman *et al*, 2006). Verbal Autopsy involves interviewing individuals (such as relatives or caregivers) who were close to the deceased, and if possible, who cared for the individual around the time of death, to document events that may have led to the individuals' death. The content of this corpus is derived from two questionnaires for conducting Verbal Autopsy interviews to obtain cause of death information, the first for stillbirths and infants (less than 12 months of age) and the second for women of reproductive age.

2 Verbal Autopsy

Of the estimated 57 million deaths per year worldwide, 67 per cent are not medically certified with cause of death due to weak or negligible death registration systems ((Soleman *et al*, 2006).. During the Year 2000 Summit, the United Nations set out the Millennium Development Goals for improving world health and these included a determined focus on reducing child mortality and maternal deaths(Sachs and McArthur 2005). To achieve this goal, there is a crucial need for cause of death information from hospitals as well as deaths that occur outside hospitals. Meanwhile, globally, over 40 countries have employed Verbal Autopsy as a means to ascertain the likely cause of death(Fottrell *et al*. 2007). Cause of death information is vitally important for developing health interventions and disease treatment research. Analysis of Verbal Autopsy data may be invaluable in revealing preventable illness, for example a locally high death rate due to neonatal infection may help regional health managers respond with strategies such as targeted education on infection control. This information may help to

inform national and international health managers, policy makers and researchers about trends in causes of death in order to develop strategies, design interventions, and carry out sound budgetary allocations.

3 Problems associated with current automatic analysis of Verbal Autopsy

Typically information gathered using Verbal Autopsies is captured on paper using standard questionnaires which are then passed to physicians who review them to determine the most likely cause of death. The standard practice worldwide has been the use of a minimum of two physicians to give two independent assessments of each Verbal Autopsy, even though there is some evidence to suggest that one physician may be enough for this process (Danso et al, 2011). The use of physicians for this is characterised by several limitations: high cost; intra-physician reliability; repeatability; and time (Byass *et al*. 2010). The cost of manual review and assignment of cause of death to Verbal Autopsy documents has not been formally evaluated. The problem is compounded where there is a shortage of medical personnel and generally this is the case in places where Verbal Autopsy is used.

Consequently, there is a growing interest in research in the use of computational approaches to classify causes of death, to address the limitations associated with time consuming and expensive physician reviews (Byass *et al*. 2010). Verbal Autopsy questionnaires contain both structured data and open history narrative. Our literature review found that the computational approaches published so far have only made use of the structured data available while physicians have access to and make use of both the structured information and the open history narrative (Danso et al, 2013).

4 Source of the corpus

This corpus is obtained from two large field trials carried out in Ghana, which led to the establishment of a Verbal Autopsy surveillance system which ran between December 2000 and July 2010. The surveillance system covered 7 contiguous, predominantly rural, districts within the Brong-Ahafo region of Ghana. The objective of the ObaapaVitA trial was to assess the effect of weekly low-dose vitamin A supplementation in women of reproductive age in Ghana on pregnancy-related mortality, female mortality more generally, and peri-natal and infant mortality Kirkwood *et al*. (2010a) . The objective of the Newhints trial was to develop a feasible and sustainable community-based approach to improve newborn care practices in order to improve the survival of newborns. Data was collected during four-weekly surveillance, which included recording all stillbirths, deaths in infants up to one year, and women of reproductive age using Verbal Autopsies conducted by field supervisors Kirkwood *et al*. (2010b) .

The corpus contains real Verbal Autopsy text as obtained from the interview and transcribed onto the Verbal Autopsy form. The sample contained all stillbirths and deaths in infants to the age of 12 months, which is referred to in this paper as the infant subcorpus. Additionally, it also contains text about the causes of all deaths in adult women between the age of 15 and 45, which is referred to in this paper as the women subcorpus. The corpus contains a total of approximately 2.5 million words in 11,741 documents (Danso et al, 2013).

References

Byass, P, K Kahn, E Fottrell, MA Collinson and SM Tollman. 2010. Moving from data on deaths to public health policy in Agincourt, South Africa: Approaches to analysing and understanding Verbal Autopsy findings. *PLoS Medicine* 7(8).

Danso, S, E Atwell, O Johnson, G ten Asbroek, Karen Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, Z Hill, J Fenty, S Amenga-Etego, S Owusu-Agyei and B R Kirkwood. 2011. Verbal Autopsy corpus for machine learning of cause of death. *Proceedings of the Corpus Linguistics Conference*. Birmingham, UK.

Danso, S; Atwell, ES; Johnson, O; ten Asbroek, A; Soromekun, S; Edmond, K; Hurt, C; Hurt, L; Zandoh, C; Tawiah, C; Fenty, J; Etego, S; Agyei, S; Kirkwood, B. 2013. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. *ICAME Journal* 37.

Fottrell, E, P Byass, TW Ouedraogo, C Tamini, A Gbangou, I Sombie, U Hogberg, KH Witten, S Bhattacharya, T Desta, S Deganus, J Tornui, AE Fitzmaurice, N Meda and WGraham. 2007. Revealing the burden of maternal mortality: a probabilistic model for determining pregnancy-related causes of death from verbal autopsies. *Population Health Metrics* 5: 1

Kirkwood, B R, L Hurt, S Amenga-Etego, C Tawiah, C Zandoh, S Danso, C Hurt, K Edmond, Z Hill, G ten Asbroek, J Fenty, S Owusu-Agyei, O Campbell and P Arthur. 2010a. Effect of vitamin A supplementation in women of reproductive age on maternal survival in Ghana (ObaapaVitA): a cluster-randomised, placebo-controlled trial. *Lancet*. 375: 1640-1649.

Kirkwood, BR, A Manu, C Tawiah-Agyemang, G ten

Asbroek, T Gyan, B Weobong, ER Lewandowski, S Soremekun, S Danso, C Pitt, Hanson, S Owusu-Agyei and Z Hill. 2010b. NEWHINTS cluster randomised trial to evaluate the impact on neonatal mortality in rural Ghana of routine home visits to provide a package of essential newborn care interventions in the third trimester of pregnancy and the first week of life: trial protocol. *Trials* 11: 58.

Sachs, JD and JW Mcarthur. 2005. The millennium project: a plan for meeting the millennium development goals. *Lancet* 365: 347-353.

Soleman, N, D Chandramohan and K Shibuya. 2006. Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization* 84: 239-245.

Representation of female body shape and size in newspaper discourse: A corpus-based study

Lisa Da Silva
Edge Hill University
lisa.dasilva@go.edgehill.ac.uk

1 Introduction

Research concerning the imbalance in gender representation and misrepresentation in the media has evolved greatly since Tuchmann (1978:169) first discussed "the symbolic annihilation of women by the mass media". Taking a corpus-based approach, this research reports on the extent to which newspaper discursive practices support gendered ideologies regarding body size. The research also comments on the ways that discoursal practices enable this construction.

2 Background

A number of key points become apparent when examining the literature in this area. Certain body sizes are constructed as unacceptable in a variety of ways; for example, by associating them with either negative behaviour and personality characteristics, or (specifically in the case of sizes considered too large) a resulting lack of success either personally or professionally (Ferris 2003; Wykes & Gunter 2005; White & Kurz 2008). Media discourses (in particular the newspaper discourse of the *Daily Mail* and *Daily Express*) are contradictory in nature as to what body size is acceptable (Wykes & Gunter 2005; Gill 2006). Finally, it is clear that the images and discourse around body size in the media can have a detrimental psychological effect on readers and viewers (Guendouzi 2004; Bessenoff 2006) and thus the topic merits further research.

3 Methodology

In order to explore to what extent and how gendered discoursal practices are constructed and maintained the study employed a methodology combining Corpus Linguistics and Critical Discourse Analysis approaches (e.g. Baker et al. 2008; Mautner 2005).

The study used the British National Corpus to focus on a) determining which words describing body size have a high frequency and b) whether a discrepancy exists in the frequency that these words are used to refer to either men or women. The words used in the search query were

identified in an earlier pilot study as frequently occurring in discourse regarding body size (Da Silva 2012).

The corpus data was collected using a search query (restricted to the newspaper section of the BNC) comprising words within the semantic field of body size as shown in Table1. Concordance lines were analysed manually in order to establish any emerging patterns of the words in relation to gender. Instances where occurrences did not explicitly refer to either men or women are listed in the final column ("Other"). Closer qualitative analysis of the occurrences noted in the analysis aimed to answer the research questions using the approach of Critical Discourse Analysis (Wodak & Meyer 2009).

4 Results

Table 1 shows the results of the quantitative analysis.

Word	Total	Women	Men	Other
fat	318	37	25	256
thin	218	6	8	204
slim	135	17	23	95
overweight	50	8	19	23
slender	33	0	0	33
plump	23	4	3	16
chubby	19	0	6	13
anorexia	16	11	0	5
trim	14	2	3	9
voluptuous	14	5	1	8
flabby	13	0	0	13
obesity	13	2	2	9
flab	11	5	2	4
obese	9	1	4	5
skinny	9	3	3	3
tubby	7	3	3	1
anorexic	5	3	0	2
curvy	4	3	0	1
curvaceous	3	2	0	1

Table 1. Words describing body size: frequency by gender

The quantitative analysis, combined with the manual examination of concordance lines, revealed the following points.

Words referring to both shape and size (such as *curvy*, *curvaceous* and *voluptuous*) were used more frequently to describe to women.

Chubby, which was used with an overall positive meaning, did not occur in reference to women, but only in reference to men, children or babies. Contrary, *skinny* tends to be used with negative meaning, in that it is aligned to ill-health. *Flabby* was only used to describe specific body parts, and frequently patterned with *thighs*, *bum*, *bottom* and *backside*.

The use of figurative language was a frequent feature of the discourse representation. For example, flab only appeared as the subject of a metaphorical fight, battle, or war.

5 Conclusions

Overall, there was little difference in the frequency of references to body size for men (102) and women (112). However, women's body shapes and sizes were presented in a negative light more frequently than men's. Men were more frequently described as *thin*, *slim*, *overweight* and *chubby*. In contrast, women were more frequently described as *fat*, and descriptions of body weight using *anorexia* and *anorexic* only occurred in reference to women.

Finally, body size and body shape tended to presented as areas of conflict, producing either winners or losers.

References

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press". *Discourse & Society*. 19 (3): 273-306.

Bessenoff, G., R. 2006. "Can the media affect us? Social comparison, self-discrepancy, and the thin ideal". *Psychology of Women Quarterly*. 30: 239-251.

Da Silva, L. 2012. Representations of female body size in newspaper discourse. Unpublished dissertation (PGCert. in Research). Edge Hill University.

Ferris, J. 2003. "'Parallel Discourses and "Appropriate" Bodies: media constructions of Anorexia and Obesity in the cases of Tracey Gold and Carnie Wilson". *Journal of Communication Enquiry.* 27 (3): 256-273.

Gill, R. 2006. *Gender and the Media*. Cambridge: Polity Press.

Guendouzi, J. 2004. "'She's very slim': talking about body-size in all-female interactions". *Journal of Pragmatics.* 36 (9): 1635-1653.

Mautner, G. 2005. "Time to get wired: Using web based corpora in critical discourse analysis". *Discourse & Society* 16 (6): 809-828.

Tuchman, G. 1982. "The symbolic annihilation of women by the mass media". In S. Cohen and J. Young (eds.) *The manufacture of news: social problems deviance and the mass media.* London:

Constable.

Whitehead, K. and Kurz, T. 2008. “Saints, sinners and standards of femininity: discursive constructions of anorexia nervosa and obesity in women's magazines” *Journal of Gender Studies* 17 (4): 345-358.

Wodak, R. and Meyer, M. 2009. “Critical Discourse Analysis: History, Agenda, Theory and Methodology”. In: R. Wodak and M. Meyer (eds.) *Methods of Critical Discourse Analysis.* 2nd ed. London: Sage.

Wykes, M. and Gunter, B. 2005. *The Media & Body Image*. London: Sage Publications.

Collocational priming of idiomatic expressions: norms and exploitations

Natalya Dubois Marysheva
University of South Brittany
natalya.dubois@univ-ubs.fr

1 Introduction

Intertextuality has been much discussed in terms of what is available on the surface. Semantic prosody (Louw 1993, 2000) and lexical priming (Hoey 2005) open the way into more subconscious aspects of intertextuality. Collocational resonance (Williams 2008) has already shown how subtle patterns of meaning variation can be shown over time, the aim here is to carry this research forward through the analysis of idioms of literary origin. Although fixedness is the key property of fixed expressions and idioms (FEI), Moon (1998) points out that almost 40 per cent of FEI have lexical variations. Exploitations of idioms allow speakers/authors to benefit from the stylistic manipulation of lexis and semantics of FEI, providing some sort of familiarization, and typically providing humour. Moon argues that exploitations ultimate only work with full perception of both vehicles and tenors in the metaphors, and a vestige of lexical form, which correlates with Hoey's lexical priming theory. In this study, I am mostly interested in wordplay and in what Melčuk terms “artistic deformations”.

In order to illustrate this work, we take the much used expression *Fools rush in where angels fear to tread*, which is in fact a line from Alexander Pope’s poem *An Essay on Criticism*, written in 1709, although few speakers of English are likely to be aware of this. This expression is simply one of the many hackneyed phrases of the English idiomatic pantheon that is much used, and much exploited in journalism and elsewhere.

2 Corpus exploration

The corpora we use for the analysis are the British National Corpus and the web-as-corpus EnTenTen8 [1] . To create the lexicographical prototype (Hanks 1994, 2000) needed to track variations of meaning of the nouns constituting the expression, we begin with the definitions of the Oxford English Dictionary. According to this definition, “*an angel*” is a spiritual being, divine messenger, superior to men in power and

[1] both of which are available at https://the.sketchengine.co.uk/

intelligence, a rebellious spirit (the fallen angel), a guardian, a member of a celestial hierarchy, a person of exemplary conduct and virtue.

To go further, Corpus Pattern Analysis and WordSketches were used to create and explore the lexico-syntagmatic prototypes and the collocational networks of the nouns "*angel*" and "*fool*". The first stage is to analyse the nature of angels and of fools. To begin, with the word *angels*, the question is what these celestial beings can actually do, other than *rush in*. We next need to know what can be done with or to an angel, what angels are like and what they are associated with.

Once we have prototypical definitions of angelness, we can isolate a few frequent FEI, namely: *fallen angel, avenging angel, be on the side of the angels, angels in the house*, and then move on to a similar analysis of "*fools"*.

3 Exploiting "angels"

The exploitations found in the EnTenTen8 reveal that "*angels"* is frequently substituted in this expression by words conveying the idea of *people of exemplary behaviour, more intelligent than the average*, without taking into account *the rebellious spirit* sense:

> **Politicians *rushed in*** where **philosophers** have feared to tread.
>
> **William Derham *rushed in*** where **Huygens** dared not tread.
>
> So **many incompetent ones *rush in*** where **those much wiser** fear to tread.
>
> **Lots of journalists *have rushed in*** where **more sensible people** dared not tread.

Another method we applied in our study is to analyse the verbs constituting the expression through the FrameNet[1] and the PDEV (Pattern Dictionary of English Verbs)[2]. Both *"rush"* and "*tread"* belong to the frame "self-motion", and "*fear*" belongs to the frame "*fear*". According to the entry for "rush in" of the PDEV based on Corpus Pattern Analysis, we observe that the notion of "great haste" is essential to the meaning of "rush in":

No.	*Pattern / Implicature*	
2	pv **[[Human]] rush [NO OBJ] {in | into [[Location]]}**	conc.
	[[Human]] enters [[Location]] in great haste	exploit.

Figure 1. PDEV entry for "*rush in*"

The analysis of corpus data provided by EnTenTen8 shows that only "*go swiftly"*, "*jump in"* and "*pounce"*, that replace "*rush in*" in the exploitations, convey the idea of haste. The others are about courage and audacity: '*venture', 'dare to go', 'baldly go', 'go trespassing*', being a contrast to "*fear*".

This brief analysis shows how collocational networks and lexicographical prototypes provide a deeper understanding of exploitations of idiomatic expressions. Much more work will ensue.

References

Hanks, P. "Corpus pattern analysis". *Euralex Proceedings*. Vol. 1. 2004. 87–98.

Hanks, P. "Linguistic norms and pragmatic exploitations or, why lexicographers need prototype theory, and vice versa". *Papers in Computational Lexicography: COMPLEX* 94 (1994): 89–113.

Hoey, M. *Lexical priming: A new theory of words and language.* Routledge, 2005.

Louw, B. "Contextual prosodic theory: Bringing semantic prosodies to life". *Words in Context. A tribute to John Sinclair on his Retirement. Birmingham: University of Birmingham, Cédérom* (2000)

Louw, B. "Irony in the text or insincerity in the writer". *Text and technology: In honour of John Sinclair* (1993): 157–176.

Moon, R. *Fixed expressions and idioms in English: A corpus-based approach*. Clarendon Press Oxford, 1998.

Williams, G. "Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles". *International Journal of Corpus Linguistics* 3.1 (1998): 151–171.

Williams, G. C. "Verbs of Science and the Learner's Dictionary". *Proceedings of the 13th EURALEX International Congress. Barcelona, Universitat Pompeu Fabra.* 2008. 797–806.

[1] https://framenet.icsi.berkeley.edu
[2] http://deb.fi.muni.cz/pdev/action=patterns&id=rush

Query logs as a corpus

Ann-Marie Eklund
Univ. of Gothenburg
ann-marie.eklund@gu.se

Dimitrios Kokkinakis
Univ. of Gothenburg
dimitrios.kokkinakis@gu.se

1 Introduction

This paper provides a detailed description of a large Swedish health-related query log corpus and explores means to derive useful statistics, their distributions and analytics from its content across several dimensions. Information acquisition from query logs can be useful for several purposes and potential types of users, such as terminologists, infodemiologists / epidemiologists, medical data and web analysts, specialists in NLP technologies such as information retrieval and text mining but also public officials in health and safety organizations.

Analysis of web interaction logs can provide useful information regarding the use of a site considering when and how users seek information for topics covered by the site; Bar-Ilan *et al.* (2009). Such information can be used both for a general understanding of public health awareness and information seeking patterns, and to optimize search indexing, query completion and presentation of results for improved public health information. For an overview of some common applications of log analysis and the methods of analyzing them see Oliner et al. (2011). To our knowledge there is no similar resource for Swedish. Therefore, we intend to provide a detailed, in-depth description of the content and features of this new large corpus, which is an important step towards understanding the breadth and depth of usage patterns, the properties of the resource and the challenges involved in working with such type of data.

The use of web logs for search engine optimization is part of the research often called Search Analytics. This research also covers analyses to reveal trends and search patterns to address topics of interest to, for instance, governments and public information providers. Therefore, deeper mining into queries can reveal more important information about search engine users and their language use and also to reveal new information from the search requests; *cf.* Medelyan, 2004. The basis for Search Analytics is made of different kinds of logs of search terms and presented and chosen results by web site users. According to Mat-Hassan & Levene (2005) the objectives of a web log analysis are to:

- investigate a searcher's performance
- establish the profile of an effective searcher
- establish a user's searching characteristics and
- understand a user's navigational behaviour, including the number of search terms entered and the number of click-throughs viewed.

In the context of health-related information, search analytics has been used to study, in addition to traditional search optimization, how search behaviour relates to disease outbreaks Hulth *et al.*, 2009. At a syntactic level queries may contain e.g. synonyms and hyponyms, and to be able to study patterns of search behaviour at a more abstract level, we map the syntactic terms to semantic concepts.

2 The Log Corpus and its annotation

The corpus consists of the interaction logs, i.e. search queries and clicked links, for the period October 2010 to September 2012. The data is provided by the Swedish health web site *Vårdguiden*, the Stockholm Health Care Guide <http://www.vardguiden.se/> (via Euroling AB). The site is the official health care portal of the Stockholm County and is also sponsored by the Stockholm County Council.

The total number of queries is 67 million, where 27/2,2 million are unique (before and after case normalization). The corpus has been automatically annotated with two medical semantic resources, a named entity recognizer as well as part-of-speech information. The semantic resources are the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and the Medical Subject Headings (MeSH), a thesaurus, which is organized as a hierarchy from general categories, such as *Diseases* and *Organisms*. The named entity annotation includes ontological categories such as *Person*, *Location* and *Organization*.

The poster will provide a detailed description of various corpus statistics and the analysis we will provide will enable to gain new insights into the language used in the logs, particularly the terminology and general vocabulary and even, to a certain degree, provide an indication of the search strategies applied by the users of the web site service from where the logs are obtained. Our findings can serve as background work that can be incorporated in search engines or other web-based applications to personalize search results, provide specific site recommendations and suggest more

precise search terms, for example by the automatic identification of laymen/novices or domain experts; (*cf.* Ryen White *et al.,* 2008). In the long run, we are also interested to aid our understanding of user's search behavior and categorizing their information need(s) through the use of vocabulary patterns and interpret the knowledge that exists therein, therefore visual analytics techniques can be an important and effective mechanism for achieving these goals.

Acknowledgements

We are thankful to Adam Blomberg**,** CTO**,** Euroling AB for providing the log data. We are also thankful for the support by the Centre for Language Technology (http://clt.gu.se).

References

Judith Bar-Ilan, J., Zheng Zhu and Mark Leven. 2009. Topic-specific analysis of search queries. Proceedings of the 2009 workshop on Web Search Click Data (WSCD). Pp. 35–42. ACM,NY, USA http://doi.acm.org/10.1145/1507509.1507515

Anette Hulth, Gustaf Rydevik and Annika Linde. 2009. Web queries as a source for syndromic surveillance. PLoS One 4(2).

Mazlita Mat-Hassan and Mark Levene. 2005. Associating search and navigation behavior through log analysis: Research articles. J. Am. Soc. Inf. Sci. Technol. 56, 913–934 (July 2005), http://dl.acm.org/citation.cfm?id=1067990.1067995

Olena Medelyan. 2004. Why Not Use Query Logs As Corpora? Proceedings of the Ninth ESSLLI Student Session. Paul Egré and Laura Alonso i Alemany (eds). Nancy, France.

Adam Oliner, Archana Ganapathi, Wei Xu. 2011. Logs contain a wealth of information for help in managing systems. Queue vol. 9, no. 12. Issue: Advances and Challenges in Log Analysis. ACM.

Ryen W. White, Susan Dumais and Jaime Teevan. 2008. How Medical Expertise Influences Web Search Interaction. Proceedings of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA.

Reading multimodality: a report of an investigation into the multimodality of data representation in a corpus of medical articles

Mel Evans
University of Birmingham
m.evans@bham.ac.uk

Caroline Tagg
University of Birmingham
c.tagg@bham.ac.uk

The significance of multimodality in human communication is now recognised across linguistic disciplines as diverse as stylistics, second language acquisition, and sociolinguistics. Within English corpus linguistics, the shift towards multimodal analysis is evident in recent developments in spoken corpora such as the CHILDES corpus and the University of Nottingham's multimodal corpus (see Knight et al. 2008). However, despite the prominence of the 'multimodal turn', many corpus studies of written language continue to rely primarily on monomodal text, due in part to the practical difficulties involved in incorporating other modes. How can corpus-driven, written language research successfully account for a range of key meaning-making resources, such as typography, layout and space, colour, texture, image, and embedded video and audio? Our poster reports on the early findings of a project which addresses key theoretical and methodological concerns surrounding multimodality in corpus linguistics.

The project, entitled 'The likely impact of Elsevier's new interface on reader engagement with representations of data in medical research articles', investigates the different modes and forms of data representation in a corpus of academic medical journal articles, in order to establish the relationship between these modes, authorial stance and reader processing. Our study primarily addresses the following research question: 'How is Elsevier's new online interface impacting on the ways in which readers process and interpret data representations (figures, tables and other images) in research articles?' The project is part of an ongoing collaborative research initiative between the Centre for Corpus Research at the University of Birmingham and Elsevier, one of the world's largest academic journal publishers. Elsevier's extensive back-catalogue provides an exciting opportunity to conduct large-scale synchronic and diachronic research into academic discourse. The outcomes of our investigation are anticipated to, firstly, raise

awareness of reader engagement and practices with multimodal discourse, specifically within the medical academic community; secondly, offer new insights for Elsevier into the effect of their new interactive online interface and its impact on reader engagement with data representation in medical research articles; and thirdly, advance theories and methods used in corpus analysis, particularly regarding multimodal data.

The project comprises a number of stages or phases, which include the initial identification and description of the representation of data in the medical journal articles; interviews with authors and editors to establish the impact of editorial policy; eye-tracking experiments with readers to identify processing norms for the traditional and innovative article interfaces; and finally the automation of the process of initial data analysis and the development of a fully-annotated multimodal corpus. In our poster, we report the findings from the first stage of the investigation. It will document: (1) the key properties of the corpus; (2) the rationale and manual methods used to annotate and process the multimodal elements; (3) the discussion of preliminary results collected on the linguistic and other modal features relating to data representation.

References

Knight, D., Adolphs, S., Tennent, P. and Carter, R. 2008. "The Nottingham Multi-Modal Corpus: A Demonstration". *LREC* 2. Available online at http://www.lrec-conf.org/proceedings/lrec2008/pdf/13_paper.pdf

The difference between English and English: Examining varieties on the basis of register

Jennifer Fest
RWTH Aachen University
jennifer.fest@rwth-aachen.de

English, as the most widespread language of today's world, is spoken in many different places and on many different levels and occasions, and the sheer number of usages makes it almost impossible to clearly identify individual varieties, let alone define their status. Studies so far have mainly focused on descriptive analyses based on either geographical and political factors or grammatical and lexical diversity (e.g. Trudgill and Hannah 1982; Jenkins 2003; Kirkpatrick 2010) in order to examine the role English plays in a given society or country. These conditions alone, however, do not necessarily allow a deep insight into the stage of the variety's development. While English can be an official language, this does not mean that it is also used in many other, everyday contexts; moreover, grammatical or lexical differences in comparison to the standards of British or American English very often originate from the contact to other languages spoken in the area and do not solely depict a limitation of the variety.

In any given language, different fields of application will result in variations according to use (Halliday and Hasan 1989; Matthiessen 1993), and on the basis of this assumption it can be said that the diversity of registers that are applied in a speech community might be an important indicator for the status of a variety (Mukherjee and Schilk 2012). If a variety is put to ever more use and occurs in ever more environments and fields in a community, it can safely be presumed that most members of the society feel comfortable with it or are at least able to understand it and identify certain differences in registers.

Following this, the study at hand aims at analysing two varieties of English, namely Kenyan and Hong Kong English, in comparison to the native standards of Great Britain, the USA and Australia. Since it would not be possible to cover all existing registers in a community, the corpus is restricted to the language found in newspapers, a field which addresses a large and diverse audience and at the same time covers numerous topics. The register dimensions of tenor and field of discourse (Halliday and Hasan 1977) will therefore be the main focus of the analysis and will be combined with theoretical frameworks

from the field of media studies as well as recent findings on the impact and opinion-shaping effects of mass media (Lasswell 1948; Schenk 2007).

In order to create a representative picture of the genre of newspaper writing and its different registers, the corpus for this study consists of 800 articles per variety, including 160 each for the areas of economy, features, spot news, sports and politics, and thus adds up to 4,000 articles in total. In the progress of this work, the data will be tagged for parts of speech to allow for more detailed and sophisticated queries, yet preliminary analyses with the untagged corpus have already rendered first glimpses at the differences between the varieties and their usage and diversity with regard to registers and have indicated that a variance of different registers in this field might not only give an impression about the variety as a whole, but also show in which areas of life or for which target groups in particular the usage of English is most common and important.

References

Halliday, M.A.K. and Hasan, R. 1977. *Cohesion in English*. 2nd ed. London: Longman.

Halliday, M.A.K. and Hasan, R. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Jenkins, J. 2003. *World Englishes*. London: Routledge.

Kirkpatrick, A. 2010. *The Routledge Handbook of World Englishes*. London: Routledge.

Lasswell, H. D. 1948. "The Structure and Function of Communication". In L. Bryson (ed.) *The Communication of Ideas. A Series of Addresses*. New York: Harper & Brs.

Matthiessen, C. 1993. "Register in the Round: Diversity in a Unified Theory of Register Analysis". In M. Ghadessy (ed.) *Register Analysis. Theory and Practice*. London: Pinter Publishers.

Mukherjee, J. and Schilk, M. 2012. "Exploring Variation and Change in New Englishes: Looking into the International Corpus of English (ICE) and Beyond." In T. Nevalainen and E. Traugott (eds.) *The Oxford Handbook of the History of English*. Oxford: Oxford University Press.

Schenk, M. 2007. *Medienwirkungsforschung*. 3rd ed. Tübingen: Mohr Siebeck.

Trudgill, P. and Hannah, J. 1994. *International English*. 3rd ed. London: Arnold.

A comparison of metaphors of love across three music genres, based on the lyrics of the top charting albums of 2011 in the UK

Stephanie Furness-Barr
University of Portsmouth
sfurness80@hotmail.com

'Love', the focus of this on-going research, is apparently 'all around' (Werner, 2012, p. 44, pre-publication version), being a topic that occurs very frequently in mainstream music. It is also an idea that is frequently expressed in metaphor. Metaphors of love (Lakoff and Johnson, 1980; Deignan, 1997; Tissari, 2001; and Kövecses, 2010, amongst others), and music lyrics have been extensively studied. There has likewise been corpus research of lyrics, including that of Murphey (1992), Schnieder and Meithaner (2006), Kreyer and Mukherjee (2009), Kreyer (2012) and Werner (2012). I have yet to discover, however, research that has employed corpus techniques to discover how love and its metaphorical expressions are portrayed across music genres.

This is an exploratory investigation of metaphorical expressions of love in the lyrics of three genres of music. Given that Country, R&B, and Rock and Metal are musically distinct, and tend to appeal to different audiences, it is reasonable to wonder if the lyrics in these genres characterise love, a common theme in popular music, any differently. Thus the aim of this research was to address the following questions:

- What metaphors of love can be found in the three genres?
- How are metaphors of love similar and different across the genres?

To find out, I composed three small corpora, consisting of the lyrics from up to 27 albums per genre, each of which reached the top, or near the top, of their respective charts in 2011 (See Table 1), according to 'The Official Charts Company'. The data from this UK based organisation derives from "real sales to British consumers, of audio and video releases across a wide range of retailers" (Official Charts Company, n.d), which is to say all significant retailers. Data includes downloads but excludes streaming. Note that the dominance of certain albums in the number one chart position in RnB necessitated the inclusion of albums that reached the number two spot in this corpus, while similarly, the dominance of certain albums in the number one and two chart positions

in Country meant that those albums reaching the third spot be included in this corpus.

CORPUS	Chart Positions	No. of Albums	No. of Songs	Ave. No. of Songs/Album
Rock & Metal	1 only	25	337	13.48
R&B	1 and 2	25	418	16.72
Country	1, 2 and 3	27	373	13.81
Whole Corpus		**77**	**1128**	**14.65**

Table 1: Chart, Album and Song Data

Table 2 provides the primary data for the three corpora. Note that R&B has a significantly higher number of tokens. This can be accounted for in part by the fact that a) one of the albums was a compilation of three CDs of 18 songs each, b) there were a greater number of songs per album (See Table 1 above); c) songs were, in general, longer than in the other two genres, with a fair degree of repetition.

CORPUS	Tokens used for word list	Types (distinct words)	Type/Token Ratio (TTR)
Rock & Metal	79,290	4,709	5.94
R&B	204,042	11,691	5.73
Country	90,248	5,289	5.86
Whole Corpus	**373,580**	**14,738**	**3.95**

Table 2: Corpus Data

Once compiled, using Sketch Engine and Wordsmith 6.0, the subsequent exploration of the corpora focussed on

1. discovering the number of instances of LOVE, its use syntactically, and common collocates, such as BE IN LOVE, MAKE LOVE, FALL IN LOVE;
2. identifying more 'creative' metaphorical expressions of LOVE, by:
    - searching for structures typically used in metaphor, eg. LOVE IS.., LOVE LIKE…, LOVE HAS…;
    - using techniques outlined by Baker (2006) to extract metaphor from corpora; and
    - examining collocates to the right and left of the word LOVE
3. establishing those metaphors within a typology
4. comparing and contrasting use of these expressions across the three genres.

Initial analysis has shown similarity and variation across the three corpora. In terms of frequency, as Table 3 shows, all three genres refer to love. R&B has a significantly higher number of instances of LOVE overall, at 56.5% of the total, which is not surprising given the much greater size of that corpus. This is followed by Country at 31.7% and Rock & Heavy Metal at 11.8%. However, there is a higher frequency of instances per number of tokens in Country, at 0.95%, than R&B (0.75%) and Rock & Heavy Metal (0.4%).

LOVE	Frequency	Tokens	% of Corpus	% LOVE
Rock & Metal	317	79,290	0.40	11.8
R&B	1,521	204,042	0.75	56.5
Country	855	90,248	0.95	31.7
Whole Corpus	2,693	373,580	0.72	100.0

Table 3: Instances of LOVE

In terms of metaphor 'types', under the provisionally established categories, each genre includes expressions in which, for instance,

LOVE IS…

- AN INANIMATE OBJECT
- AN ANIMATE / ORGANIC OBJECT
- (A/AN)(UNPREDICTABLE) FORCE

but that the perspective of the 'object' or 'force' is sometimes substantially different. Observations include the fact that:

- R&B seems to favour the metaphor LOVE AS AN INTOXICANT
- Rock & Metal does not seem to have a dominant love metaphor, but the highest number of expressions occur under: LOVE AS AN INANIMATE OBJECT and LOVE IS (A/AN UNPREDICTABLE) FORCE
- With some exceptions, the vast majority of love metaphors used in Country music seem to give quite a positive or hopeful view of love.

These observations represent only a partial review of study findings. A more comprehensive summary and discussion of results is forthcoming. It is not as yet my intention to discuss the implications of the results of this study. There is a body of research, however, in which the intersection of music and lyrics is discussed from psychological perspectives. Articles from, for example, Sellnow and Sellnow (2001) and Greitemeyer (2011) offer insight into this area.

There are a number of limitations to this research, including the fact that as official lyrics were not easily accessible, those used in the corpora are not wholly reliable representations. The lyrics were of necessity copied from lyrics websites to which the general public contributed. Moreover, generalisations cannot be made with such small corpora, and based only one year of top albums. The results can thus only be

indicative, but could point the way to lead to further research.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Deignan, A. (1997). Metaphors of Desire. In K. Harvey & C. Shalom (Eds.) *Language and Desire: Encoding sex, romance and intimacy* (pp. 21-42). London: Routledge.

Greitemeyer, T. (2011) Exposure to music with prosocial lyrics reduces aggression: First evidence and test of the underlying mechanism. *Journal of Experimental Social Psychology* 47, pp. 28–36. DOI:10.1016/j.jesp.2010.08.005

Kövecses, Z. (2010). *Metaphor* (2nd ed). Oxford: Oxford University Press.

Kreyer, R. (2012). Love is like a stove – it burns you when it's hot: A corpus-linguistic view on the (non-) creative use of love-related metaphors in pop songs. In S. Hoffmann, P. Rayson and G. Leech (Eds.). *English Corpus Linguistics: Looking back, Moving forward: Papers from the* 30th International Conference on English Language Research on Computerized Corpora (ICAME 30) (pp. 103-115). Amsterdam: Rodopi.

Kreyer, R. & Mukherjee, J. (2009) The Style Of Pop Song Lyrics: A Corpus-Linguistic Pilot Study. *Anglia – Zeitschrift für englische Philologie*, 125(1), pp. 31-58. DOI: 10.1515/ANGL.2007.31

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago.

Murphey, T. (1992). The Discourse of Pop Songs. *TESOL Quarterly*, 26(4) pp. 770-774. Retrieved from http://www.jstor.org/stable/3586887 .

The Official Charts website (n.d). Retrieved from: http://www.officialcharts.com/

Schnieder, E. W. & Miethaner, U. (2006). When I *started to using* BLUR : Accounting for Unusual Verb Complementation Patterns in an Electronic Corpus of Earlier African American English. *Journal of English Linguistics*, 34, pp. 233-256. DOI: 10.1177/0075424206293381

Sellnow, D. & Sellnow, T. (2001). The "Illusion of Life" Rhetorical Perspective: An Integrated Approach to the Study of Music as Communication. *Critical Studies in Media Communication*, 18(4), pp. 395–415

Tissari, H. (2001) Metaphors We Love By: On the Cognitive Metaphors of *Love* from the 15th Century to the Present. *Studia Anglica Posnaniensa*, 36, pp. 217-242.

Werner, V. (2012). Love is all around: a corpus-based study of pop lyrics. *Corpora*, 7(1), pp. 19–50. DOI: 10.3366/corp.2012.0016 (pre-publication version)

An alternative perspective to the analysis of recurrent phraseology: lexical bundles and phrase frames in the language of hotel websites

Miguel Fuster
IULMA-Universitat de València
miguel.fuster@uv.es

1 Introduction

For this paper we have analysed a corpus of British hotel websites in an attempt to find a suitable perspective that may explain the role of recurrent multiword units (MWUs) which are representative of this genre. We have focussed on the determination of common 4-word lexical bundles. This aspect has been sorted out very rapidly by means of a concordancer. However, we have been able to observe that next to lexical bundles, the close phraseological pattern of phrase frames with a similar number of words should enter the picture. It is our view that lexical bundles and phrase frames are options which complement each other quite well and help determine more faithfully the main phraseological types which hospitality website writers have at their disposal.

2 The language of tourism

To start with, the identity of the language of tourism as a genre in itself, or preferably as multiple genres has been discussed extensively in the literature. Calvi (2010) refers to it as a macro-genre (see also Sanmartín-Sáez 2012; Suau-Jiménez 2012). On the other hand, within the language of tourism, the kind of discourse which is exhibited by hospitality websites represents the lowest level of specialization (see further Vargas Sierra 2008), since these sites have been designed by service providers to reach final customers themselves, thus bypassing intermediaries. In the marketing business, hotel websites are conceived as powerful marketing service tools which contain informational, transactional and relational features (see Ab Hamid et al. 2010). To be truly useful, the features of these websites need to be constantly revised and, more importantly, updated.

3 Hospitality websites and corpus data

The reasons provided above explain why the language data obtained from hotel websites does not remain immutable. Later visits clearly show changes of various sorts and, in that sense, these

sources differ from more traditional sources. Thus, any corpus based on these sources is almost automatically outdated and cannot be verified (or falsified) by later scholarship. Despite such changes, on the whole, it may be argued that the sites reflect linguistic features which are fairly uniform or representative of the internet genre.

4 The vocabulary of hotel websites

In general, the vocabulary in most of the sections of these webpages could be described as non-specific. Nevertheless, the literature mentions some outstanding quantitative and qualitative features of this genre. For instance, there has been research on the role played by nouns and adjectives in tourism English (see Mapelli 2008, Pierini 2009, Manca 2010; Ning and Yu 2011; Edo-Marzá 2012). For instance, self-mention through nouns or noun phrases which specify the hotel name is a common feaures feature, no doubt used to promote the hotel, its staff or service. Likewise, there is an extensive use of lexical verbs and adjectives with crucial functions which should be examined from the perspective of the whole word sequence.

5 Phraseological features in the language of tourism

Various other discursive aspects of its language have been examined. However, in spite of abundant work on the subject, little attention has been devoted to its recurrent phraseology. Scholars have noticed that both academic and specific genres exhibit a number of frequent fixed expressions which help to characterize or define their discourse. Particularly, there has been interesting work on the use of lexical bundles, or n-grams, in academic and specialized contexts (see Biber, et al. 1999; Biber, Conrad and Cortes 2004; Chen and Baker 2010; Forchini & Murphy 2010; Hyland 2008). These word combinations of words are fixed albeit not necessarily idiomatic. Most research on lexical bundles to date has focused on combinations of three to five members, where both form and (discursive) function are discussed. A closely related type of word combination is that of phrase or p-frames (see Fletcher 2003-2006), where at least one of the elements of the lexical phrase is variable (not fixed) (see overview of the types in Granger and Meunier 2008).

6 Some conclusions about the phraseology of hospitality websites

In our view, both lexical bundles (fixed expressions) and phrase frames (partially fixed expressions) should be treated together on an equal footing as phraseological constituents of this hospitality website genre. Both types, in fact, appear to merge by forming lexico-syntactic expressions where clear functional affinities emerge. I hope to show through the quantitative and qualitative examination of our corpus of British hospitality websites, that the role played by nouns, verbs and adjectives can be more fully understood when they are examined as constituents of these fixed and semi fixed multiword expressions.

References

Ab Hamid, NR, Md Akhir, R. and Mashudi, PM. 2010. "An Assessment of the Internet's Potential in Enhancing Customer Relations". *International Journal of Arts and Sciences* 3(12): 265-281.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

Biber, D., Conrad, S. and Cortes, V. 2004. "*If you look at* ... Lexical Bundles in University Teaching and Textbooks". *Applied Linguistics* 25 (3): 371-405.

Calvi, M.V. 2010. "Los géneros discursivos en la lengua del turismo: una propuesta de clasificación". *Ibérica* 19: 9-32.

Chen, Y.H. and Baker, P. 2010. "Lexical Bundles in L1 and L2 Academic writing". *Language Learning & Technology* 14 (2): 30-49.

Edo-Marzá, N. 2012. "Páginas web privadas e institucionales: el uso de la adjetivación en un corpus inglés-español de promoción de destinos turísticos". In Sanmartín-Sáez, J. (ed) (2012b), pp. 51-80.

Fletcher, William 2003–2006. "Exploring Words and Phrases from the British National Corpus". Available online at http://pie.usna.edu.

Forchini, P. & Murphy, A. 2010. "N-grams in comparable specialized corpora". In Römer, Ute & Rainer Schulze (eds.) *Patterns, Meaningful Units and Specialized Discourses*. Amsterdam, NLD: John Benjamins Publishing Company, 87-103.

Granger, S. and Meunier, F. 2008. "Disentangling the phraseological web". In Granger & Paquot (eds.) 2008 *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins, 27-49.

Hyland, K. 2008. "As can be seen: Lexical bundles and disciplinary variation". *English for Specific Purposes* 27: 4–21.

Manca, E. 2010. "From phraseology to culture: Qualifying adjectives in the language of tourism". In Römer, U. and Schulze, R. (eds.) *Patterns, Meaningful Units and Specialized Discourses*. Amsterdam, NLD: John Benjamins Publishing Company, 105-122.

Mapelli, G. 2008. "Las marcas de metadiscurso interpersonal de la sección 'turismo' de los sitios web de los ayuntamientos". In M. V. Calvi, M.V., Mapelli, G. and Santos López, J. (eds.). *Lingue, culture, economia: comunicazione e pratiche discorsive*, Milano, FrancoAngeli, 173-190.

Sanmartín-Sáez, J. 2012a. "De las normativas turísticas a las páginas electrónicas de promoción de hoteles: la clasificación hotelera desde la perspectiva lingüística". In Sanmartín-Sáez, J. (ed) (2012b), 81-124.

Sanmartín-Sáez, J. (ed.) 2012b. *Discurso turístico e Internet*, Madrid: Iberoamericana-Vervuert.

Suau-Jiménez, F. (2012). "Páginas web institucionales de promoción turística: el uso metadiscursivo interpersonal en inglés y español". In Sanmartín-Sáez, J. (ed) 2012b., 125-154.

Vargas-Sierra, C. 2008. "La sistematización terminográfica: una propuesta metodológica para la elaboración de diccionarios traductológicos". Available online at http://193.145.233.67/dspace/bitstream/10045/13212/1/1453%20Vargas.pdf.

Towards a multilingual specialised corpus for business translators

Daniel Gallego-Hernández
University of Alicante
daniel.gallego@ua.es

Francisco José García-Rico
Secret Escapes
fgarciarico@gmail.com

Ramesh Krishnamurthy
Aston University
r.krishnamurthy@aston.ac.uk

Paola Masseau
University of Alicante
paola.masseau@ua.es

Miguel Tolosa-Igualada
University of Alicante
miguel.tolosa@ua.es

The aim of this poster is to present the COMENEGO project (Corpus Multilingüe de Economía y Negocios) (Multlingual Corpus of Business and Economics). The main objective of this research project is to design a virtual platform that allows the exploitation of a multilingual specialised corpus (economics and business). This corpus may be conceived as an on-line linguistic tool designed especially for translators so that these professionals and others such as translator trainers and trainees can reduce or eliminate the time involved in the compilation of ad hoc specialised corpora. The project involves four main stages which are not sequential:

1) designing the virtual platform: this stage entails the analysis of users' needs and implementation of tools according to them (translators, translator trainers, researchers, etc.) and also in keeping with the results of other stages; in this way the platform should allow not only users and usage rights management, but also corpus exploitation techniques such as concordances, filters, etc.).

2) obtaining copyright permissions: the pilot COMENEGO corpus is currently designed as an in-house tool for teaching (exercises for terminology, translation, revision of translations, etc.) and research (extraction and management of terminology and specialized phraseology). This stage involves applying for permission from the different organizations from which the pilot corpus texts were retrieved so that the corpus can be distributed through the platform and provide open access.

3) integrating texts representative of the professional practice of business translation into

the virtual platform: this stage requires carrying out a survey of professional translators and organizations in order to complement and confirm (or reject) the selection of texts in the current pilot corpus; the survey investigated not only current practices in the translation of texts, but also, among other things, the perceived role of corpora in business translation.

4) carrying out discourse analysis of the textual resources: this stage aims to analyze the pilot corpus in order to reveal imbalances and deficiencies which should be addressed, and also to confirm or reject the classification of the corpus texts (which were initially classified in categories established by pragmatic and subjective parameters) so that it can be implemented in the platform).

This poster will briefly present the results of project so far, and discuss future research that could be carried out.

References

Gallego Hernández, Daniel (2012): "Comenego (Corpus Multilingüe de Economía y Negocios): anatomía de un proyecto", *Seminario sobre traducción e interpretación económica e institucional: docencia, investigación y profesión*, Universidad de Alicante.

Gallego Hernández, Daniel (2012): "Proyecto COMENEGO: algo más que un corpus multilingüe de economía y negocios", *X Jornadas de Redes de Investigación en Docencia Universitaria*, Universidad de Alicante.

Gallego Hernández, Daniel and García Rico, Francisco José and Masseau, Paola and Miguel Tolosa Igualada (2012): "COMENEGO (Corpus Multilingüe de Economía y Negocios): hacia la alimentación de una plataforma virtual para traductores", *IV Congreso Internacional de Lingüística de Corpus CILC2012*, Universidad de Jaén.

Krishnamurthy, Ramesh and Daniel Gallego Hernández (2012): "Discursive analysis of textual resources of COMENEGO", *IV Congreso Internacional de Lingüística de Corpus CILC2012*, Universidad de Jaén.

Gallego Hernández, Daniel and Ramesh Krishnamurthy (2011a): "COMENEGO (Corpus Multilingüe de Economía y Negocios): corpus estable vs. metodologías ad hoc (web as/for corpus) aplicadas a la práctica de la traducción económica, comercial y financiera", In Carrió Pastor, M. L. and Candel Mora, M. A. (eds.): *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III Congreso Internacional de Lingüística de Corpus*, Valencia: Universitat Politècnica de València.

Gallego Hernández, Daniel & Ramesh Krishnamurthy (2011b): "COMENEGO (Corpus Multilingüe de Economía y Negocios): design, creation and applications", *Corpus Linguistics 2011: Discourse and Corpus Linguistics*, University of Birmingham.

Gallego Hernández, Daniel and Ramesh Krishnamurthy (2010): "Rates of Exchange: first report on COMENEGO (Corpus Multilingüe de Economía y Negocios)", *III Jornadas Internacionales de Fraseología Contrastiva: Fraseología, Opacidad y Traducción*, Universidad de Alicante.

Tracing salience in the Prague Dependency Treebank

Eva Hajičová, Barbora Hladká, Jan Václ
Charles University
{hajicova, hladka}@ufal.mff.cuni.cz
janvacl@centrum.cz

1 Motivation

Most of the natural language processing systems require context to be taken into account to get adequate results. In general, context is information directly present in a document processed, i.e. knowledge of words and the relations among them, as well as information given by a broader context of situation.

We believe that salience concept of the Functional Generative Description (FGD; Sgall et al., 1986) could be one of the theories which shed light on this broad area of phenomena as well as we consider its hypotheses to be a solid basis for empirical testing.

We work with the notion of the degrees of salience (activation) of the items in the stock of shared knowledge together with the representation of the dynamic development of the discourse by means of changes of these degrees.

2 Salience of the items

The knowledge-based *salience* algorithm was designed to capture a dynamic character of the stock of knowledge assumed by the speaker to be shared by her/him and the hearer(s): not only the repertoire of items it includes is changed but also their activation (salience) (Hajičová and Vrbová 1982), (Hajičová 1993), (Hajičová and Hladká and Kučová 2006).

3 Data

The linguistic data we have used for our experiment are those of the Prague Discourse Treebank 1.0[1] being the extension of the Prague Dependency Treebank 2.5.[2]

We read out as much information as possible from the sentence underlying structure represented in the form of a dependency tree capturing underlying syntactic structure, the information structure of the sentence, i.e. its topic-focus articulation, coreference and bridging relations. In total, we work with 3,165 documents consisting of 49, 431 sentences.

[1] http://ufal.mff.cuni.cz/discourse/
[2] http://ufal.mff.cuni.cz/pdt2.5

4 Salience graphs and their interpretation

When one is to interpret the numerical data, salience degrees here, visualization of them can help a lot. We visualize the development of salience degrees during a discourse using *salience plots*.

The salience plots indicate the ways in which a dynamic account of discourse structure may be applied. First, a certain segmentation of the texts analyzed is displayed, viewing it either vertically or horizontally: one can imagine that vertical lines can be drawn between those parts of discourse in which certain items keep a higher degree of activation and do not 'fade away' too far. On the other side, horizontal lines can be imagined to indicate certain thresholds for the possibility of a weaker (pronominal) referential expression to be used, or the necessity for a stronger reference by a noun or a more descriptive noun group. Also the topic of a segment of the discourse can be determined on the basis of the groupings of items on the top of the schema for the given segment.

We find out an evidence for these hypotheses in both the Prague Discourse Treebank data and the salience plots.

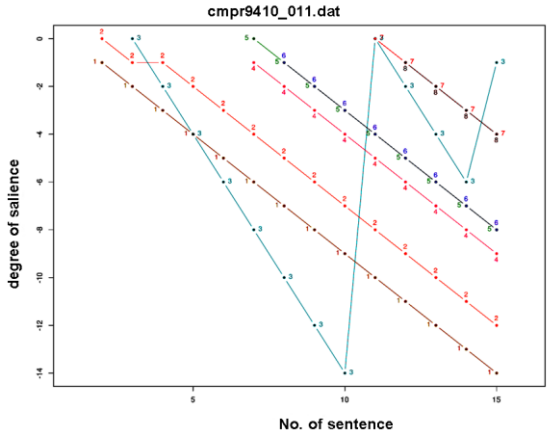


Figure 1

References

Hajičová, E. (1993): *Issues of sentence structure and discourse patterns*. Karolinum-Charles University Press, Prague, Czech Republic.

Hajičová, E., Hladká, B., Kučová L.(2006): An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In: *Proceedings of the Workshop on Constraints in Discourse*, Copyright © National University of Ireland, Maynooth, Ireland, pp. 82-89.

Hajičová, E., Vrbová, J. (1982): On the role of the hierarchy of activation in the process of natural language understanding. *Proceedings of the COLING'82*, 107−113.

Sgall, P., Hajičová, E., Panevová, J (1986): *The*

meaning of the sentence in its semantic and pragmatic aspects. Dordrecht:Reidel / Prague:Academia.

An initial approach on medical term formation in Japanese through the usage of corpora

Carlos Herrero-Zorita
Autonomous University of Madrid
car.herrero@estudiante.uam.es

1 Introduction

When speaking about Natural Language Processing (NLP), the Japanese language proves to be undoubtedly challenging. One of the primary problems is the distinction between word-formation using affixation, or by compounding. Although Japanese has bound morphemes, sometimes the distinction from free morphemes is not so clear, as kanji ideograms represent concepts or ideas that have full meaning on their own, which can be problematic for some NLP tasks. The objective of this research is to propose an initial exploration on how Japanese words are formed in medical terminology, through the usage of corpora and the Sketch Engine [1] (Kilgarriff et al. 2004) as query tool. Since compounding is extremely productive in Japanese, our hypothesis is that medical specialised terms would be formed by agglutinated strings of free morphemes, instead of an affixation process such as in English and Spanish.

2 Antecedents

Japanese compounding has been widely studied (Kageyama 1982; Shibatani 1990), and it has been a primary focus for NLP researchers in specialised texts (Hisamatsu and Nitta 1996; Han et al. 2005; Kudo 2007; among others). However, the area of medical terminology has not been so widely covered. Our research proposes an initial description of the formation of medical terms in Japanese using real-life texts, which may become useful for further research on the NLP area. We have chosen the Sketch Engine as it has proven to be a very reliable tool when using Japanese corpora (Srdanović and Nishina 2008).

3 Methodology

We have used as empirical evidence the Japanese corpus from the MultiMedica project at the Computational Linguistics Laboratory of the Autónoma University of Madrid. The corpus was

[1] http://www.sketchengine.co.uk/

compiled and tagged in the Laboratory, lead by Dr. Antonio Moreno Sandoval (Moreno-Sandoval and Campillos-Llanos, 2013), and it consists of texts from several specialised magazines: *Kampo Medicine* (Japanese oriental medicine), *Kansenshogaku Zasshi* (infectious diseases magazine), *Kanzo* (magazine about diseases of the liver), *ORLTokyo* (Japanese otorhinolaryngology) and *Sanfujinka no shinpo* (advances on obstetrics magazine). The corpus is formed by a total number of 3,746 documents and 1,131,304 tokens. It was then uploaded to Sketch Engine, which tagged it with the morphological analyser ChaSen[1].

The starting point was a list of English medical affixes collected from different sources and re-elaborated by us, consisting on 467 Greco-Latin prefixes and affixes. Each of them was translated, using an online Japanese-English medical dictionary[2] that allowed the user to search for the beginning and the ending of the words. After a search is made (e.g. *cervic-*), the dictionary provides a list with all the possible terms in English (*cervicalgia*, *cervicitis*, *cervicobrachial*, etc.). Each of these terms is introduced in the general Japanese medical search tool of the same dictionary, and the kanji string(s) that correspond to the affix are obtained (e.g. 頚 'neck' and 頚部 'neck region').

After we acquired the string(s) of kanji, they were introduced in Sketch Engine's Word Sketch entry form to check if they appeared in the corpus. At this point we encountered one main problem: the oversegmentation issue (Hisamatsu & Nitta, 1996). Since words in Japanese are not separated by blank spaces as in other languages, the POS taggers have problems when deciding when does a word begin and end, especially in an agglutinative language such as Japanese. In this case, medical terms formed by more than two kanji that do not appear in common dictionaries are split by ChaSen into smaller recognisable morphemes. For example, if 鼓室 'tympanic cavity' (equivalent to the prefix *tympano–*) is searched using Sketch Engine, it will not appear in the corpus, as ChaSen has divided it into 鼓 'hand drum' and 室 'room'. In order to successfully find it in the corpus, we would have to include a space between both characters. After each English affix has been translated into Japanese, we obtained a complete list of medical morphemes, named MEDICAL_JP.

Following this, we classified the items in MEDICAL_JP into prefixes, suffixes, and free morphemes. To that end, we extracted a list of all the prefixes and suffixes that appear in the corpus[3], and observed which ones appear in MEDICAL_JP. In this way, we created a list of possible medical prefixes and suffixes, and classified the rest of items into free morphemes (頚 and 頚部 are free morphemes).

4 Results and conclusion

The majority of items from MEDICAL_JP are free morphemes, which confirms our hypothesis that compounding is the most productive operation in the formation of Japanese medical terms (See Table 1), a very different process from other languages such as English and Spanish (Moreno-Sandoval et al. 2013):

	Types	Percentage
MEDICAL_JP	548	100%
Free Morphemes	444	81.02%
Medical Prefixes	43	7.84%
Medical Suffixes	61	11.13%

Table 1: Type Distribution in Japanese Medical Term Formation

References

Han, D., Ito, T., and Furugoori, T. (2003). A Deterministic Method for Structural Analysis of Compound Words in Japanese. *IEICE Transactions on Information and Systems, Pt.2 (Japanese Edition), J86-D-2*(5), 706–714.

Hisamitsu, T., and Nitta, Y. (1996). Analysis of Japanese compound nouns by direct text scanning. In *Proc. of the 16th Conference on Computational Linguistics – Volume 1* (pp. 550–555). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/992628.992723

Kageyama, T. (1982). "Word formation in Japanese". In Lingua 57, 215-258.

Kilgarriff, A., Rychly, P. Smrz, P. and Tugwell, D. (2004) The Sketch Engine. *Proc EURALEX*, Lorient, France. 105-116.

Kudo, M. (2007). A lexical semantic study *of four-character Sino-Japanese compounds and its application to machine translation* (Thesis). Dept. of Linguistics – Simon Fraser University.

Moreno–Sandoval, A. and Campillos-Llanos, L. (2013)

[1] http://chasen-legacy.sourceforge.jp/

[2] http://www.medo.jp/

[3] Using as query option jp_tag=="接頭詞-名詞接続" for the prefixes and jp_tag=="名詞-接尾-一般" for the suffixes.

Design and annotation of MultiMedica – a multilingual text corpus of the biomedical domain. *Proc. of the 5th International Conference on Corpus Linguistics* University of Alicante, Spain.

Moreno-Sandoval, A., Campillos-Llanos, L., González-Martínez, A. and Guirao-Miras, J. M. (n.d.) An affix-based method for automatic term recognition from a medical corpus of Spanish. *Proc of the International Corpus Linguistics Conference 2013*. Lancaster University, United Kingdom.

Shibatani, M. (1990). *The Languages of Japan*. Cambridge: Cambridge University Press.

Srdanović, E.I., Erjavec. T, Kilgarriff, A. (2008) "A web corpus and word-sketches for Japanese". In『自然言語処理』(Journal of Natural Language Processing) 15/2, 137-159.

Srdanović, E.I., Nishina, K. (2008) "The Sketch Engine corpus query tool for Japanese and its possible applications". In『日本語科学』(Japanese Linguistics) 23, 59-80.

Classifying fictional texts in the BNC using bibliographical information

Henrik Kaatari
Uppsala University
henrik.kaatari@engelska.uu.se

This poster introduces a new resource for users of the *British National Corpus* (henceforth BNC, see Burnard 2007) by presenting a sub-genre classification of the fictional texts included in the BNC. The sub-genre classification will be made publicly available in the form of a spreadsheet (similar in form to Lee's (2001) genre classification of the full BNC).

David Lee's (2001, 2002) genre classification of the BNC has many merits, the greatest being that some genres are hierarchically nested to allow for super- and sub-genre distinctions. Furthermore, in his BNC Index, Lee provides an abundance of information about the texts included in BNC. However, the 'fiction prose' genre (w_fict_prose) constitutes an anomaly in Lee's rigid classification scheme and it does so in two ways. First, the 'fiction prose' genre is not subcategorized into sub-genres such as 'mystery fiction', 'adventure fiction' etc. Second, unlike most of the written texts included in the BNC, Lee provides very little information (such as keywords and subject headings) on the texts included in the 'fiction prose' genre, making a further sub-genre classification virtually impossible. As a consequence, the 'fiction prose' genre is very broad compared to much more narrow genres such as 'newspapers broadsheets editorials,' making inclusions of these genres in subcorpora problematic since they are working at different levels of generality.

To address the lack of information regarding fictional texts in the BNC World Index, all titles in the 'fiction prose' genre that are marked as targeting an adult audience (i.e. excluding children's stories and titles targeting teenagers, based on the coding of the texts in the BNC Index) have been manually searched for in three different on-line catalogues in order to extract as much information about each title as possible. The catalogues consulted include Copac,[1] the Library of Congress catalogue (LOC) and the London Library Consortium catalogue (LLC).[2] All available information obtained from these

[1] Copac is an online library catalogue which merges the catalogues of major British and Irish academic libraries.
[2] www.copac.ac.uk; www.loc.gov; www.londonlibraries.gov.uk

searches has been stored in a spreadsheet, including Copac keywords and GSAFD genre headings (Guidelines on Subject Access to individual works of Fiction, Drama, etc. (the American Library Association 2000)). Based on this information, a sub-genre classification has been conducted. In classifying the material, I have relied only on the information from the three catalogues, with the aim of arriving at an objective categorization. I have thus steered away from trying to elicit information about the titles from other sources in a manner which would involve me reading extracts or summaries and on that basis categorize the titles. Since many titles have several different subject and genre headings the classification process was far from straightforward. The sub-genres and their distribution are given in Table 1.

Sub-genres	**Number of texts**	**Word total**
W_fict_prose_mystery	70	2,785,696
W_fict_prose_misc	66	2,628,636
W_fict_prose_general	66	2,443,043
W_fict_prose_romance	50	2,387,480
W_fict_prose_adventure	31	1,180,497
W_fict_prose_historical	27	1,066,401
W_fict_prose_sci/fi	20	814,507
W_fict_prose_humour	11	421,650
W_fict_prose_horror	6	232,853
W_fict_prose_child	4	118,617
W_biography	3	111,128
Total	**354**	**14,190,508**

Table 1. The sub-genre classification.

In total there are 354 texts in the 'fiction prose' genre targeting an adult audience. Three of these texts have been reclassified as 'biographies' (an already established genre in the BNC) and four texts as targeting a child audience (together with a number of other fictional texts already included in the BNC). The remaining texts have been classified into different sub-genres, most of them reflecting established fictional genres such as 'mystery,' 'romance' and 'adventure,' whereas texts abut which little or no information exists have been classified as miscellaneous (misc).

This new resource should be seen as an extension of Lee's BNC Index. It supplements the index and increases the flexibility with which the BNC can be used in at least two major ways. First, sub-genres become available for scholars interested in a particular sub-genre of fiction or interested in comparing multiple sub-genres. Second, the creation of subcorpora can be achieved in a more systematic way as the 'fiction' genre is now working at the same level of generality as genres such as 'newspapers' and 'academic prose' in that there are a number of different sub-genres available for all these genres. Scholars can use the sub-genres as they stand, and since all information (such as Copac keywords, GSAFD genre headings and shelfmark information) from the catalogues is available in the spreadsheet, they are also free to make their own classification as they see fit. The classification is fully compatible with the XML-edition of the BNC and can be obtained upon request via e-mail.

References

American Library Association. 2000. *Guidelines on subject access to individual works of fiction, drama, etc*. Chicago: American Library Association.

Burnard, L. 2007. *Reference guide to the British National Corpus (XML-edition)*. Available online at http://www.natcorp.ox.ac.uk/docs/URG/

Lee, D. 2001. "Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle". *Language Learning and Technology* 5(3): 37-72.

Lee, David. 2002. *Notes to accompany the BNC World (bibliographical) Index*. Available online at http://www.uow.edu.au/~dlee/home/BNCWIndexNotes.pdf

Identification of linguistic features for predicting L2 proficiency levels: Using Coh-Metrix and machine learning

Yuichiro Kobayashi
Japan Society for the Promotion of Science
kobayashi0721@gmail.com

Toshiyuki Kanamaru
Kyoto University
kanamaru@hi.h.kyoto-u.ac.jp

1 Introduction

Automated essay scoring is the ability of computer technology to evaluate and score written prose (Shermis and Burstein, 2003). The goal is to classify a large set of texts into a small number of discrete grades. Therefore, it can be considered a problem of text categorization or machine learning (Larkey and Croft, 2003). In machine learning, objectively measurable features in the essays are used as "exploratory variables" for the prediction of scores defined as "criterion variable".

2 Purpose

The purpose of this study is to explore how second language (L2) texts written by learners at various proficiency levels can be classified. By applying natural language processing and machine learning techniques to the assessment of L2 texts, the present paper proposes a new method for automated essay scoring, and the result is verified by the comparison with that of the existing essay scoring system.

3 Data

Corpus used in this study consists of 69 L2 essays written by Japanese university students. The total number of words is about 16000. The topics are based on two TOEFL exam samples, "New Product" (expository essay) and "Money on Technology" (persuasive essay). Writers were required to write their essays on computer, and all the essays were evaluated by e-rater (Attali and Burstein, 2006). In this study, the scores (Lv. 2 to Lv. 6) graded by e-rater were regarded as the proficiency levels of writers, and used as criterion variable (There is no Lv. 1 essay in our corpus).

L2	L3	L4	L5	L6
3	8	17	33	8

Table 1. Data

4 Coh-Metrix

Explanatory variables for this study are lexical indices related to breadth of lexical knowledge (word frequency, and lexical diversity), depth of lexical knowledge (hypernymy, polysemy, semantic co-referentiality, and word meaningfulness), and access to core lexical items (word concreteness, familiarity, and imagability). These indices are extracted from L2 texts using the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, and Cai, 2004).

5 Random forests

After extracting these features, random forests (Breiman, 2001) was employed to predict the proficiency levels using the features extracted above. Random forests can be defined as an algorithm for statistical classification and machine learning, and it is known as a powerful method for text classification and feature extraction (Hastie, Tibshirani and Friedman, 2009). Building a classifier in RF consists of two steps, bootstrap and ensemble learning. In the step of bootstrap, RF divided the input data into *n* sets of bootstrap samples randomly. Then, it constructs each classification tree based on each bootstrap sample. As a result, *n* classification trees are generated, independently. In the step of ensemble learning, all results of each tree are synthesized by majority decision. One of the most prominent features of random forests is its high accuracy in respect to classification of data set. Random forests is also very useful in that it gives estimates of what exploratory variables are important in the classification.

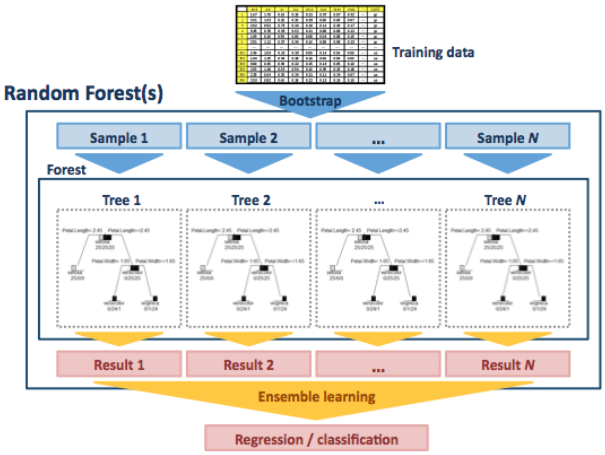


Figure 1. Random forests

6 Results and discussion

As a result of random forests with the out-of-bag error estimate, 53.62% of L2 essays were correctly classified.

	L2	*L3*	*L4*	*L5*	*L6*	*Accuracy*
L2	0	1	2	0	0	0.00
L3	0	1	6	1	0	12.50
L4	0	2	7	8	0	41.18
L5	0	0	4	29	0	87.87
L6	0	0	0	8	0	0.00

Table 2. Confusion matrix

According to the MeanGiniDecrease obtained as a result of random forests, the strongest predictors of an individual's proficiency level were number of words (READNW), number of sentences (READNS), average words per sentences (READASL), mean of location and motion ratio scores (SPATC), incidence of intentional actions, events, and particles, (INTEi) and noun phrase incidence score (DENSNP).

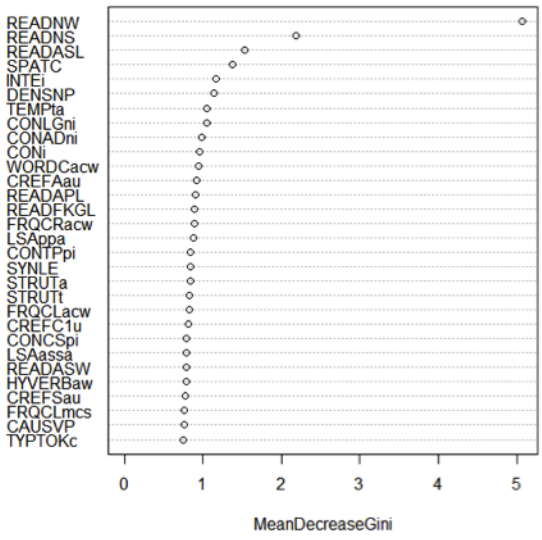


Figure 2. MeanGiniDecrease

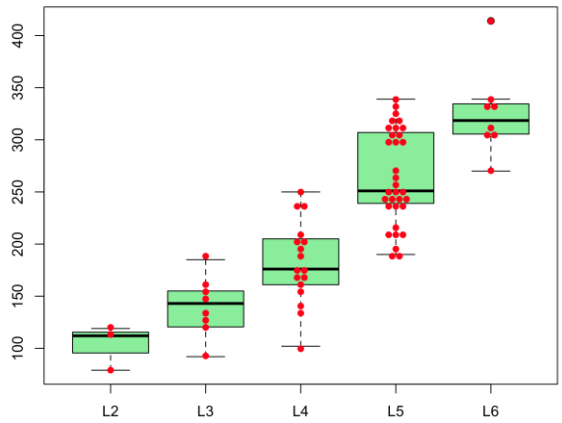


Figure 3. Boxplot with beeswarm (READNW)

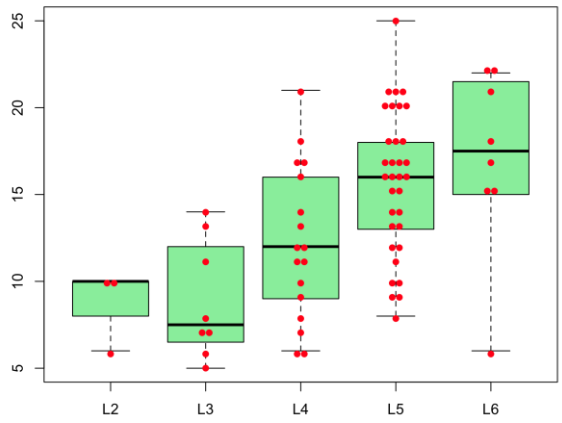


Figure 4. Boxplot with beeswarm (READNS)

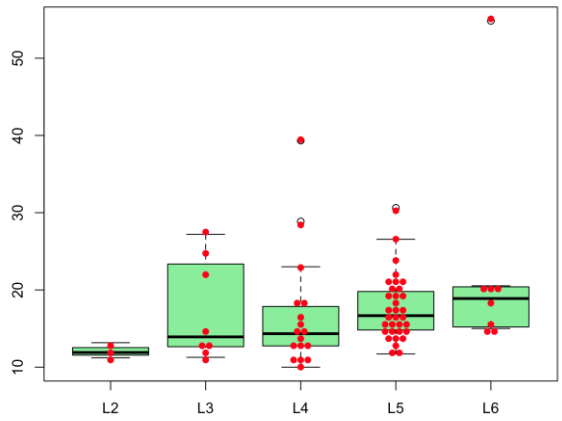


Figure 5. Boxplot with beeswarm (READASL)

7 Final remarks

The method used in this study will be useful for L2 learner profiling research (e.g. Hawkins and

Filipović, 2012). It can identify variables which serve as "criteria" for particular L2 proficiency levels. Moreover, it can elucidate the mechanism of existing automated essays scoring systems, such as e-rater.

References

Attali, Y., and Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment,* 4(3), 1-30.

Breiman, L. (2001). Random forests. *Machine Learning* 24, 123-140.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers,* 36, 193-202.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Second Edition. New York: Springer-Verlag.

Hawkins, J. A., and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: Cambridge University Press.

Larkey, L. S. and Croft, W. B. (2003). A text categorization approach to automated essay grading. In Shermis, M., and Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 55-70). Hillsdale: Lawrence Erlbaum Associates.

Shermis, M. D., and Burstein, J. (2003). Introduction. In Shermis, M., and Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii-xvi). Hillsdale: Lawrence Erlbaum Associates.

Corpus-driven terminology

Dominika Kováříková
Charles University
dominika.kovarikova@ff.cuni.cz

1 Introduction

In the last 30 years, there have been many attempts to approach terminology from a formal point of view for the purpose of automatic identification of terms in texts (Kageura and Umino 1996; Korkontzelos et al. 2008). Automatic term recognition (ATR) focuses on the extraction of terminology based on features such as frequency, distribution and other attributes processable by computers.

The ultimate goal of ATR is usually the highest possible success rate, but it can also be viewed as an outstanding resource for the theory of terminology and for the description of a term (Šrajerová 2009).

The terminology extraction should be as corpus-driven as possible for at least two reasons: 1. it is possible to use any text for the research, not just linguistically tagged corpora, and 2. the less linguistic premises are used in the research, the more theory-independent the conclusions are.

2 ATR method using data mining

The presented ATR method is based on data mining techniques. The approach has the following advantages:

- it can be corpus-driven to a great extent,
- the success rate of the term identification is very high, and
- the method is able to evaluate the importance of the individual attributes for the data mining procedure which provides an insight into characteristics of a term.

Data mining is defined as an (semi)automatic process of discovering meaningful patterns in data (Witten and Frank 2005). After a training procedure, it provides tools for finding a specific type (such as term or non-term) in given data. The data mining methods successfully deal with substantial quantities of data and are thus suitable for exploring the large data available in language corpora.

The corpus data used in the presented research is in fact much more extensive than just the number of text positions or their frequency in a corpus: they include various statistical,

distributional and other attributes that can be calculated for individual types or tokens. The attributes selection is based on previous research (e.g. Bečka 1972; Yang 1986; Šrajerová et al. 2009; Čermák 2010; Cvrček 2012).

The new ATR method based on data mining is able to identify terms and/or assign a degree of terminological value to each text position in any data provided in the appropriate format.

The material for this research was extracted from the Czech National Corpus, but the method is usable in other languages as well providing there is a suitable material available.

3 Corpus-driven approach

This particular approach to terminology extraction is corpus-driven to a great extent. It is possible to regulate the extent of linguistic knowledge and interpretation (in form of linguistic labelling) used in the individual experiments and thus to explore its influence on the success rate.

Even though linguistic labelling provides an increase of the success rate of some degree, the benefits of the corpus-driven approach are so compelling that neither lemmatization nor morphological tags are used in the presented research.

However, the method requires manual labelling of terms and non-terms in the training data. Whether the word is labelled as a term or a non-term, is based on the knowledge of terminology and on existing terminological dictionaries.

4 Success rate of the ATR

The method based on data mining is highly successful in extracting terminology from texts. The most successful data mining methods are able to correctly classify around 95% of the running words in text as terms or non-terms. The success rate changes in various academic disciplines – it is generally higher in natural and applied sciences and lower in humanities and social sciences.

Table 1 shows the success rate of a data mining method (J48graft) in labelling one-word terms and non-terms in the training data. Note that the conclusions must be based on the comparison with a simple method ZeroR which indicates the number of non-terms in the training texts (the greater the difference between the two methods, the better the result). Both methods are components of a data mining tool Weka (Hall et al. 2009).

Discipline	ZeroR	J48graft
computer science	75.3%	95.6%
literature	90.0%	94.7%
medicine	74.8%	94.9%
sociology	86.4%	95.4%
all 4 disciplines	81.8%	94.4%

Table 1: Success rates of the ATR method

The high success rate is not the main goal of the research, it is however very important because it indicates that the conclusions based on the results of the method are quite reliable.

5 Evaluation of the attributes

Some of the data mining methods are able to evaluate the importance of the individual attributes for the term recognition process. This feature ranking provides an insight into (mainly statistical and distributional) characteristics of a term and is an important resource for the theory of terminology.

The feature ranking suggests that among the characteristics that distinguish terms from non-terms is their frequency, distribution throughout corpus, distribution in individual academic disciplines, and context. Specifically, the more frequent the word is in a given academic discipline in comparison to the general corpus (fiction and journalism), the higher is the probability it is a term. The more evenly the word is distributed throughout the whole corpus, the lower the probability of it being a term. If the word occurs only in a small number of academic disciplines, it is more likely to be a term. High entropy of the right context (very variable right context) of the word occurs mainly in case of non-terms (more in Kováříková 2013).

6 Data mining in linguistic research

It is likely that there are other linguistic topics that can benefit from the data mining and its ability to deal with large quantities of data. Among such topics are author identification, stylometry, part-of-speech determination, research of grammatical categories, study of context etc.

7 Conclusions

Data mining shows outstanding results in the domain of automatic term recognition. The approach is corpus-driven to a great extent, and the method has quite high success rates (especially in natural and applied sciences). The data mining tools are able to evaluate various attributes in terms of their importance for ATR. Such knowledge is helpful in describing (formal)

characteristics of a term.

The high success rate of the presented method as well as the fact that there is no linguistic labelling needed in the actual term extraction makes it possible to apply the results, e.g. in an online tool (based on the Czech National Corpus) for extracting terms from any inserted text.

References

Bečka, J.V. 1972. "The lexical composition of specialized texts and its quantitative aspect". *Prague Studies in Mathematical Linguistics* 4: 47-64.

Cvrček, V. 2012. *Kvantitativní analýza kontextu*. Habilitation thesis, Charles University in Prague.

Čermák, F. 2010. *Lexikon a sémantika*. Prague: NLN.

Český národní korpus – SYN2010. Prague: Institute of the Czech National Corpus, Charles University in Prague. Available online at http://www.korpus.cz.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. "The WEKA data mining software: An update". *SIGKDD Explorations* 11 (1).

Kageura, K. and Umino, B. 1996. "Methods of automatic term recognition: A Review". *Terminology* 3 (2): 259-289.

Korkontzelos, I., Klapaftis, I.P. and Manandhar, S. 2008. "Reviewing and evaluating automatic term recognition techniques". In A. Ranta and B. Nordström (eds.) *GoTAL 2008*. Berlin: Springer-Verlag.

Kováříková, D. 2013. *Automatické vyhledávání termínů pomocí data-miningových metod*. Unpublished PhD thesis, Charles University in Prague.

Šrajerová, D. 2009. "Automatic term recognition as a resource for theory of terminology". *CL 2009 Proceedings*. Available online at http://ucrel.lancs.ac.uk/publications/cl2009/

Šrajerová, D., Kovářík, O. and Cvrček, V. 2009. "Automatic term recognition based on data-mining techniques". *CSIE 2009 Proceedings*. Los Angeles.

Witten, I.H. and Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Yang, H. 1986. "A new technique for identifying scientific/technical terms and describing science texts: an interim report". *Literary and Linguistic Computing* 1 (2): 93-103.

Learner corpus of L3 acquisition

Hui-Chuan Lu
National Cheng Kung University
huichuanlu1@gmail.com

An Chung Cheng
University of Toledo
accheng99@gmail.com

1 Introduction

Expanding on the results of the corpus construction of CATE (Corpus de Aprendices Taiwaneses de Español/ Taiwanese' Learner Corpus of Spanish) in the past seven years, this paper sets out to develop a sub-corpus, COATE (Corpus Oral de Aprendices Taiwaneses de Español/ Taiwanese Learners' Oral Corpus of Spanish). The other objective of this paper is to study the acquisition of Spanish tense and aspect in the written and spoken texts with semantic analysis based on the created corpora.

2 A unique learner corpus

The research team of the CATE has compiled 2400 texts, 340,000 words, for the written learner corpus CEATE (Corpus Escrito de Aprendices Taiwaneses de Español/ Taiwanese Learners' Written Corpus of Spanish) in 2005-2012. The unique CATE included written work of L1 Chines-speaking learners of Spanish as a third language (L3), who learned English as a second language (L2) at schools.

The addition of speech data in the sub-corpus, COATE provides richer data and search results of more variety for study of L3 acquisition than the CEATE. In the process of constructing the COATE, the research team collected oral data from three levels (beginning, intermediate, advanced) of learners of Spanish in Taiwan, where Mandarin-Chinese is spoken in daily life and English is learned at middle or high schools. All participants had signed consent form to authorize the future usage of compiled data, provided linguistic profile and language learning experience related to Spanish. Since the participants are college students, the Wisconsin Placement Test wildly used in the US colleges was administered to identify learners' Spanish proficiency level for posterior analysis. In the audio laboratory, the participants recorded oral narrative by describing a picture series with key words provided at the side of each picture, illustrating a fairy tale, Little Red Riding Hood in a period of 18 minutes. The speech data then were transcribed and annotated with assisted tools.

Hispanic native speakers annotated the learner errors. Furthermore, the research team POS-tagged and annotated the verbs of two Spanish past aspects (preterite and imperfect) according to the categories of verbal predicates (Vendler, 1967).

To strengthen the research value of the corpus of multilingual learners, CPEIC (Corpus Paralelo de Español, Inglés y Chino/ Trilingual Parallel Corpus of Spanish, English and Chinese) was also constructed as a base line for comparison and linguistic analysis for the research of the acquisition of L3 Spanish learners.

3 The study of tense/aspect acquisition of L3 Spanish learners

The learner corpus provides rich data of learner performance in written and oral narrative. One area of the research with the CEATE and COATE is the study of the development of past tense and aspect of Spanish. Under the theoretical framework of Lexical Aspectual Hypothesis proposed by Andersen (1986, 1991), the study examined the developmental patterns of L3 Spanish acquisition. The research question is: Do the developmental patterns differ between written and oral production of L3 learners with respect to the usages of lexical aspect in the acquisition of Spanish tense? Besides the comparison between written and oral learners corpus, the research team took a trilingual parallel corpus CPEIC as a reference. The results of the contrastive analysis of the three paralleled native languages in Spanish, English and Chinese will serve as a baseline to examine the influence of the input of learners' first and second language on the acquisition of their third language.

For the implication of the findings, this study will suggest approaches and strategies for teaching the Spanish past tense and aspect by comparing the language output of learners of Spanish and that of native speakers, as well as conducting contrastive analysis of natural data by native speakers of Spanish, English and Chinese.

In conclusion, this paper combines the construction of corpus, the study of corpus-based foreign language acquisition and contrastive analysis of multiple languages. The fruitful results of the current research will lead to an effective foreign language teaching on the target forms.

References

Andersen, R. (1986). El desarrollo de la morfología verbal en el español como segundo idioma. In J. Meisel, (Ed.) *Adquisición del lenguaje/ Aquisição da linguagem*. Frankfurt: Vervuert.

Andersen, R. (1991). Developmental sequences: The emergence of aspect marking in second language acquisition. In T. Huebner and C.A. Ferguson (Eds.). *Crosscurrents in second language acquisition and linguistic theories*. Amsterdam: John Benjamins.

Vendler, A. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

PMSE: text categorization – a case study

Jiří Mácha
ÚČNK
jm@petamem.com

Jiří Václavík
ÚČNK
jv@petamem.com

1 Introduction

This poster presents a new corpus tool called PMSE and its first major application, categorization of textual documents.

2 PetaMem Scripting Environment

The PetaMem Scripting Environment (PMSE) is a generic tool for batch processing of large text corpora and can also serve as middleware for other corpus processing tools.

The software suite is written in Perl and consists of components which enable the user to acquire statistical information from the given texts (corpora). Great attention was paid to maintaining as universal applicability of the tool as possible. Similar to UNIX philosophy, PMSE is a building kit of small units that can be combined together in a different order, not a tool focused on one specific task or functionality.

The crucial parts of PMSE are represented by scripts intended for computing of language-statistical information, e.g.:

- Frequencies of token occurrences
- Token probabilities
- Generating of n-grams
- Computing numerous association measures for n-grams of various length (e.g. MI-score, t-score)
- Computing various distance measures among pairs of n-grams

The whole tool chain of PMSE is designed to process texts from the very beginning (downloading texts, converting formats, removing formatting etc.) to the final task – computation of the statistical characteristics of the data and their visualization.

PMSE is designed to be language independent, it works with plain texts encoded in UTF-8. PMSE is an ongoing effort and still under development, the available functionality, however, already allows for interesting real-world applications.

Future plans for development involve e. g.: graphical web interface, conversion between various corpora formats and annotation tags.

3 Text Categorization

The authors will present an application case for PMSE – a Text Categorization project (abbreviated: TextCat). The general task for the TextCat app is to categorize various documents in any language. The authors will present an example of already finished categorization of parallel texts in about 20 European languages. Resulting dendrograms (one for each language from the parallel corpora) show signs of similar structures. The evaluation of individual dendrograms is part of future linguistic interpretation.

The modularity of the source code allows the user to change the behaviour of all procedural steps, especially since TextCat is extensible by simple plugins. Also, a great attention was paid to the performance of the software (efficiency as well as parallel processing).

The linguistic criteria addressed and delivered by TextCat may be defined by the user. The categorization process has several steps:

Extract text from all documents.
Pre-process all the texts, extract n-grams of any size.
Filter n-grams according to specific criteria .
Filter files, exclude the inconvenient (some files could be damaged, too short, too long, non-relevant etc.).
Create and precompute all existing groups (each text belongs to one group) and compute distances for all the possible pairs.
Find the closest groups and join them into a parent group.
Repeat the previous two steps until only one group remains.
Visualize a binary tree representing relations among the texts (dendrogram).

TextCat is a modular framework which could perform categorization on any criteria – that is why it has a high coverage. It could be used for language identification, corpus sorting, forensic linguistics etc.

References

Huang, A. 2008. *Similarity Measures for Text Document Clustering*. Available on-line at http://favi.com.vn/wp-content/uploads/2012/05/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

CLEG and "die Deutschen"

Ursula Maden-Weinberger
Lancaster University
u.weinberger@lancs.ac.uk

1 Introduction

This presentation accompanies the public release of a new resource for learner corpus research: CLEG – a *C*orpus of *LE*arner *G*erman, and exemplifies its use in an exploration of British students' views of Germany and the German people.

CLEG is a significant expansion of learner corpus resources for the German language. At present, most of the research effort in learner corpora is focused on English and while the amount of learner corpus resources for English has been growing substantially over the last decade, other languages receive less attention – one such language is German. (The learner corpus database of the Centre for English Corpus Linguistics at Louvain-la-Neuve lists 73 English learner corpora – most of which publicly available – as opposed to 9 German learner corpora). To date, the only notable publicly available German learner corpus is the *Fehlerannotiertes Lernerkorpus (FALKO)* (error-annotated learner corpus), collected and managed at the Humboldt Universität, Berlin (Reznicek et al. 2012).

While research into German learner language is, of course, a valuable and insightful enterprise in itself, there is another reason why the expansion of resources and research involving languages other than English is crucial. Research already carried out on CLEG data has provided evidence that could be used to triangulate findings from other L2 corpus studies and thus explore general, universal tendencies in learner language. These are trends that are not just L1-, but also L2-independent. One such undertaking with CLEG data has revealed that learners, regardless of L1 or L2, exhibit a general tendency to overuse personalised expressions in writing (Maden-Weinberger 2012).

2 The CLEG corpus

CLEG is an approx. 300,000 word corpus of advanced second language learner writing. Contributors and texts are tightly controlled for a number of relevant criteria. The learners are undergraduate students of German at Lancaster University who had achieved an A-level in German (this equates to between 5-7 years of school tuition in German and a CEFR level of B1 to B2). They are all native speakers of British English and between 18 and 22 years old. Texts were collected over a period of four years from all three years of the undergraduate programme (called Year A, Year B and Year C). Most students in the first two years (Year A and Year B) have spent a few weeks in Germany on vacation or as part of a school exchange. The students in the final year (Year C) have all spent between six and twelve months in a German-speaking country as part of their "year abroad", which is a compulsory part of the degree scheme for all language majors in the third year of study.

The texts chosen for the corpus can be classified as "expository-argumentative". This is defined in a operational way as texts where the task instructions imply "the presentation and weighing up of arguments, writer's criticism or systematic outlines of abstract concepts" (Lorenz 1999:12). Incidentally, expository-argumentative texts are also the kind of texts that learners are asked to produce most frequently throughout their study course at Lancaster University. This means that this collection criterion yielded the largest amount of reasonably homogeneous texts from all year groups. Metadata on learner and text profiles are stored in a database, which enables researchers to link each text to the relevant text and learner information and create sub-corpora according to different specifications if desired.

3 Truly longitudinal data

As the data was collected over the course of four years, all of the data is quasi-longitudinal, however there is a core of one entire cohort of students whose data is truly longitudinal from their first year through to the end of the four year degree programme. This is a distinctive feature of CLEG, which provides unique opportunities for corpus research on developmental aspects of learner writing.

4 CLEG and "die Deutschen"

In a "taster" example, the development of students' views on Germany and the German people are explored through the collocations around the item "deutsch*" in the three year-groups. The following examples provide a first glance at the kind of statements to be found in the student texts:

Die Deutschen...

[the Germans...]

...sind sehr freundlich. (Year A)

[...are very friendly]

...berichten viel über Fußball. (Year B)
[...give football a lot of coverage]
...sind Experten in Bezug auf technischen Qualität. (Year C)
[...are experts in technical quality]

Although in its infancy at the moment, this study should provide some clear evidence of a development of opinions about Germans from the British students and it will be particularly interesting to investigate the impact of the time spent in Germany before the final year of study.

References

Lorenz, G. (1999). Adjective intensification – Learners versus Native Speakers. A corpus study of argumentative writing. Amsterdam – Atlanta: Rodopi

Maden-Weinberger, U. 2012. "Personal expressions in Learner German – it's all about the bigger picture". Presentation at the Teaching and Language Corpora Conference, Warsaw, 11-14 July 2012

Reznicek, M., Lüdeling, A. and Schwantuschke, F. 2012. Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.01. Berlin

Conditionals in 18th-century philosophy texts: A corpus-based study[1]

Leida Maria Monaco
Universidade da Coruña
leidamaria.monaco@udc.es

Luis Puente Castelo
Universidade da Coruña
luis.pcastelo@udc.es

If a conditional clause can be defined as "a rhetorical device for gaining acceptance for one's claims" (Warchal 2010: 141), it is not surprising that conditionals play a very important role in argumentative strategies in virtually any register within both spoken and written discourse. This is especially the case in scientific register, where conditionals appear to be particularly frequent (Athanasiadou & Dirven 1997; Ferguson 2001: 69), not only in order to express the relationship between a phenomenon and its consequence, but also to state a hypothesis and/or speculate on possible outcomes of events. Moreover, it has been noted that conditionals are frequently used as downtoning devices by means of which a claim would presumably sound less assertive or categorical, functioning thus as metadiscoursive strategies, or hedges, and therefore playing a mediating role in the relationship between the authors and their discourse community (Hyland 1994, 1998a, 1998b; Declerck & Reed 2001; Carter-Thomas & Rowley-Jolivet 2008; Warchal 2010: 141-142).

Within scientific discourse, the argumentative function of conditionals may be noted best in highly speculative fields such as that of Philosophy, as "conditionals feature prominently in deductive argument – for example in the classical rules of inference known as modus ponens and modus tollens" (Ferguson 2001: 62). This argumentative function is expected to be particularly prominent in earlier philosophical texts, as these stem from the decaying roots of scholasticism, a knowledge framework which paid special attention to the logical form of the argument, frequently demanding the use of "common logical terms to organize discourse and build up the arguments" (Taavitsainen 1999: 249).

On the other hand, conditionals are one of the

[1] This research was funded by the Consellería de Educación e Ordenación Universitaria (I2C plan, reference number Pre/2011/096, co-funded 80% by the European Social Fund) and the Ministerio de Ciencia e Innovación (FPU grant, reference number AP2009-3206). These grants are hereby gratefully acknowledged.

linguistic parameters used in Biber's (1988, 1995) multi-dimensional analysis of register variation, namely in his Dimension 4 labelled "Overt expression of persuasion". According to Biber (1988: 111, 148-151), the linguistic features co-occurring in Dimension 4 have a common underlying suasive function. Likewise, in further multi-dimensional studies, such as Biber's multi-register overview of 18th century texts (2001) and the new multi-dimensional analysis of present-day English academic writing (Biber et al. 2004), conditional subordination appears to be grouped with linguistic features conveying an 'oral', vs. a 'literate', type of discourse. This may appear contradictory at first sight, as the relatively high use of conditionals in scientific discourse conflicts with its grouping with oral features in Biber's most recent work. A reason for such a characterisation of conditionals is presumed to lie in their inherent speculative function, which could be described otherwise as self-persuasion from an introspective point of view or as a mechanism to direct the discourse and to put forward the author's ideas to the readers from a pragmatic perspective (Biber 1988: 111).

The aim of this study is to describe the wide-ranging field of conditional sentences in eighteenth-century philosophy texts, describing their diachronic evolution, the uses of the different varieties and their pragmatic functions in the text and in the construction of scientific discourse. For this purpose we have adopted Warchal's (2010) functional classification, which covers up to eight different functions in the use of conditionals in scientific writing, ranging from conditionals which frame the logical argumentation of the discourse (i.e. content conditionals) to the different types of hedging conditionals used to speculate or to downtone the certainty of claims (i.e. epistemic and speech act conditionals) as is customary in the discourse of modern science. Our working hypotheses are that there should be a correlation between the use of specific conditional subordinators in the different texts and the scores of those texts for Biber's (1988) Dimension 4 and that there should be a diachronic evolution in the importance of the different functions of conditional sentences along the century as scientific discourse drifts apart from the old scholastic trends towards the modern scientific methods.

This research is carried out on a ca. 200,000-word corpus corresponding to the eighteenth-century half of *CEPhiT* (*Corpus of English Philosophy Texts*), a subcorpus of the *Coruña Corpus of Scientific Writing* (Moskowich & Crespo 2007; Moskowich 2011). The whole of *CEPhiT* contains forty 10,000-word text samples spanning from 1700 to 1900, at a rate of two per decade. We consider that *CEPhiT* is a representative corpus, as its compilation has been completed by taking into account different parameters such as the sex of the authors, their place of education, as well as the different text types used for writing Philosophy in English during the time span it covers. Our corpus comprises a total of twenty eighteenth-century texts, among which thirteen are treatises, six are essays and only one is a textbook, written by three female and seventeen male authors. Both the genre and the sex categories are borne in mind for the analysis of our results.

In order to find the different conditional uses, the corpus has been searched for selected conditional particles by using the *Coruña Corpus Tool* (Moskowich & Parapar 2008), a search engine specifically designed to work with texts from the *Coruña Corpus*, which allows for multi-word and metadata-based searches. However, given that *CEPhiT* had not yet been annotated for grammatical categories while this research was carried out, the lists of occurrences obtained automatically with the tool have had to undergo manual disambiguation in order to eliminate all non-conditional uses of the particles.

Once our search has been completed, the different conditional uses are examined from a double perspective to comply with both our objectives. First, the function of each conditional structure in the text is described in order to examine the diachronic evolution of functions. After that, the different occurrences of each subordinator are counted in an attempt to find correlations with each text score in Biber's Dimension 4, using data from previous research conducted on eighteenth-century *CEPhiT* (Crespo 2011). This research is a starting point for a further comparative study of conditionals across different scientific registers.

References

Athanasiadou, Angeliki and René Dirven. 1997. *On conditionals again*. Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2001. Dimensions of variation among eighteenth-century speech-based and written registers. In Biber, Douglas, and Susan Conrad (eds.), *Variation in English: Multi-Dimensional Studies*, 200-214. Essex: Pearson Education.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua. 2004. *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton, NJ: Educational Testing Service.

Carter-Thomas, Shirley, and Elizabeth Rowley-Jolivet. 2008. *If*-conditionals in medical discourse: From theory to disciplinary practice. *Journal of English for Academic Purposes* 7: 191-205.

Crespo, Begoña. 2011. Persuasion markers and ideology in eighteenth century philosophy texts. *Revista de Lenguas para Fines Específicos* 17: 199-228.

Declerck, Renaat, and Susan Reed. 2001. *Conditionals: A comprehensive empirical analysis*. Berlin: Mouton de Gruyter.

Ferguson, Gibson. 2001. If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes* 20: 61-82.

Hyland, Ken. 1994. Hedging in academic writing and EAP textbooks. *English for Specific Purposes* 13(3): 239-256.

Hyland, Ken. 1998a. *Hedging in scientific research articles*. Amsterdam: Benjamins.

Hyland, Ken. 1998b. Persuasion and context: the pragmatics of academic metadiscourse. *Journal of Pragmatics* 30: 437-455.

Moskowich, Isabel. 2011. "The golden rule of divine philosophy" exemplified in the Coruña Corpus of English Scientific Writing. *Revista de Lenguas para Fines Específicos* 17: 167-197.

Moskowich, Isabel, and Begoña Crespo. 2007. Presenting the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing. In Pérez Guerra, Javier et al. (eds.) *Of Varying Language and Opposing Creed: New Insights into Late Modern English*. 341-357. Bern: Peter Lang.

Moskowich, Isabel, and Javier Parapar. 2008. Writing science, compiling science: The *Coruña Corpus of English Scientific Writing*. In Lorenzo Modia, María Jesús (ed.), *Proceedings from the 31st AEDEAN Conference*. 531-544. A Coruña: Universidade da Coruña.

Taavitsainen, Irma. 1999. Dialogues in English Medical Writing. In Jucker, Andreas H.; Gerd Fritz and Franz Lebsanft (eds.), *Historical Dialogue Analysis*. 243-268. Amsterdam/Philadelphia: John Benjamins.

Warchal, Krystyna. 2010. Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. *Journal of English for Academic Purposes* 9: 140-150.

The Czech preposition *v/ve* and its English equivalents

Renata Novotná
Charles University
renata.novotna@ff.cuni.cz

The linguistic data used for this study was searched in the English-Czech parallel corpus built as a part of the InterCorp project. The number of occurences of the preposition *v* in this corpus is 69114. It was impossible to study this material as a whole, therefore 15 random samples of 100 sentences each were extracted automatically, i. e. 2% of the material.

The emphasis of this study was laid mainly on the valency functions, i.e. adverbal and adnominal ones. According to Syntagmatics and Paradigmatics of the Czech Word (Čermák and Holub 2005) the adverbal function was divided into two groups:

a) valency – cases such as *spočívat v* (to be based on), *změnit se v* (to change into) etc., where the preposition *v* expresses an abstract relation between the verb and the noun

b) addition/complement – such as *ležet v /blátě/* (to lie in the mud), in which the place, but not the abstract relation is expressed by the preposition; in this case the preposition *v/ve* can be replaced by other prepositions, such as *na*, e.g. *ležet na /zemi/* (to lie on the ground) etc.

The adnominal function was concentrated on nouns and adjectives, i. e. their valency. As noun valency is slightly different from verbal valency, abstract and concrete nouns were treated separately:

a) valency of abstract nouns, such as *život v cizině* (the life abroad)

b) valency of concrete nouns, e. g. *klíč v zámku* (the key in the lock)

The valency of the adjectives was studied as a separate group, e. g. *obratný v řeči* (skilful in conversation).

While the valency of verbs, nouns and adjectives expressed by the preposition is the inherent attribute of the lexical item with the categorial feature and therefore the verb/noun/adjective with the valency preposition can be taken as one lexical item or one lexeme; the adverbial function of the preposition is independent and forms a collocation, e.g. *v létě – in summer*.

Besides the groups given above, fixed collocations and idioms were also studied, such as *v souvislosti s* (in connection with), *držet v zajetí* (to hold captive), *převracet něco v mysli* (to turn

sth over in one's mind), *mít v něčem prsty* (have a hand in sth), *být v tahu* (to be a goner) etc.

According to the English equivalents given in the 15 extracted samples the groups of lexical valency, collocations and idioms were arranged to the following scheme:

I – lexemes:

1) verb

a) valency – e.g. *pokračovat v předčítání – to go on with the talk*

b) complement – e.g. *ležet v posteli – to lie in the bed*

2) noun

a) valency of the abstract noun – e.g. *záliba v určitém druhu krutosti – taste for a special kind of cruelty*

b) valency of the concrete noun – e.g. *kamery ve výtahu – cameras in the elevator*

In some cases the English equivalent is expressed not by a preposition but by an adjective, where the essential meaning of the adjective-noun combination in English is the purpose:

c) valency of the concrete noun in Czech expressed by adjective in the attribute position in English, e.g. *pohovka v obývacím pokoji – living room sofa*

3) adjective valency – e.g. *ponořený v sadech – buried in orchards*

II – collocations:

1) adverbials – e.g. *v listopadu – in November*

2) fixed collocations – e.g. *v podstatě – in general*

III – idioms: This is also a separate group as it can include both collocations and sentences, e.g. *obrátit oči v sloup – to roll someone's eyes heavenward*, *při tom mi tuhla krev v žilách – it made my blood run cold*

IV – mixed group: This group incorporates the various possible differences in combinations of one-word and multi-word units between the two languages:

1) collocation – lexeme, e.g. *ve skutečnosti – really, v naději – hoping*

2) idiom – lexeme – e.g. *mít v úmyslu – intend, říkat si v duchu – to wonder*

The last group is composed of quasi-idioms, i.e. combinations of verbs and abstract nouns having one-word equivalents in the other language:

3) quasi-idiom – lexeme – e.g. *být v pokušení – to be tempted, být v rozpacích – to be embarassed, být v šoku – to be shocked*

V – Particular differences: As a separate group, cases of particular differences were also studied:

The most typical examples are the addition of such collocations as *v první chvíli* (in the first moment) etc.

The poster shows the variability of English equivalents, esp. of the verbal and noun valency. As an example we can show the variability of adverbials, such as *v Londýně – in London, ve vlaku – on the train, v Rosině příbytku – at Rosa's*; *v září – in September, v pondělí – on Monday, v pět hodin – at five o'clock* etc.

References

Čermák F. 1991. Podstata valence z hlediska lexikologického, In *Walencja czasownika a problemy leksykografii dwujęzycznej*, ed. D. Rytel-Kuc, Wrocław-Warszawa-Kraków: Wydawnictwo polskiej akademii nauk, 15-40.

Čermák F. 1996. Systém, funkce, forma a sémantika českých předložek. *Slovo a Slovesnost* 57, 30-46.

Čermák F. and Křen M. (eds) 2004. *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny.

Čermák F. 2005. Abstract Noun Collocations: Their Nature in a Parallel English-Czech Corpus. In *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, Eds. G. Barnbrook, P. Danielsson and M. Mahlberg. London,-New York: Continuum, 143-153.

Čermák F. and Holub J. 2005. *Syntagmatika a paradigmatika českého slova I. Valence a kolokabilita*. Praha:Nakladatelství Karolinum.

Čermák F. (ed) 2007. *Frekvenční slovník mluvené češtiny*. Praha: Nakladatelství Karolinum.

Čermák F. (ed.) 2007. *Slovník Karla Čapka*, 2007. Ed. F. Čermák. Praha: Nakladatelství Lidové noviny.

Čermák F. (ed) 2009. *Slovník české frazeologie a idiomatiky* I-IV. Academia, Praha.

Čermák F. and Cvrček (eds) 2009. *Slovník Bohumila Hrabala*. Praha: Nakladatelství Lidové noviny.

Čermáková A. 2009. *Valence českých substantiv*. Praha: Nakladatelství Lidové noviny.

Klégr A., Malá M. and Šaldová P. 2012. *Anglické ekvivalenty nejfrekventovanějších českých předložek*. Praha: Nakladatelství Karolinum.

Kopřivová M. 2006. *Valence českých adjektiv*. Praha: Nakladatelství Lidové noviny.

Novotná R. 2009. The Czech preposition na and its English Equivalents. In: *Intercorp: Exploring a Multilingual Corpus*, eds F. Čermák, P. Corness and A. Klégr. Praha: Nakladatelství Lidové noviny, 138-145.

Business ethics documents of French companies from an intercultural point of view: Example of a contrastive study of the French and American versions of Lafarge's *Principles of Action*

Emmanuelle Pensec
University of South Brittany
emmanuelle.pensec@univ-ubs.fr

In the context of global markets, English has become the language used for communication in large international organizations. However, globalization does not mean homogenization of outlooks and of culture. It even seems that we are assisting at the maintenance of discursive and cultural communities.

Globalization of the economy and finance has developed alongside financial scandals such as those of *Enron* or *Wordcom*. In their attempt to bring morality into their practices, companies have had to consider their stakeholders when making decisions. This corporate social responsibility is clearly discernible through the writing of ethical documents for the stakeholders and civil society in general. The companies aim to answer to the expectations of society. These documents for external communication reflect publicly the company's values. In this context, how do international companies communicate a common ethics policy to all their subsidiaries? Striking a balance between cultural diversity and business strategy is a real challenge for these companies.

It would seem that the communication, dissemination and the content of the business ethics documents changes depending on the different cultural communities. We will try to understand the reasons for the needs for such changes while the situation would have us believe that globalization makes everybody think in the same way and respond to the same thinking and same consumer standards.

Through a contrastive analysis of the French and American versions of Lafarge's *Principles of Action*, we will see how the company deals with cultural diversity while transmitting their own values and their corporate strategy. We have analysed each of the two versions of the ethical principles with corpus linguistics tools such as *Lexico* and *Antconc* so as to determine the cultural and discursive changes.

The results of our contrastive study imply that there is a different cultural approach of the company to its different locations. The use of Corpus Linguistics tools highlights that these differences address three dimensions (Hofstede, 2001):

The position of the company in relation with its economic partners
Relationships of the employees with hierarchy
The community spirit within the company

Each time the company has dealt with one of these three dimensions, we can observe several repeated words derived from one of those as in Table 1.

Dimension	*Key term in English*	*Key term in French*
The position of the company in relation with one of its economic partners	Deliver Provide	Répondre Offrir
Relationships of the employees with hierarchy	Employees Being a customer driven organization	Collaborateurs Orienter notre organisation vers le client
The community spirit within the company	Members of our communities	Citoyen

Table 1

So as to underline these cultural and discursive differences, we have compared these key terms of our corpus to French and American national corpora so as to try to understand the cultural variations. The results of the analysis imply that there is a greater tendency to community spirit (3rd dimension) both in the company and in American society than in the French version. The French version adopts a more detached stance, showing its superiority compared to its employees (2nd dimension) and other companies (1st dimension) contrary to the American version in which we observe the company as a part of a whole (3rd dimension). We hypothesize that business culture can reflect the founding values of a company. Lafarge, which can be qualified as a paternalistic company, was founded in the nineteenth century, and promotes values of mutual assistance and sharing.

Lafarge has opted for a communicative translation (Communicative translation attempts to produce on its readers an effect as close as possible to that obtained on the readers of the original (Newmark, 1981: 39)) when choosing to culturally adjust its ethical principles. This communication strategy (corporate communication) can be explained by the company's presence in the United States and the

number of its employees (14.6% of its global workforce in 2010). We can consider this as a strategic option for Lafarge.

The choice of this micro-corpus can be justified by the cultural and discursive interests in a bilingual approach. Corpus linguistics tools have helped to demonstrate the importance of the intercultural dimension in the communication strategies of companies'.

References

d'Iribarne, P. 2009. L'épreuve des différences, L'expérience d'une entreprise mondiale. Paris: Seuil.

Hofstede, G. 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Accross Nations. London: Thousand Oaks.

Moirand, S., & Tréguer-Felten, G. 2010. "Des mots de la langue aux discours spécialisés, des acteurs sociaux à la part culturelle du langage: raisons et conséquences de ces déplacements". Asp 51-52. Available online at http://asp.revues.org/465

Newmark, P. (1981). Approaches to translation. New York: Oxford.

Schlegelmilch, C. C. (1990, 4th Qtr.). "Do Corporate Codes of Ethics Reflect National Character? Evidence From Europe and the United States". Journal of International Business Studies, pp. 519-539.

Tréguer-Felten, G. (2009). Le leurre de l'anglais lingua Franca; Une étude comparative de documents professionnels produits en anglais par des locuteurs chinois, français et nord-américains. Thesis, University of Paris III.

Corpus mining tools in the PLEC project

Piotr Pęzik
University of Łódź
pezik@uni.lodz.pl

1 PLEC

The PELCRA Learner Corpus (PLEC) is a research project aimed at investigating the lexico-grammatical, phraseological and phonetic competence of Polish learners of English (Pęzik 2012). The project was launched in 2010 and it will run until late 2013. The corpus compiled in the project contains samples of learner English, such as essays, in-class and in-exam assignments, letters, MA theses and many other types written compositions authored by Poles using English as a foreign language (2.8 million words in total). It also features a time-aligned, error-annotated spoken subcorpus of learner English containing 200 000 word segments (which roughly corresponds to 25 hours of continuous recordings). These recordings are mostly informal interviews conducted with learners of English representing a variety of proficiency levels and social backgrounds.

2 Annotation

There are several tiers of corpus annotation used in the PLEC corpus. Texts and transcriptions of utterances are linked with author and speaker metadata, such as age, gender, education, proficiency level and language learning background. Linguistic annotation includes not only part-of-speech annotation, but also automatic syntactic dependency metadata. There are also two types of error annotation in PLEC. Firstly, a general, learner error taxonomy was adopted for the manual annotation of errors in a selection of the corpus. Secondly, the entire spoken component of the corpus has been annotated for word mispronunciations and used to compile an index of words commonly mispronounced by Polish learners of English (Zając & Pęzik 2012). Interestingly, the PLEC corpus also contains an automatic phraseological annotation tier. Using the HASK collocation dictionary (pelcra.pl/hask_en) (Pęzik 2013) a BNC-based collocation tagger was developed to identify and annotate instances of selected native-like phraseological units in the learner data. This type of annotation made it possible to estimate the so-called *phraseological index* of learner English

samples, which is a measure of native-like idiomaticity in non-native texts.

3 Tools

The different types of annotation available in the PLEC corpus can be mined using the online corpus mining tools developed within the project. The linguistic tier of the corpus annotation can be explored through a dedicated search engine supporting complex part-of-speech queries and a syntactic dependency browser. The search engine was implemented using a customized version of the Apache Lucene library and it can be used in other corpus projects as it scales well with the size of the collection up to billions of segments. The online interface for the corpus provides multimodal access to the written and spoken data components; users can stream audio snippets for utterances matching their queries. Learner and text profiles such as proficiency levels, domains, genres and register can be used as search criteria. Error and phraseological annotation tiers can be explored through dedicated online tools.

4 Availablity

The online corpus tools described in this abstract are available at http://pelcra/PLEC. The spoken time-aligned component of PLEC has been released under a Creative-Commons license and it can be used in further research on spoken English learning.

Acknowledgments

This research was funded in 2010-2013 by a grant from the Polish Ministry of Science and Higher Education.

References

Pęzik, Piotr. 2012. "Towards the PELCRA Learner English Corpus." In Corpus Data Across Languages and Disciplines, ed. Piotr Pęzik, 28:33–42. Łódź Studies in Language. Peter Lang.

Pęzik, Piotr. 2013. *Forthcoming*. Graph-based analysis of native and learner phraseology in HASK Collocation Dictionaries. pelcra.pl/hask_en.

Zając, Magda, and Piotr Pęzik. 2012. "Annotating pronunciation errors in the PLEC spoken learner corpus." In Proceedings ot TALC 10 Conference. Warsaw.

Applying corpus techniques to climate change blogs

Andrew Salway, Knut Hofland
Uni Research, Bergen
andrew.salway,knut.hofland@uni.no

Samia Touileb
University of Bergen
samia.touileb@gmail.com

1 Introduction

The emergence of social media has created new opportunities for social scientists to investigate how organisations, individuals and media contribute to shaping public understanding of, and opinions about, important issues. The work described here is about the application of corpus techniques to support investigations of complex and large-scale discourses in online social networks, e.g. blogs about climate change. These techniques are at the interface of corpus linguistics, text mining and information extraction.

2 Background

In the NTAP project[1] we are developing tools and methods for analysing the distribution, flow and development of knowledge and opinions across online social networks.

One innovation is that we treat the content of texts (blog posts) as key statements, rather than keywords; this contrasts with current tools for social media analysis that use word clouds and graphs of keywords over time. Our system to identify key statements, e.g. "climate change causes rising sea levels", depends on the identification of information structures, e.g. "X causes Y". Corpus techniques are crucial for us to ensure that the structures characterise how information about core concepts is typically expressed in the given domain and text type. As well as supporting statement extraction, the distribution of information structures within corpora will be analysed in investigations of discourse style.

In a blog corpus, each blog post can have metadata including 'time-date', 'author' and 'in/out hyperlinks'; there are also blogroll links between blogs. Our second innovation is to fuse the analysis of text features and network features,

[1] NTAP: Networks of Texts and People, funded by the Norwegian Research Council, 2012-15, http://www.ntap.no

through interactive visualizations. Given instances of a key statement in multiple posts, along with instances of supporting and opposing statements, we can explore the occurrence of these statements over the network of blogs, and over time. Thus we aim to give social scientists new affordances for understanding what influences the diffusion of key statements; for more on this, see (Salway et al. 2012).

3 Progress to date

We are currently crawling the blogosphere and have so far downloaded the complete content of about 3,000 English-language blogs that include posts mentioning "climate change" or a related term. The crawl started from about 20 handpicked blogs: the terms to determine topic relevance were extracted from some of these, using a keyword list and word clusters. Later, once we have a set of key statements relating to climate change, we plan to use a blog search engine to locate further material. Effort is now focussed on transforming raw html into a database of blog posts with associated metadata (author, date, hyperlinks, etc.).

In parallel, work has begun on the identification of information structures that characterise how information about climate change is expressed in blogs. Using, for now, the Yahoo BOSS blog search engine, we are exploring both "top-down" and "data-driven" methods.

Working in a top-down manner, we assumed that statements about the causes and effects of climate change would be important. A linguistic account of how causality can be expressed in English was used to generate a set of 14,000 queries: "causes climate change", "is the result of climate change", etc. We assumed that the number of hits per query is a crude indication of the information structures that are used commonly. Further, taking the snippets returned by the search engine, we collated the text around the query phrases, i.e. where we expect to find the stated causes and effects. Keyword and n-gram analysis of this text seems promising for elucidating further domain concepts, like "rising sea levels", "greenhouse gas emissions, that represent typically stated effects and causes of climate change.

With a top-down method there is a danger of never knowing what information structures are missed. Thus, we wish to work also in a data-driven manner and induce information structures directly from texts. As a step in this direction, we obtained 3352 non-overlapping n-grams (n<=8) containing "climate change", by iteratively querying the blog search engine: all n-grams contained the seed "climate change", and returned 10 or more hits. We are currently running automatic grammar induction techniques over such sets of n-grams, as a way to condense the information about the linguistic context around the term "climate change". Early results with ADIOS (Solan et al. 2005) and ABL (van Zaanen 2001) are promising, i.e. they generate word groups like "G1 = (cope, deal), G2 = (adverse, devastating, negative)" and combinations like "to G1 with the G2 impacts of climate change". Moving forwards, we plan to integrate these techniques with interactive visualizations, and human input, into a method for identifying information structures.

Acknowledgements

This research was supported by a grant from the Norwegian Research Council's VERDIKT program.

The generation of keyword lists, word clusters and n-grams was done using AntConc (Anthony 2011).

References

Anthony, L. (2011). AntConc (Version 3.4.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Salway, A., Diakopoulos, N. and Elgesem, D. 2012. "Visualizing Information Diffusion and Polarization with Key Statements". *Procs. SocMedVis 2012*.

Solan, Z., Horn, D., Ruppin, E. and Edelman, S. 2005. "Unsupervised Learning of Natural Languages", *PNAS* 102 (33): 11629-11634.

van Zaanen, M. 2001. "ABL: Alignment-Based Learning". *Procs. 18th Int. Conf. Computational Linguistics (COLING)*: 961-967.

Contrastive analysis of moves and steps taken in writing medical notes

Wenli Tsou
National Cheng Kung University
wtsou@mail.ncku.edu.tw

Hui-Chuan Lu
National Cheng Kung University
huichuanlu1@gmail.com

Sheng-Yun Hung
National Cheng Kung University
yunbe0811@hotmail.com

1 Introduction

Globalization has drawn more and more attention in applied linguistics and become one of the main themes in language teaching including the field of English for Specific Purpose (ESP). By taking advantage of technology development in corpus linguistics, the researchers compiled the Corpus of English for Medical Purpose (CEMP), which consists of two sub-corpora, 50 notes taken by Taiwanese students and 50 notes taken by experts and published in medical journals. The creation of the corpus aims to discover the gap between the two types of writers and provide suggestions for future improvement in teaching English writing for medical purpose.

2 Methodology

The researchers compared the medical notes taken by students with those of experts to contrast the differences for further analysis. Corpus Tool of UAM (Universidad Autónoma de Madrid) was used to annotate the written notes with 4 main moves which are chief complaint, history of present illness, record of patient visits, inter-hospital patient transfer, and related sub-steps.

In total, 4,058 tokens were annotated according to the scheme including the related features of those four moves listed above. In addition, the researchers analyzed the moves and steps with significant difference (P value ≤ 0.05) focusing on the elements needed to be paid more attention and to be improved in the process of professional medical training.

3 Findings

The following results are presented. Firstly, for the sub-steps "life style" and "patient history", experts used past tense while students tended to use present tense. Past tense was always used by experts when they described the facts about the patients. A possible reason could be that Taiwanese students do not pay attention to tense usages corresponding to time changes.

Secondly, in the steps "treatment procedure" and "doctor's prescription and advice", experts used passive voice more often, while active voice was used more frequently in the step "doctor's initial observation and diagnosis". Nevertheless, students used different voices in all thee three sub-steps. Thus, the findings of the study conclude that passive voice should be learned for prescriptions and advices while active voice for describing doctor's initial observation and diagnosis.

Finally, the phrase "the patient" was often used by experts to initiate the sentences in the sub-steps such as "treatment result, other history, life style, record of past and current visit, doctor's initial observation and diagnosis". In the sub-step "physical examination", to describe vital signs, the adjective "normal" was used in the expert corpus while "stable" was used very often in the student corpus to describe the examination results according to the normal standard. In addition, although both experts and students used "showed" and "revealed" to describe the lab findings, experts used "are shown in table X" which was rarely found in the student corpus. This difference could be associated with less information in students' texts that was presented with tables.

For language teaching of EMP, the CEMP can provide authentic and empirical examples with contextual information based on selected linguistic features of moves and steps taken in professional doctors' notes. By offering a systematic and corpus-based learning list, medical students do not only learn the syntactic patterns as well as lexical choices of professional usages, but also improve their writing skills on the medical notes. Future work for the study is to add more data to CEMP in order to obtain more persuasive generalization based on more representative data. Furthermore, the research model presented in the current study can be extended to ESP fields other than EMP.

References

Biber, D., Conrad, S., & Reppen R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Busch-Lauer, I (1995). Abstracts in German medical journals: A linguistic analysis. *Information Processing & Management*, 31(5), 769-776.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-

based methodologies. *English for Specific Purposes*, 24(3), 321-332.

Granger, S., Petch-Tyson, S., & Hung, J. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

UAM CorpusTool. http://www.wagsoft.com/CorpusTool/features.html

Generic pronouns in Latvian student-composed essays in English: A comparison of the BNC (British National Corpus) and BCML (Balanced Corpus of Modern Latvian)

ZigridaVincela
University of Latvia
zigrida.vincela@lu.lv

1 Background

The use of the third person pronouns denoting gender-indefinite (henceforth *generic*) nouns differs in the Latvian and English languages.

The Latvian language has an inflectional category for gender; all generic nouns are classified in comprehensive masculine or feminine gender inflection use patterns. The nouns that refer individually to a broad mixed-sex human group are labelled by linguists (Meistare, 2008; Nītiņa, 2008) as false generic nouns. They follow the masculine gender inflection pattern and involve the use of generic *he (all its forms)*.

In the English language, there is no inflectional category for the gender of nouns, so the gender is signalled by the third person singular pronouns (*he, his, him, himself* for the male antecedents and *she her, hers, herself* for the female antecedents). As the English language does not have a sex-neutral third person singular pronoun, the generic *he* or pronouns such as *anybody* have been traditionally used in the cases when a noun refers individually to a broad mixed-sex group of humans. However, during the last four decades the generic *he* has been questioned because it is often perceived to be male and thus implies discriminatory pronoun use in favour of males. The research and discussion of generic *he* including the possible alternatives to this masculine generic pronoun have been in the focus of linguists research as well.

To eliminate the bias in texts of various genres, the generic *he* avoidance strategies have been discussed in contemporary grammars of the English language. For example, Biber et al. (1999) in the *Longman Grammar of Spoken and Written English*, present two generic *he* avoidance strategies: (1) The use of coordinated pronoun forms *he or she, his or her, he/she, his/her*; (2) The use of plural *they* forms. While Downing and Locke (2006), in *English Grammar: A University Course*, distinguish four alternative strategies

enumerated in the order of their preference: (1) The use of *they* with both singular and plural verb forms; (2) The disjunctive use of *he* or *she*; (3) The combination of *s/he*; (4) The use of *she* as unmarked form.

2 Why this study?

The afore proposed linguists' alternatives to the use of the generic *he* in the English language can cause a double challenge to Latvian students, firstly, due to the differences in generic *he* use in Latvian and English, and secondly, due to the variety of the generic *he* avoidance strategies in contemporary English. The actual use of the generic *he* and its avoidance strategies preferred by the Latvian students in their English essays have not been investigated. This study, therefore, employs corpus analysis techniques to fill in the gap. The aims of the study are as follows:

1 To detect the generic *he* (*his, him, himself*) and its avoidance frequency in the corpus of Latvian student-composed essays in English and reveal the avoidance strategies preferred by students.

2 To compare the generic *he* and its avoidance strategies in students' essays with the generic *he* and the respective avoidance strategies found in *BNC* by Adami (2009).

3 To compare the noun pronoun patterns in the students English essays with the Latvian generic *he* use in *BCML (Balanced corpus of modern Latvian)* to uncover the possible influence of Latvian noun/pronoun patterning on the generic *he* in the students' English essays.

3 Results

The analysed material, the students' essays (138, 481 words), has been selected from a corpus of student-composed texts compiled at the Department of English Studies, University of Latvia.

The selected essays are arranged in three groups: timed pre-essays (written before the students' participation in writing course), untimed essays (written as writing course assignments) and timed post-essays (written after the students' participation in the seminars of writing course). Two of the seminars that were devoted to the practice of comparison and/or contrast essay composition aimed also to focus the students on the generic *he* and its avoidance strategies in English. The students after exploring contemporary grammars on generic *he* avoidance, searched, extracted and discussed relevant examples from BNC texts. Then, they were encouraged to apply their findings about generic *he* and its avoidance strategies in their own comparison and/or contrast essays as well as in the rest of the upcoming assignment essays of the course.

In order to find out the generic *he* avoidance strategy changes in the students' pre-essays, untimed essays and post-essays, the following pronoun cases (using *AntConc*) have been searched, retrieved from the student composed essays and then counted as separate tokens: generic *he*, *s/he, he or she, he/she* and singular *they*.

The results at the current stage of the study are as follows:

1 The generic *he* is less frequently used (Table 1) by the students in their timed post-essays, whereas they are the most frequently used in the pre-essays. The students tend to prefer the generic *he* in their pre-essays with the antecedents that are used with the generic *he* in Latvian (e.g. *child*). The avoidance of the generic *he* is the most prominent in the timed post-essays that may be explained by the influence of the generic *he* use and avoidance strategy discussion in academic writing seminars.

Essays	Generic *he*	*he/she he or she*	*S/he*
Timed pre-essays 30, 400	121.55	91.98	0
Untimed essays 69.257	119.84	86.63	31.76
Timed post-essays 38, 784	87.66	183.06	12.89

Table1: Pronouns in students' essays (per 100, 000 words)

2 As to generic *he* avoidance strategies, the students have preferred the use of coordinated forms: *he or she*, *he/she*, *his/her*, *him/her*. From the range of these forms, the slashed versions prevail. The use of *s/he* has not been found in the timed pre-essays. This coordinated form is occasional in the untimed and timed post-essays. The use of *s/he* is also occasional in *BNC* (Adami 2009).

3 The generic *she* does not occur in students' essays. Moreover, the texts show only a few cases of singular *they* (an example from untimed essays *Every student tried to represent what they thought.)***.** The singular *they* is also the least frequent in Adami (2009) findings of *BNC*.

The generic *he* use and its emerging avoidance

calls for a further comparative, corpus-based research of these linguistic features in the texts of non-native students of English.

References

Adami, E. 2009."To each reader his, their or her pronoun. Prescribed, proscribed and disregarded uses of generic pronouns in English". In A. Renouf and A.Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments*. Amsterdam, New York: Rodopi.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman

The Balanced Corpus of Modern Latvian (2012) The Institute of Mathematics and Computer Science, University of Latvia. Available online at http://www.korpuss.lv

Downing, A., and Locke, F. 2006. *English Grammar: A University Course.* London and New York: Routledge.

Meistare, A. 2008."Use of Generic Nouns".*Letonika. Linguistics Articles I*.

Nītiņa, D. 2008. "Problems of Working on the Latvian Grammar".*Linguistica Lettica18.*

A critical exploration of the use of English general extenders in a corpus of Japanese learner speech at different levels of speaking proficiency

Tomoko Watanabe
University of Edinburgh
tomoko.watanabe@ed.ac.uk

This paper aims to provide critical explorations of Japanese learners' use of English general extenders (*and so on*, *or something* and *and stuff*) at different speaking proficiency levels in an English speaking test both in the quantitative and qualitative approaches of corpus linguistics.

General extenders are one of the frequently occurring forms of English vague language in spoken discourse (Biber et al. 1999; Carter et al. 2011). The use of general extenders by learners of English has also been argued; for instance, De Cock et al. (1998) have identified *and so on* to be overused in learners' spoken texts, and general extenders have been found as rare linguistic forms in spoken discourse by Japanese learners of English (Shirato and Stapleton 2007). In these studies, however, the individual-speaker level has not been taken into consideration. As has been pointed out in corpus linguistic studies by Harrington (2008) and Murphy (2011), the generalised quantitative result in a specific corpus as a whole data set is not enough to explain how representative the use of general extenders can be of individual speakers. This can be said to learner corpora, as Durrant and Schmitt (2009:168) say that 'taking each text as an individual case' could avoid misleading results. In order to see the representativeness and potential causes of the results, the study attempts an in-depth individual-level analysis across the speakers' speaking proficiency and contexts, more specifically, task types in the speaking test, in the learner corpus.

The spoken data occurring in the speaking test comes from the National Institute of Information and Communications Technology Japanese Learner English Corpus (Izumi et al. 2004). This study consists of three phases: frequently occurring forms of general extenders firstly in the whole data and then at each speaking proficiency level and context, typical patterns of the general extenders, and their typical functions.

In the first phase of the study, single-word and multi-word cluster lists are run for the whole data to identify frequently occurring general extender forms and to have a big picture of their occurrence

in the speaking test. Then the data is stratified into the examinees' speaking proficiency levels in order to see which forms are typical to each level. A fine-grained analysis considers their general frequency in relation to the number of examinees that use them and, with reference to the examinees who use them, their densities in each of their spoken texts. This level of analysis reveals, for instance, a typical occurrence of *and so on* at the lower intermediate levels but not at the very beginners' or higher levels. The stratification of the data is also made across contexts in the speaking test. It shows that, for instance, *or something (like that)*, which is typical to intermediate and higher levels, occur in the interview and description much more typically than in the narrative or role play. The in-depth analysis suggests that the speaking proficiency levels and contexts may be potentials that impact on the frequency of the general extenders.

The second phase of analysis moves to identifying typical linguistic patterns of the general extenders, focusing on co-occurring words with them, the number of exemplars for each of them, and their positions in the examinees' turn. Each of the analysis levels uncovers differences among the general extender forms and across the speaking proficiency levels and contexts; for instance, in the context of the interview, the lower the speaking proficiency level, the more *and so on* occurs at the turn-final position, while the higher their level, the more *or something (like that)* occur at the turn-internal position.

Backed up with the quantitative findings, the final phase of the study explores typical pragmatic features of each of the general extenders, with looking at the full set of concordance lines. It reveals each of their interpersonal and interactional functions in the speaking test, concurring with the literature which has argued their multifunctionality and context-dependency (Overstreet 1999; O'Keeffe 2004; Cheshire 2007); for example, *and so on* can function to make up for the examinees' linguistic awkwardness and *and stuff* can function to make what is said emphatic.

The study argues the dynamics of the use of general extenders across the examinees' speaking proficiency levels and contexts in the speaking test. In more general sense it holds the integrated approach to the data as a whole and individual-level for the exploration of general extenders in learner corpora.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Carter, R., McCarthy, M., Mark, G., and O'Keeffe, A. 2011. *English grammar today: an A-Z spoken and written grammar*. Cambridge: Cambridge University Press.

Cheshire, J. 2007. "Discourse variation, grammaticalisation and stuff like that". *Journal of Sociolinguistics 11*(2): 155–193.

De Cock, S., Granger, S., Leech, G., and McEnery, T. 1998. "An automated approach to the phrasicon of EFL learners". In S. Granger (ed.) *Learner English on computer*. London: Longman.

Durrant, P., and Schmitt, N. 2009. "To what extent do native and non-native writers make use of collocations?" *International Review of Applied Linguistics in Language Teaching 47*(2): 157–177.

Harrington, K. 2008. "Perpetuating differences? Corpus linguistics and the gendering of reported dialogue." In K. Harrington, L. Litosseliti, H. Sauntson, and J. Sunderland (eds.) *Gender and language research methodology*. Basingstoke: Palgrave Macmillan.

Izumi, E., Uchimoto, K., and Isahara, H. 2004. *Nihonjin 1200nin no eigo supiikingu koopasu [A spoken corpus of 1200 Japanese-speaking learners of English]: the NICT JLE Corpus*. Tokyo: Alc.

Murphy, B. 2011. "Gender identities and discourse". In G. Andersen and K. Aijmer (eds.) *Pragmatics of society*. Berlin/Boston: De Gruyter Mouton.

Overstreet, M. 1999. *Whales, candlelight, and stuff like that: general extenders in English discourse*. Oxford: Oxford University Press.

O'Keeffe, A. 2004. "'Like the wise virgins and all that jazz': using a corpus to examine vague categorisation and shared knowledge". In U. Connor and T. A. Upton (eds.) *Applied corpus linguistics: a multidimentional perspective*. Amsterdam: Rodopi.

Shirato, J., and Stapleton, P. 2007. Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: pedagogical implications arising from an empirical study in Japan. *Language Teaching Research 11*(4): 393–412.