

# Abstracts

Corpus Linguistics 2011  
Discourse and Corpus Linguistics

20-22 July 2011

International Convention Centre (ICC)  
Birmingham, England

Hosted by the Centre for Corpus Research



UNIVERSITY OF  
BIRMINGHAM



# Sponsors

The conference organizers would like to thank the following organizations for sponsoring CL2011:



**CAMBRIDGE**  
**UNIVERSITY PRESS**



**UNIVERSITY OF**  
**BIRMINGHAM**

# Prefatory Note

Abstracts are ordered alphabetically by surname of the first mentioned author and grouped by type: plenaries, full papers, pecha kuchas, colloquia, workshops, and posters.

# Plenaries

## Plenary 1

**Susan Hunston (University of Birmingham)**

'Flavours' of corpus linguistics: the case of evaluative language

This paper begins by asking whether an agenda set by discourse studies can be addressed using corpus linguistics. It takes as an example the case of evaluative language. This topic has been tackled in both discourse studies (exemplified in this paper by Martin and White's Appraisal theory) and corpus studies (exemplified here in its most extreme form by Sentiment Analysis, but including also investigations of stance, semantic prosody and local grammar). The paper then extends the discussion to more general issues of difference between 'text' and 'corpus' as sites of investigation, outlining some of the difficulties identified if a corpus is conceptualised as a very large text and arguing that a corpus, unlike a text, cannot be analysed.

## Plenary 2

**Paul Baker (Lancaster University)**

Discourse, news representations and Corpus Linguistics

Corpus linguistics methods are increasingly becoming popular in research which examines discourse, ideology and attitudes in naturally occurring texts. It is argued that the use of large numbers of texts along with automatic procedures like keywords can help to reduce researcher bias. Baker et al (2008) provided a nine stage model of corpus-assisted critical discourse analysis which advises alternating between various qualitative and quantitative techniques to produce and test new hypotheses. This talk illustrates and evaluates how this model worked in practice when used in a recent research project to examine the representation of Muslims in the British press. I then describe an inter-analyst consistency experiment, where 5 analysts were separately asked to analyse the same corpus (news articles about foreign doctors). I focus on explaining why the analysts achieved different/similar results, and which techniques appeared to be most 'productive' in terms of eliciting interesting findings.

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 and Society* 19(3): 273-306.

## Plenary 3

**Stefan Th. Gries (University of California, Santa Barbara)**

Quantitative and exploratory corpus approaches to registers and text types

Following Baker's (2006) set of definitions of *discourse*, in this plenary I will be concerned with discourse, a notion of discourse that involves "different types of language use or topics" and that is related to notions such as *genre*, *style*, and/or *text type*. With that orientation, I will cover the following three different aspects, which I consider relevant for a large number of corpus-based studies to discourse, but also to the other ways in which *discourse* is understood.

1. As many others before me, I will briefly address the dichotomy of qualitative vs. quantitative research that is so commonly made in discourse analysis, sociolinguistics, socio-cultural linguistics, to name but a few disciplines. Unlike many others before me, however, my way of addressing this dichotomy will not consist of the usual we-need-both-types-of-approaches-complement-each-other call to arms, but I will attempt to make the point that, in a very trivial sense, qualitative work *is* ultimately based on quantitative work and, thus, needs to take lessons from quantitative methods into consideration.
2. I will summarily discuss a variety of previous case studies with an eye to demonstrate the

potential of bottom-up approaches to register-like corpus parts. On the one hand, I will outline ways in which corpus divisions into registers as made by corpus compilers can be tested for their discriminatory power. On the other hand, I will discuss how the study of any linguistic phenomenon in corpora should feature bottom-up analytical steps in order to identify the most promising registers/text types for a particular linguistic phenomenon.

3. In an attempt to extend previous work in this area, I will then exemplify a new idea how such bottom-up approaches can be made more comprehensive. Rather than using only a particular phenomenon in question as a diagnostic for the corpus parts to distinguish, as mentioned above and argued for in previous work, I will explore an approach to identifying homogenous parts in corpora that can include more diagnostics, assign different weights to them depending on which diagnostics are considered (more) important, and use these diagnostics in a cluster-analytic approach (with various ways to follow up and ascertain the discriminatory power of the results).

To the extent possible, I will discuss how such methods reflect and underscore similarities between corpus linguistics in general, register/text type-based studies in corpus linguistics in particular, and psycholinguistic theories. Linguistic elements to be discussed include lexical items, various types of *n*-grams, grammatical patterns/constructions, and argument structure constructions, plus maybe more; data to be discussed are from the BNC, the BNC Baby, the ICE-GB, plus maybe more.

# Full Papers

**Kirsten Ackermann (Pearson), Douglas Biber (Northern Arizona University), and Bethany Gray (Northern Arizona University)**

An Academic Collocation List

This paper presents a frequency list of the most common and pedagogically relevant collocations in written academic English discourse, derived from the written Academic component of the Pearson International Corpus of Academic English (PICAЕ). The list can be used, for example, in lexicography, test item writing, and EAP material development.

PICAЕ is a corpus of over 37 million words, comprising a written component (32.5 million words) as well as a spoken one (4.6 million words). The corpus covers American, Australian, British, Canadian and New Zealand English. PICAЕ was designed with reference to the question, what English does a non-native speaker need in order to be successful in academic settings where English is the main language. Spoken data include lectures, seminars and broadcasts. Written data comprise textbooks and journal articles reflecting a broad range of academic disciplines, as well as university, student and alumni journals, and study and career information.

For this project only the written Academic component of the Pearson International Corpus of Academic English (PICAЕ) was used, which comprises over 25 million words covering 28 major academic subjects.

For the compilation of such a collocation list, the concordance program MonoConc was used to first obtain a simple list of words occurring more than 12 times in the corpus.

This list was processed by a computer program written in Pascal to index each word, and the sub-dimensions for each word using variables like frequency, number of texts that the word has occurred in, and frequencies in each of the four academic disciplines (humanities, social science, natural and formal science, professions and applied science).

The output of this program was a reference list of important academic words, which occurred at least 5 times per million words and in at least 5 different texts. A stop- list was created containing the frequent function words that express purely grammatical meaning. This list was used by the collocations and bundles programs written in Perl to exclude those words from subsequent analysis.

The collocation program itself used the reference list of important academic words as input, again using a large multi-dimensional hash or database table. As each potential collocate is located in the texts, the program added the entry to the data structure if it did not exist, or increment frequency and distributional counts if it already existed.

This paper will explain the motivation for the academic collocation project. It will shortly introduce the Pearson International Corpus of Academic English. The paper will then look at the methodology applied and problems encountered. Lastly it will discuss potential usages of the Academic Collocation List.

**Amal Alsaif and Katja Markert (both University of Leeds)**

Annotating Discourse Connectives in MSA: Disagreement Cases in the LADTB

Discourse relations such as CAUSAL or CONTRAST relations between textual units play an important role in producing a coherent discourse. They are widely studied in theoretical linguistics (Halliday and Hasan, 1976; Hobbs, 1985), where also different relation taxonomies have been derived (Hobbs,

1985; Knott and Sanders, 1998; Mann and Thompson, 1988; Marcu,2000). Discourse relations can be signalled by explicit lexical indicators, so-called discourse connectives (Marcu, 2000; Webber et al., 1999; Prasad et al..2008a). Our study is based on Leeds Arabic Discourse Treebank “ LADTB- a recent annotation effort of discourse connectives of MSA (Alsaif and Markert,2010). It provides a new annotation layer above the existing layers of annotation (syntax and morphology) in the Arabic news corpus Penn ATB, Part1 (Maamouri and Bies, 2004) by annotating all discourse connectives, the relations they signal and the two arguments they relate.

In the first such study for Arabic, 107 potential discourse connectives and 18 discourse relations were analyzed following similar annotation principles of Penn DTB project for English (Prasad et al..2008a); taking into account properties specific to Arabic. In particular, we deal with the fact that Arabic has a rich morphology: we therefore include clitics, prepositions and nouns as connectives as well as a wide range of nominalizations as potential arguments. A dedicated discourse annotation tool is developed for Arabic which is based on plain text; for unrestricted discourse annotation. Both the human identification of discourse connectives and the determination of the discourse relations they convey are reliable (Alsaif and Markert,2010). We measure also the inter-annotator agreement of identifying the text spans of the arguments individually in different ways (i) the exact match of textual units in the argument text (ii) the average of overlapping syntactic tree nodes of the two arguments and (iii) the match of syntactic heads of arguments. We show that although there is no high agreement on the exact textual units, annotators reliably agree on the syntactic heads a part from disagreements for some connectives at beginning of paragraphs. The syntactic head seem to be in the most cases the text expressing the core proposition in the discourse.

We report the disagreement and ambiguity cases in our human annotation in terms of identifying (i) discourse connectives,(ii) relations and (iii) related arguments. Our results show that Arabic has a higher ambiguity than in English; connectives in PDTB are almost unambiguous a part from few discourse connectives such as since, while. Moreover, Arabic discourse tends to use longer and more complex sentences with many complements than in English. Thus annotators have disagreed in the definite boundaries of the arguments. In addition, there is a common usage of a coordinating conjunction wa/and at beginning of each paragraph if not every sentences, particularly in the news writing, without any specific discourse function rather than conjunction. Defining variant disagreement cases would help in understanding the language features in a comparative study with other languages and improving further annotation studies.

A. Al-Saif, K. Markert.2010. ‘The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic’. In *International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman London.

J.R. Hobbs. 1985. *On the coherence and structure of discourse*. Center for the Study of Language and Information, Stanford, Calif.

E.H. Hovy. 1993. ‘Automated discourse generation using discourse structure relations’. *Artificial intelligence*, 63(1-2):341-385.

A. Knott and T. Sanders. 1998. ‘The classification of coherence relations and their linguistic markers: An exploration of two languages’. *Journal of Pragmatics*, 30(2):135-175.

M. Maamouri and A. Bies. 2004. ‘Developing an Arabic treebank: Methods, guidelines, procedures, and tools’. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based*

*Languages (COLING)*, Geneva.

W.C. Mann and S.A. Thompson. 1988. 'Rhetorical structure theory: Toward a functional theory of text organization'. *Text*, 8(3):243-281

D. Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008a. 'The Penn discourse treebank' 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

**Laurence Anthony (Waseda University, Japan)**

Introducing Corpus-Based Methods into a Large-Scale Technical Writing Program for Scientists and Engineers

This paper describes the introduction of corpus tools and methods into a large-scale technical writing program for undergraduate and graduate university students of science and engineering. This novel approach is shown to greatly improve students' understanding of discourse structure, increase their productive vocabulary knowledge, and provide them with useful skills that they can apply in writing tasks long after graduation.

The target university in the study operates the largest science and engineering English language program in Japan and possibly the world. Currently, the program serves approximately 6000 undergraduate students, 4000 graduate students, and also many post-doctorate researchers. Starting in 2004, elective courses in undergraduate- and graduate-level technical writing were radically overhauled to meet the growing challenges of students in a globalized society. This led to a rapid increase in the number of students from diverse disciplines and with diverse ability levels, and also an increase in the number of part-time instructors who did not have a background in science and engineering. To deal with the challenges of a more diverse student population and a greater number of non-expert part-time faculty, a decision was made to introduce corpus-based methods into the classroom. This would allow instructors without backgrounds in science and engineering to offer meaningful classes to their students, and more importantly, would give students useful analytical skills that they could apply throughout their professional careers.

In the presentation, I will first explain some of the key design choices that have led to a successful implementation of the new program, including 1) the selection of suitable tools and corpora that can be easily used by both teachers and students inside and outside the classroom, 2) the development of a printed textbook with step-by-step exercises and sample texts, and 3) the introduction of teacher training sessions and teacher feedback systems. Next, I will describe some of the challenges that full-time faculty in the program have faced when coordinating the program and introducing concepts from corpus linguistics to non-expert part-time faculty. Then, I will report on quantitative and qualitative measures of improvement in student writing, before discussing the teacher and student feedback on the program. Finally, I will suggest ways in which the technical writing program can be implemented and subsequently improved at other institutions around the world.

**Laurence Anthony (Waseda University, Japan), Kiyomi Chujo (Nihon University, Japan), and Kathryn Oghigian (Waseda University, Japan)**

A Freeware, Open-Source, Web-Based Framework for Distribution and Analysis of Single and Parallel Corpora

In recent years, an increasing number of corpora have been made available to corpus researchers. Many of these are released as raw texts that can work with standalone concordancers such as WordSmith Tools (Scott, 2010), ParaConc (Barlow, 2010), and AntConc (Anthony, 2010). However, more and more corpora are being released through dedicated Web-based interfaces. This trend is

partly due to the convenience of an 'anytime-anywhere' environment and also due to the fast processing offered by server-based programs. A further and perhaps more important reason for the growth in Web-based tools is that they allow researchers to avoid restrictions associated with distributing copyrighted materials (Hemming and Lassi, 2010).

One major problem with releasing a corpus via a Web-based interface is the effort required to develop, test, and manage the server-side analysis tool. Each new project requires the corpus researcher to consider the corpus database architecture, the design of the interface, the development programming language, and ultimately the difficult task of coding itself. Some corpus projects are fortunate to have programmers and designers as part of the team, but this usually results in a long development time frame. More commonly, project members are forced to outsource the tool development to commercial software developers. In this case, the costs can become very high and the resulting tool inflexible.

In this paper, we introduce a freeware, open-source Web-based framework that allows corpus researchers to easily release their single or parallel corpora to the wider field with only minimal knowledge of servers, databases architectures, and programming languages. In essence, researchers download a setup file from a Website and drag and drop this into a folder on a standard hosting server. Then, after launching the setup file in a browser, a script proceeds to index the raw texts and setup the system for users. There are many advantages to such an open-source platform including: 1) corpus researchers no longer have to 'reinvent the wheel' creating a Web interface each time a new project is started, 2) the framework reduces development time and costs allowing researchers to focus their efforts on developing better corpora, 3) the framework gives complete control back to the corpus researchers, allowing them to tweak the system to their own needs, and 4) the corpus community can work together to improve the framework by adding different database architectures, attractive interface skins, and new functionality.

In the presentation, we will introduce a pilot version of the framework and show how it has already been used to create an effective environment for novice teachers and students at two university institutions, where they interact with single corpora of science and engineering texts, and parallel corpora of newspaper texts.

Anthony, L. (2010). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.antlab.sci.waseda.ac.jp/>

Barlow, M. (2010). MonoConc Pro (Version 2.2) [Computer Software]. Houston, US: Athelstan. Available from <http://athel.com/>

Hemming, C. and Lassi, M. (2010). Copyright and the Web as Corpus. Retrieved September 21, 2010, from <http://hemming.se/gslt/copyrightHemmingLassi.pdf>

Scott, M. (2010). WordSmith Tools (Version 5) [Computer Software]. Liverpool, UK: Lexical Analysis Software. Available from <http://www.lexically.net/wordsmith/version5/index.html>

#### **Antti Arppe (University of Helsinki & University of Alberta)**

##### From modeling lexical synonyms to constructional alternations

Recently, large-scale corpus-based studies (Arppe 2008 and Divjak 2010) have explored the phenomenon of near-synonymy exploiting a battery of univariate and multivariate statistical techniques, based on a wide range of contextual linguistic features at the morphological, syntactic and semantic level. Noteworthy is that this research has demonstrated an extension of the application of statistical methods from dichotomous linguistic settings to polytomous ones, i.e.,

concerning more than two possible alternative outcomes with a similar meaning; however, these studies have focused on the lexical level of semantic similarity.

Although corpus-based work has also been conducted on the syntactic level concerning constructional alternations (alternatively synonymous structural variants, see Biber et al. 1998), e.g. concerning the English possessive constructions (Gries 2002 and Rosenbach 2003), the English verb-particle placement (Gries 2003a), and the English dative alternation (Gries 2003b and Bresnan et al. 2007), such work has predominantly been restricted to dichotomous alternatives. Nevertheless, it is clearly evident in general grammatical descriptions, e.g. Biber et al. (1999), that there are often more than two possible constructional alternatives, which clearly motivates a shift of interest in also constructional studies from pairs to sets with more than two alternative members.

This paper demonstrates the application of multivariate statistical analysis to the English (1) active vs. (2) be-passive vs. (3) get-passive alternation (see e.g. Biber et al. 1999), focusing on those verbs which occur in all three alternative syntactic constructions. In addition to explanatory variables based on current literature, i.e. register, stative vs. dynamic distinction, long vs. short form, length of subject/agent phrase, and preferences of the node verb itself, the underlying linguistic analysis also incorporates the scrutiny of the morphological and syntactic structure as well as semantic subcharacterizations of the context associated with the verbs in question.

In the statistical analysis, the application and results of both (1) polytomous logistic regression (see e.g. Arppe 2008) as well as its novel extension to corresponding (2) polytomous mixed-effects logistic regression modeling are demonstrated. The key benefit of these two statistical methods is that they allow us to (1) estimate the relative weights of the linguistic explanatory variables in natural terms as odds, as well as to (2) model the impact of their joint occurrence in various combinations as expected probability distributions for the alternative constructions; moreover, with the latter model we can (3) directly incorporate the effect of extralinguistic factors such as individual speaker/writer preferences.

The results support a probabilistic view of the relationship between linguistic usage and the underlying linguistic system, in which only a minority of linguistic choices are categorical - instead, most contexts exhibit degrees of variation as to their outcomes, resulting in proportionate choices over longer stretches of usage in texts or speech.

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.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Bresnan, J., A. Cueni, T. Nikitina, and R. H. Baayen 2007. 'Predicting the Dative Alternation'. In: *Cognitive Foundations of Interpretation*. Boume, G., I. Kraemer, and J. Zwarts. Amsterdam: Royal Netherlands Academy of Science, pp. 69-94.

Divjak, D. 2010. *Structuring the lexicon: a clustered model for near-synonymy*. Berlin: Mouton de Gruyter. Cognitive Linguistics Research.

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*. London: Continuum.

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.

#### **Ingo Bachmann (University of Duisburg-Essen)**

How do we learn what a successful gay and lesbian relationship is (and needs)? – The discursive construction of gay and lesbian relationships in dating and relationship guides

In Western societies there has been a change in gay and lesbian politics from identity issues to relationship acceptance and recognition (cf. Weeks 2000: 213). The focus in linguistic analyses of gay and lesbian discourses has been mostly centred on identity and community construction (e.g. Baker 2005; Koller 2008), which needs to be complemented with a view on how relationships are constructed. This paper aims to shed more light on the way same-sex relationships in contemporary society are discursively formed, setting them apart from or putting them on a level with heterosexual couples.

Dating and relationship guides prove to be a fruitful source for that endeavour. They tell the reader how to meet and recognise an appropriate partner, how to turn a date into a lasting relationship and how to lead and maintain a good and healthy relationship. Three corpora containing online dating and relationship guides/advice for gay men, lesbians and a heterosexual readership have been compiled and semantically tagged, using the software Wmatrix. A comparison of the tagged corpora results in key semantic concepts for each corpus. The automatic semantic tagging is completed with a manual categorisation of keywords into semantic sets (e.g. words referring to aspects of time in building relationships or expressions denoting emotional or sexual satisfaction). These emerging semantic sets serve as a starting point for a more detailed collocational and contextual analysis.

What can be demonstrated, among other things, is that depending on the sexual identity of the readership, the focus of the guides is different. This difference underlies the constructed nature of the need for diverse advice for gay, lesbian and heterosexual people. It indicates that gay, lesbian and heterosexual dating behaviour and relationships are commonly regarded as facing different issues (e.g. how to find Mr or Mrs Right, how to turn a fling into a lasting relationship or which role to play in a relationship).

Baker, P. 2005. *Public Discourses of Gay Men*. Abingdon & New York: Routledge.

Koller, V. 2008. *Lesbian Discourses: Images of a Community*. London: Routledge.

Weeks, J. 2000. *Making Sexual History*. Cambridge: Polity Press.

**Minhee Bang (Sangmyung University) and Seoin Shin (Hallym University)**

A corpus study of the use of English loan words in Korean: the case of noksaek, grin and green

This paper is part of a larger study of the use of English loan words in Korean language. While there are loan words which have no Korean equivalent such as Internet, smartphone, twitter, and therefore are phonetically translated into Korean, increasingly, loan words which do have Korean equivalents are phonetically translated into Korean, entering everyday language use. These English loan words co-exist along with their Korean counterparts. The issue of what motivates people to use loan words over their mother tongue counterparts may traditionally be a sociolinguistic inquiry. One much discussed motivating factor is the status of English as a privileged language; the use of English is seen as a status symbol. What is aimed in this study is to take a quantitative, bottom-up approach in investigating what other factors may be involved in when phonetically translated English loan words are preferred over their Korean equivalents. The investigation will involve three way analysis of the use of a phonetically translated English loan word, its Korean equivalent, and its English counterpart, for example, grin, noksaek, and green, whose results are presented in this paper. The analysis will focus on the use of the words in the environmental context. We hope that the use of corpus data and methods can shed light on the following questions:

1. Are there any collocational patterns associated uniquely with the phonetically translated English loan word (grin) and its Korean counterpart (noksaek)?
2. Are there any contextual patterns associated uniquely with the phonetically translated English loan word (grin) and its Korean counterpart (noksaek)?
3. Is there any kind of transfer from English to Korean in the use of the phonetically translated English loan word (grin) and its Korean counterpart (noksaek)?

The study will be expanded on three other sets of words: mihonmo/ single mom, jigoochon(yi)/ global, kekuhada/ clean. Asking these questions may provide an answer to in what context phonetically translated loan words are preferred over their Korean equivalents and what motivation lies in the preference. Furthermore, it will be interesting to see how much fine-tuned textual evidence can be obtained through the use of corpus methods. For analysis, a corpus of Korean newspapers collected from the KINDS, a Korean newspaper database and various texts collected from the three popular Korean search engines Naver, Nate, and Daum is used. The Bank of English corpus and the Corpus of Contemporary American English are consulted for comparison.

**Sabine Bartsch (Technische Universität Darmstadt)**

Untypical animacy: a historical study of subjects in science writing

Subjects of the type exemplified in the following examples are a pervasive feature of contemporary English science writing:

- (1) Three factors explain the under-representation of women in editorial boards ....
- (2) The data show nearly all the features observed ....
- (3) PC1 mostly explains differences in the NIR spectra related to O–H vibrations (water content).
- (4) The tangential contact force considers additionally the friction at the contact between the particles.

The choice of subject in these sentences breaches with conventional expectations concerning the subject of verbs of saying and cognition in the active voice which are commonly expected to occur with an animate agent as subject. This violation of semantic constraints on the type of subject

expected by default with certain semantic verb classes (e.g. *verba dicendi*, *verba cogitandi*) suggests an interpretation as instances of the phenomenon Margaret Berry (1975) has termed „untypical animacy“. Master (1991) describes this phenomenon as a “subject-verb mismatch” and Banks (1996) discusses it as a case of metaphor. The use of inanimate subjects with verbs regularly requiring animate agent subjects is a frequent feature of contemporary science writing and is hypothesized to be associated with the avoidance of an overt mentioning of the human scientist as agent which is also associated with a preference for other features such as the use of agentless passives. This contributes to conveying a stance of scientific detachment in the interest of scientific objectivity.

In contrast to this feature of contemporary science writing, science writing in the 17th and 18th century is found to display a much lower frequency of this phenomenon but displays a comparatively higher frequency of first person subjects as has been shown in previous studies (e.g. Banks 2008). This is commonly associated with the different standing of the individual natural philosopher as a trusted authority in scientific investigation as well as with the epistolary style of a substantial proportion of early science writing (cf. Atkinson 1999). These observations raise the issue of tracing changing argument structure patterns in the history of modern science writing.

This paper presents a corpus study of semantic roles realizing the subject in active voice uses of verbs of saying and verbs of cognition in science writing. The study is based on a small corpus of scientific articles from different disciplines since the late 17th century. The paper presents results of a qualitative and quantitative analysis of the distribution of different semantic realizations of subject role of these classes of verbs and seeks to interpret its observations in the light of the development of science writing to-date.

Atkinson, D. (1999): ‘Scientific Discourse in Sociohistorical Context’. *The Philosophical Transactions of the Royal Society of London, 1675 – 1975*. MahWah, New Jersey, London: Lawrence Erlbaum Associates, Publishers.

Banks, D. (1996): ‘The Passive and Metaphor in Scientific Writing’. *Cuadernos de Filología Inglesa*, 5/2, 1996, pp. 13 - 22.

Banks, D. (2008): *The Development of Scientific Writing: Linguistic Features and Historical Context*. London, Oakville: Equinox.

Berry, M. (1975): *Introduction to Systemic Linguistics. Vol. 1, Structures and Systems*. London: Batsford.

Levin, B. (1993): *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago, London: The University of Chicago Press.

Master, P. (1991): ‘Active Verbs With Inanimate Subjects in Scientific Prose’, *English for Specific Purposes*. Vol. 10, pp. 15-33.

**Sabine Bartsch (Technische Universität Darmstadt), Elke Teich (Universität des Saarlandes), and Christoph Tragl (Technische Universität Darmstadt)**

Patterns of cohesion in informationally dense texts

Scientific writing is often perceived to deliver a large amount of information in a compact manner. This impression is commonly attributed to features at the level of lexico-grammar, such as domain-specific terminology, complex nominal groups, and a high lexical density. However, there are features at the level of text that contribute to (high) information density, too. One such feature is cohesion, i.e. the choices that make a text hang together in terms of reference, ellipsis, substitution,

lexical cohesion, and conjunction (cf. Halliday & Hasan, 1976).

In this paper we examine patterns of cohesion in scientific abstracts. Abstracts are chosen because they present information in an extremely aggregated manner, thus exhibiting a particularly high information density (cf. Swales 1990; Biber 2006). The corpus of abstracts under study is a subcorpus of the Darmstadt Scientific Text Corpus (DaSciTex) (Teich & Holtz 2009) consisting of around 2.000 texts from nine scientific disciplines (17 million words). We have investigated abstracts from four disciplines included in DaSciTex (computer science, linguistics, biology and mechanical engineering), taking samples of ten abstracts per discipline. The corpus is automatically pre-annotated by means of the tool Little Cohesion Helper (LCH) (Bartsch et al., 2009), which was developed for the purpose of cohesion annotation. LCH automatically identifies lexical cohesive chains within a text on the basis of lexical semantic relations represented in the Princeton WordNet (Fellbaum et al. 1998). Since automatic annotation does not achieve a hundred percent accuracy, the annotation needs to be manually corrected. The other types of cohesion (reference, ellipsis, substitution and conjunction) have been manually annotated by means of MMAX2 (Müller & Strube 2006). Since the output of LCH is mapped into the XML-format used by MMAX2, the integration of annotations is straightforward.

The aim of our study is twofold. First, we are interested in possible differences and commonalities in the usage of cohesive devices across disciplines (types of cohesion, density and length of cohesive chains). The second aim is to investigate whether there are any differences in patterns of cohesion in scientific abstracts vs. other, less technical texts. To this end, we have carried out a comparative study using parts of the FLOB corpus. In terms of the usage of cohesive devices, we expect that abstracts exhibit a relatively frequent use of lexical cohesion, rather little use of reference and rather few occurrences of cohesive conjunction. In the case of relative uniformity of cohesive patterns across disciplines and relative distinctness to the texts taken from FLOB, we can conclude that the abstract constitutes a text type/genre on its own (independent of register); in the case of relative diversity across disciplines, we conclude that we encounter a case of register (i.e. domain-specific) variation (with no discrete text type/genre). In the paper we present the results of both studies and their interpretation in terms of register vs. genre attribution.

Bartsch, Sabine et al. (2009): "ObamaSpeeches.com. Building and Processing a Corpus of Political Speeches. A Student Project." Poster presentation at the Herbsttagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL) 2009 an der Universität Potsdam, September 2009.

Biber, Douglas. (2006): *University language : a corpus-based study of spoken and written registers*. Amsterdam, Philadelphia: John Benjamins.

Fellbaum, Christiane (1998, ed.): *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Halliday, MAK & Ruqaiya Hasan (1976): *Cohesion in English*. Harlow: Longman.

Müller, Christoph & Michael Strube (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 ).

Swales, John. (1990): *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Teich, Elke & Mônica Holtz. (2009): 'Scientific registers in contact: An exploration of the lexicogrammatical properties of interdisciplinary discourses'. *International Journal of Corpus Linguistics* 14(4), 524-548.

**Monika Bednarek (University of Sydney)**

'Read less, more TV': A corpus linguistic perspective on television discourse

This paper engages with data from one of the most popular types of texts that we encounter: the discourse of fictional television. With a few recent exceptions (e.g. Rey 2001, Mittmann 2006, Quaglio 2008, 2009, Bednarek 2010), corpus linguistics has so far not attended to the language of television series. This is despite the fact that in many countries we spend more hours watching television than consuming other media. In Australia, for instance, the average time spent watching television is more than 3 hours every day (Dick 2010). Other reasons for studying television dialogue include its influence on viewer's language (Mittmann 2006) and its function as 'transmodern teacher' (Hartley 1999), with fictional television series offering models to viewers of how things are done (Wodak 2009). Indeed, the linguistic study of television discourse can be seen as an important emerging area of research, with current studies (e.g. Baker 2005, Bubel 2006, Wodak 2009, Richardson 2010, Toolan in press) drawing on existing research in (critical) discourse analysis, conversation analysis, pragmatics, sociolinguistics, and stylistics. In this paper I aim to show how we can approach the study of television dialogue from a corpus linguistic perspective. To do so I introduce short case studies on various television series drawing on my own and others' research. These case studies focus respectively on the following areas of research:

- The difference between television dialogue and unscripted language (e.g. research on *Gilmore Girls*, *Golden Girls*, *Friends*, *Dawson's Creek*);
- The linguistic construal of characterisation (e.g. research on *The Big Bang Theory*, *Lost*);
- The construal of mainstream ideologies (e.g. research on *Gilmore Girls*, *The West Wing*);
- The construal of gender and sexuality (e.g. research on *Will & Grace*, *Star Trek*)

Rather than describing each case study in detail, my paper focuses on how these studies demonstrate the value of a corpus linguistic exploration of television dialogue. At the same time, I will argue that the study of television discourse is best approached from different perspectives and profits from bringing together both qualitative and quantitative approaches. A final point I will make concerns the advantages of drawing on such research in the teaching of linguistics in general and in teaching corpus linguistics in particular.

Baker, P. 2005. *Public Discourses of Gay Men*. London/New York: Routledge.

Bednarek, M. 2010. *The Language of Fictional Television. Drama and Identity*. London/New York: Continuum.

Bubel, C. 2006. *The Linguistic Construction of Character Relations in TV Drama: Doing Friendship in Sex and the City*. Unpublished PhD dissertation, Universität des Saarlandes, Saarbrücken, Germany. Available at <http://scidok.sulb.uni-saarland.de/volltexte/2006/598/>

Dick, T. 2010. Behold the new golden age of TV. *The Sydney Morning Herald News Review*, 23-24 October 2010: 6-7.

Hartley, J. 1999. *Uses of Television*. London/New York: Routledge.

Mittmann, B. 2006. 'With a little help from Friends (and others): Lexico-pragmatic characteristics of original and dubbed film dialogue'. In Houswitschka, C. Knappe, G. & A. Müller (eds). *Anglistentag 2005, Bamberg – Proceedings*. Trier: WVT. 573-585.

Quaglio, P. 2008 'Television dialogue and natural conversation: Linguistic similarities and functional differences'. In Ädel, A. & R. Reppen (eds). *Corpora and Discourse. The Challenges of Different Settings*. Amsterdam/Philadelphia: John Benjamins. 189-210.

Quaglio, P. 2009. *Television Dialogue. The Sitcom Friends vs. Natural Conversation*. Amsterdam/Philadelphia: John Benjamins.

Rey, J. M. 2001 'Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993'. In Biber, D. & S. Conrad (eds). *Variation in English: Multi-dimensional studies*. London: Longman.138-156.

Richardson, K. 2010 *Television Dramatic Dialogue. A Sociolinguistic Study*. Oxford: Oxford University Press.

Toolan, Michael. in press. 'I don't know what they're saying half the time, but I'm hooked on the series': incomprehensible dialogue and integrated multimodal characterisation in *The Wire*. In Piazza, R., Bednarek, M. & F. Rossi (eds). *Telecinematic Discourse: Approaches to Language in Fictional Cinema and Television*. Amsterdam/Philadelphia: John Benjamins.

Wodak, R. 2009. *The Discourse of Politics in Action*. Basingstoke/New York: Palgrave Macmillan.

**Anke Beger (Flensburg University)**

Complementing discourse analysis with computer-assisted corpus analysis in the field of specialized languages: The case of ANGER metaphors in academic discourse

Applied research within the framework of Cognitive Metaphor Theory (e.g. Lakoff & Johnson 1980; Lakoff 1993) mainly consists of two different methodological approaches. Apart from a few exceptions (e.g. Cameron & Deignan 2003, 2006), the examined data comprises either small corpora which allow a more detailed and discourse-oriented analysis (e.g. Low et al. 2008; Cameron 2003, 2007), or large corpora that are analyzed with special computer software (e.g. Deignan 1999; Koller 2002). However, in the field of specialized languages, a combination of these two methods is particularly valuable in order to compare metaphor use in certain genre with the occurrence of metaphors across genres. Thus, we analyze the application of ANGER metaphors in academia by contrasting them with the occurrence of ANGER metaphors in a large computerized corpus.

For our investigation, we compiled a corpus containing approximately 32,000 words. The data consists of four psychology classes about anger and aggression that were given at an American college. This corpus is small enough to be handsearched for metaphorical expressions and analyzed with regard to discourse features. To compare the application of ANGER metaphors in academia with the occurrence of metaphorical expressions for ANGER across a variety of genres, we searched the Corpus of Contemporary American English (COCA), containing more than 410 million words, with the KWIC concordance program.

The results show that almost all ANGER metaphors in our handsearched corpus also occur across registers in the COCA. An exception is the conceptual metaphor ANGER IS A FLUID IN A CONTAINER. It is almost identical to the well-known metaphor ANGER IS A HOT FLUID IN A CONTAINER (cf.

Kövecses 2000: 21) which is realized in metaphorical expressions in the COCA. Yet, metaphorical expressions instantiating ANGER IS A FLUID IN A CONTAINER in the academic corpus do not feature the concept of HEAT, but rather communicate a “hydraulic” model of ANGER. Thus, ANGER IS A FLUID IN A CONTAINER is a genre-specific metaphor. Additionally, the metaphors which occur in both corpora differ in the relative frequency in which they occur. While ANGER IS INSANITY is one of the most frequent ANGER metaphors in the COCA, it is one of the least frequently applied ones in our academic corpus.

In complementing a detailed discourse-based examination of ANGER metaphors in a small dataset of academic discourse with a computer-assisted analysis of a large corpus comprising different genres, we are able to point out the particularities of the use of ANGER metaphors in a certain genre. This attests the value of combining different corpora and methods to research the use of metaphors in specialized languages.

Cameron, Lynne (2003) *Metaphor in Educational Discourse*. London/New York: Continuum.

Cameron, Lynne (2007) “Patterns of Metaphor Use in Reconciliation Talk”, in: *Discourse & Society* 18 (2), 197-222.

Cameron, Lynne & Deignan, Alice (2003) “Combining Large and Small Corpora to Investigate Tuning Devices Around Metaphor in Spoken Discourse”, in: *Metaphor and Symbol* 18 (3), 149-160.

Cameron Lynne & Deignan, Alice (2006) “The Emergence of Metaphor in Discourse”, in: *Applied Linguistics* 27 (4), 671-690.

Deignan, Alice (1999) “Linguistic Metaphors and Collocation in Non-literary Corpus Data”, in: *Metaphor and Symbol*

Koller, Veronica (2002) “‘A shotgun wedding’: Co-occurrence of War and Marriage Metaphors in Mergers and Acquisitions Discourse”, in: *Metaphor and Symbol* 17, 179-204.

Kövecses, Zoltán (2000) *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge: Cambridge University Press.

Lakoff, George (1993) “The Contemporary Theory of Metaphor”, in: Ortony, Andrew (eds.) (1993) *Metaphor and Thought*. Cambridge: Cambridge University Press, 202-251.

Lakoff, George & Johnson, Mark (1980) *Metaphors We Live By*. Chicago/London: The University of Chicago Press.

Low, Graham; Littlemore, Jeanette & Koester, Almut (2008) “Metaphor Use in Three UK University Lectures”, in: *Applied Linguistics* 29 (3), 428-455.

**Miguel-Angel Benitez-Castro (Department of English, University of Jaen, Spain), and Jesus Fernandez-Dominguez (Department of Spanish, University of Jaen, Spain)**

Transitivity in a learner corpus, or on how students’ experiences are shaped by their semantic choices

Over the past twenty years, the study of English learner writing has experienced considerable growth in the field of applied linguistics (Polio 2003: 35, 40-41), but most such investigations have been limited to a purely formal analysis, as witnessed by the wealth of research on error analysis (Corder 1967; Richards 1980; Carl 1998; Ellis and Barkhuizen 2005). To achieve complete

understanding of learners' interlanguage development, we should also consider the communicative potential of their written production, which is enabled by Halliday's Systemic Functional Grammar (Halliday 1994; henceforth, SFL). SFL's metafunctions (i.e. ideational, interpersonal and textual), formally encoded through the systems of transitivity, mood and theme, allow for a description of language that draws on experience, stance and textual organisation.

Among the few studies where Halliday's transitivity is explored, emphasis is laid on professional writing, as in Martínez (2001) or Melrose (2003). Research into learner writing has also referred to the ideational metafunction, but often either in conjunction with the other metafunctions (Ivanic and Camps 2001) or as secondary to other key topics (Chen and Foley 2004). Be that as it may, an in-depth analysis of transitivity in learner writing may lend revealing insights into the difficulties foreign students face, as they strive to make their meaning-form mappings more native-like and less L1-driven (Chen and Foley 2004: 204).

This paper focuses on the system of transitivity, which "[...] construes the world of experience into a manageable set of process types" (Halliday and Matthiessen 2004: 170). Each of these process types has a set of associated participants and a group of circumstances. For this study, 129 argumentative compositions by 43 Spanish first year university students were retrieved from the error-annotated learner corpus NOCE (Díaz Negrillo 2007, Díaz Negrillo 2009). These texts amount to approximately 32,000 words and were collected at three different stages of the same academic year (beginning, midway and end). This article aims at:

- i)Analysing the study sample on the basis of processes, participants and circumstances.
- ii)Comparing the occurrence of the eight processes for any differences in use in the study sample.
- iii)Assessing the students' evolution, if any, in terms of their use of transitivity patterns (processes, participants and circumstances) throughout an academic year.
- iv)Exploring the influence that topic selection (given vs. free writing) may have on student's ideational perspective.

The analysis so far evidences an overall dominance of relational processes, which is in line with the findings in Chen and Foley's (2004: 193). This paper, centred on the ideational metafunction, serves to further complement two recent studies focusing on the interpersonal make-up of the NOCE corpus (Bartley and Hidalgo 2010a, Bartley and Hidalgo 2010b). With only the textual metafunction left, research based on NOCE is on its way to providing an all-embracing systemic-functional description of Spanish English learner writing.

Bartley, L. and E. Hidalgo Tenorio. 2010a. "Modality in a learner corpus, or on how identity is articulated in narratives". Paper presented at the *Fourth International Conference on Modality in English*, Madrid (Spain), 9-11th September 2010.

Bartley, L. and E. Hidalgo Tenorio. 2010b. "'People is happy', or on the way learners of English construe identity through text and talk". Paper presented at the *Sixth International Gender and Language Association Conference*, Tokyo (Japan), 18-20th September 2010.

Carl, J. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. London: Longman.

Chen, Y. and J.A. Foley. 2004. "Problems with the metaphorical reconstrual of meaning in Chinese

EFL learners' expositions". In Ravelli, L.J and R.A. Ellis (eds.), *Analysing Academic Writing: Contextualized Frameworks*. London, New York: Continuum; 190-209.

Corder, P. 1967. "The significance of learner's errors". *IRAL* 5(4): 161-170.

Díaz Negrillo, A. 2007. *A Fine-Grained Error Tagger for Learner Corpora*. Unpublished Ph.D. thesis, University of Jaén, Jaén.

Díaz Negrillo, A. 2009. *EARS. A User's Manual*. Munich: Lincom.

Ellis, R. and G. Barkhuizen. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

Halliday, M.A.K. 1994<sup>2</sup> [1985<sup>1</sup>]. *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M.A.K. and C. Matthiessen. 2004. *An Introduction to Functional Grammar*. London: Hodder Education.

Ivanic, R. and D. Camps. 2001. "I am how I sound: Voice as self-representation in L2 writing". *Journal of Second Language Writing* 10(1): 3-33.

Martínez, I.A. 2001. "Impersonality in the research article as revealed by analysis of the transitivity structure". *English for Specific Purposes* 20: 227-247.

Melrose, R. 2003. " 'Having things both ways' Grammatical metaphor in a systemic-functional model of language". In Simon-Vandenberg, A.M, M. Taverniers and L. Ravelli (eds.), *Grammatical Metaphor: Views from Systemic Functional Linguistics*. Amsterdam, Philadelphia: John Benjamins; 417-442.

Polio, C. 2003. "Research on second language writing: An overview of what we investigate and how". In Kroll, B. (ed.), *Exploring Dynamics of Second Language Writing*. Cambridge: Cambridge University Press; 35-66.

Richards, J.C. 1980. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman.

**Silvia Bernardini<sup>1</sup>, Sara Castagnoli<sup>1</sup>, Adriano Ferraresi<sup>1,2</sup>, Federico Gaspari<sup>1</sup> and Eros Zanchetta<sup>1</sup>**

<sup>1</sup> **University of Bologna (Italy)**

<sup>2</sup> **University of Naples "Federico II" (Italy)**

Turning Wikipedia into Comparapedia: Towards a new type of comparable corpus for language professionals

Special-purpose comparable corpora are among the most valuable resources for translators. Typically, however, they are not publicly available and have to be constructed as the need arises (Varantola 2003). This solution is not ideal, since the resulting corpora are likely to be small if constructed manually and rather low-quality if an automatic procedure is used, e.g. the BootCaT method (Baroni and Bernardini 2004), or else absorb more time and effort than most translators are willing to spend on the task.

Many translation professionals and students today resort to the Web and to Web-based resources for documentation about a specialised domain, for solving content-related problems, and for finding translation equivalents. Wikipedia is one of the most popular choices, thanks to features such as its multilingual nature, size, variety of domains covered, and up-to-dateness. In terms of linguistically-sophisticated searching and handling of results, though, it suffers from the well-known drawbacks

often pointed out with reference to the Web itself (Fletcher 2004).

The paper describes the method we used to tap the potential of Wikipedia for corpus construction, and compares it with other attempts along similar lines (e.g. Gamallo Otero and González López 2010). Comparapedia (En-It) is a large bilingual corpus (over 270 million words in English and almost 140 million words in Italian), allowing on-the-fly consultation of theme-restricted comparable sub-corpora. Adapting tools and methods developed for Web-as-corpus construction (Baroni et al. 2009), all the bilingual data (i.e. explicitly linked entries) are extracted from a given Wikipedia dump, cleaned, lemmatised, part-of-speech tagged, and indexed using the Corpus WorkBench (Christ 1994). Keywords are obtained from human-inserted categories and recorded as structural attributes with each text. Other structural attributes include the article id (corresponding to its title) and the article target (the title of the matching article in the other language). Thus, Comparapedia allows users not only to search (the English and/or Italian) Wikipedia as a corpus, but also to search single texts and sub-corpora covering exactly the same topics in two languages. At the moment this is achieved through the use of keywords, even though in future work we intend to investigate the potential of Wikipedia-derived ontologies for this purpose (e.g. Nastase et al. 2010).

Current work, that the paper will also address, focuses on investigating the potential of Comparapedia as a hybrid com-parallel corpus. Since some of the entries are likely to have been translated from their matching entry (or from a third text), it should be possible to align (parts of) them, and search them as a parallel (sub-)corpus. This raises methodological and theoretical issues concerning the current status of established notions such as translated vs./and original text, parallel vs./and comparable corpus, collaborative vs./and conventional authoring.

Baroni, M. and S. Bernardini (2004) "BootCaT: Bootstrapping corpora and terms from the web". In *Proceedings of LREC 2004*. Lisbon: ELDA. 1313-1316.

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta (2009) "The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora". *Journal of Language Resources and Evaluation* 43 (3): 209-226.

Christ, O. (1994) "A modular and flexible architecture for an integrated corpus query system". In *Proceedings of COMPLEX'94*, Budapest, 1994. Online: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench>.

Fletcher, W. (2004) "Making the web more useful as a source for linguistic corpora". In Connor, U. and T. Upton (eds.) *Corpus Linguistics in North America 2002*. Amsterdam: Rodopi. 191-205.

Gamallo Otero, P. and González López, I. (2010) "Wikipedia as Multilingual Source of Comparable Corpora". In *Proceedings of the third BUCC Workshop, LREC 2010*. La Valetta, Malta, 2010. 21-25.

Nastase, V., M. Strube, B. Börschinger, C. Zirn, and A. Elghafari (2010) "WikiNet: A very large scale multi-lingual concept network". In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, La Valetta, Malta, 2010. 1015-1022.

Varantola, K. (2003) "Translators and disposable corpora". In Zanettin, F., S. Bernardini and D. Stewart (eds.) *Corpora in translator education*. Manchester: St. Jerome. 55-70.

**Cinzia Bevitori (University of Bologna at Forlì)**

'jumping on the green bandwagon': The discursive construction of *green* across 'old' and 'new' media genres at the intersection between corpora and discourse

Over the past few years, the role of language, as a privileged site of meaning-making processes in the analysis of environmental discourse, has become increasingly important (see, for example, Fill and Mühlhäusler ed. 2001) even among non-linguists (e.g. Hayer 1995, 2005, Dryzek [1997]2005, Carvahlo 2005). Drawing on earlier corpus-based research on environmental discourse (Gerbig 1997, Alexander 1999, 2008), I have explored how climate change, one of the most emblematic issues of global environmental discourse shaping our new century and its geopolitics (Giddens 2009), is discursively constructed and negotiated both in news and opinion discourse in UK and US quality press (Bevitori 2010, in press).

By incorporating typically quantitative, corpus-based techniques within qualitative, discourse analytical processes of analysis (see, *inter alia*, Partington et al. 2004, Baker 2006, Morley and Bayley 2009), the present paper will build on this work by exploring aspects of the semantics of *green* across 'old' and 'new' media genres, dealing with environmental issues.

The purpose of the study is twofold. On the one hand, the semantics of *green*, its evaluative constructions and phraseology (Hunston and Thompson eds. 2000, Hunston 2010), will be investigated in 'traditional' newspaper discourse by relying on a relatively small-scale specialised newspaper corpus (Climate Change Press (CCP) Corpus; see Bevitori 2010). From a discursive perspective, small-scale, and often ad-hoc compiled, corpora can provide a far more rigorous control over the data, allowing the researcher to shift more easily from the concordance line to a close and contextualised reading of text, or selected parts of text, as well as from text to discourse. In particular, by examining the collocational behaviour of selected lemmas, such as *green*, the study will aim to uncover traces of competing discourses, their underlying ideological assumptions and reader positioning (Martin and White 2005), emerging from the analysis of the different newspapers.

On the other, the analysis of the specialised press corpus will be integrated by the investigation of a fresh new corpus of environmental news weblogs. Despite their origins as personal diary accounts, weblogs, or blogs, have been defined as a 'new media genre' (Bruns and Jacobs 2006; Herring et al. 2005, Facchinetti and Adami 2009), which have increasingly attained widespread popularity as articulated forms of journalism or, to a certain extent, 'anti-journalism' (Lasica 2001, Grossman 2004). Larger newspaper corpora, such as SiBol 1993 and 2005, as well as the BNC, will be used for reference comparison.

Alexander, R.J. 1999, 'Ecological Commitment in Business: A computer-corpus-based critical discourse analysis', in J. Verschueren (ed.), *Language and Ideology: Selected papers from the 6th International Pragmatics Conference* (Vol. 1), Antwerp: International Pragmatics Association, 14-24.

Alexander, R.J. 2008, *Framing Discourse on the Environment: A Critical Discourse Approach*, London: Routledge.

Baker, P. 2006, *Using Corpora in Discourse Analysis*, London: Continuum.

Bevitori, C. 2010, *Representations of Climate Change. News and opinion discourse in UK and US quality press: a Corpus-Assisted Discourse Study*, Bologna: BUP.

Bevitori, C. in press, 'The meanings of RESPONSIBILITY in the British and American press on climate change: a corpus-assisted discourse perspective', in S. Gozdz-Roszkowski (ed.), *Explorations across*

- Language and Corpora, Studies in Language* (ed. Barbara Lewandowska-Tomaszczyk), Frankfurt/Main: Peter Lang.
- Bruns, A. and Jacobs, J. (eds.) 2006, *Uses of Blogs*, New York: Peter Lang.
- Carvahlo, A. 2005, 'Representing the politics of the greenhouse effect. Discursive strategies in the British media', *Critical Discourse Studies*, 2(1), 1-29.
- Dryzek, J.S. [1997]2005, *The Politics of the Earth : Environmental Discourses*, Oxford: Oxford University Press.
- Facchinetti, R. and Adami, E. 2009, 'Navigating the news online: the fluidity of styles in the sea of news blogs', in *Proceedings of the AIA XXIV National Conference: Challenges for the 21st century: Dilemmas, Ambiguities, Directions*, Rome, 1-3 October 2009.
- Fill, A., Muhlhausler, P. 2001, *The Ecolinguistics Reader: Language, Ecology, and Environment*, London: Continuum.
- Gerbig, A. 1997, *Lexical and Grammatical Variation in a Corpus. A computer-Assisted Study of Discourse on the Environment*, Bern: Peter Lang.
- Giddens, A. 2009, *The Politics of Climate Change*, Cambridge: Polity Press
- Grossman, L. 'Meet Joe Blog', *Time*, Jun. 13, 2004, at <http://www.time.com/time/magazine/article/0,9171,650732,00.html>. Accessed September 2010
- Hajer, M. 1995, *The Politics of Environmental Discourse. Ecological Modernization and the Policy Process*, Oxford / NewYork: Clarendon Press.
- Hajer, M. 2005, 'Coalitions, Practices, and Meaning in Environmental Politics: from Acid Rain to BSE', in D. Howarth, J. Torfing (eds.), *Discourse Theory in European Politics*, Basingstoke: Palgrave Macmillan, 297-315.
- Herring, S. C., Scheidt, L. A., Wright, E. and Bonus, S. 2005, 'Weblogs as a bridging genre', in *Information, Technology and People*, 18 (2), 142-71.
- Hunston, S. 2010, *Corpus approaches to Evaluation. Phraseology and Evaluative language*, London: Routledge
- Hunston, S. and Thompson, G. (eds.) 2000, *Evaluation in Texts: Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press.
- Lasica, J. D. 2001, 'Blogging as a form of journalism', *USC Annenberg Online Journalism Review*, May 24. <http://www.ojr.org/ojr/workplace/1017958873.php>
- Martin, J.R and White, P.R.R. 2005, *The Language of Evaluation: Appraisal in English*, London: Palgrave Macmillan.
- Morley, J. and Bayley, P. (eds.) 2009, *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.

**Ana Bocorny (PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul)**

The use of a specialized aviation corpus (COPAER) to build a Bilingual Online Multimedia Learner's Aviation Glossary - BOMLAG

This presentation aims to introduce the corpus based Bilingual Online Multimedia Learner's Aviation Glossary or BOMLAG. The BOMLAG project presents a set of theoretical and practical elements that will lead to the building of a collaborative online multimedia interface and to the establishment of a methodology for creating glossaries for specific users from specialized corpora. Based on these elements we intend to present the prototype of a collaborative bilingual online multimedia glossary (English – Portuguese) for students of Aeronautical Sciences (BOMLAG) which thereafter may be available online to institutional users or marketed to other universities and institutions that have interested in training pilots.

An applied project in nature, this research is based on a perspective that emphasizes the use of language in real specialized communication, having in mind the needs of a specific user. For this reason, the methodology does not take a prescriptive perspective, but rather seeks to understand and describe the terminology of an area identified in real contexts and in systematized corpora collected for this purpose. It is, therefore, carried out from the confluence of methodological principles suggested by Terminology (Cabreà & Sager, 1999), terminography, lexicography (Atkins & Rundell, 2008), specialized pedagogical lexicography (Fuertes, 2010) and (Fuertes, O. P. A., & Arribas-BanPo, A., 2008), and corpus linguistics (Gillard, P., & Gadsby, A., 1998) and (Milton, J., 1998). More specifically, it is a descriptive study, based on a specialized written corpus (operations manuals of large aircraft) from which term candidates, definitions and contexts will be extracted and then arranged in an online interface. Further information will be added to the glossary based on the profile and needs of users (students of Aeronautical Sciences). In order to achieve the general objective stated, the project will be developed in five stages: (i) design, (ii) planning, (iii) development, (iv) adequacy, and (v) socialization of knowledge.

This product will be implemented in test version and will be hosted on a server with free internet access. Besides the features already exemplified, the glossary will also contain: (i) tips for using the term in different situations and constructions, (ii) exercises / individual activities, (iii) space for posting of additional questions that learners can have on the term or terminological focus on each entry, (iv) space for the posting of contributions to the construction of the entry in the same format found in collaborative dictionaries and encyclopedias like wikidictionary and wikipedia (v) space for the posting of terms or terminological units that were not covered in the glossary. These latest posts will generate new entries, which, somehow, is also an element of cooperation from the users.

Atkins, B. T., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Cabreà, M. T., & Sager, J. C. (1999). *Terminology: Theory, methods, and applications. Terminology and lexicography research and practice*, 1. Amsterdam [u.a.: Benjamins.

Edo, M. N. (2009). *The specialised lexicographical approach: a step further in dictionary-making*. Bern: Lang.

Fuertes, O. P. A. (2010). *Specialised dictionaries for learners*. Berlin: De Gruyter.

Fuertes, O. P. A., & Arribas-BanPo, A. (2008). *Pedagogical specialised lexicography: The*

*representation of meaning in English and Spanish business dictionaries*. Amsterdam: John Benjamins Pub. Co.

Gillard, P., & Gadsby, A. (1998). 'Using a learners' corpus in compiling ELT dictionaries'. In S. Granger (Ed.), *Learner English on computer* (pp. 159–171). London: Addison Wesley Longman.

Milton, J. (1998). 'Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment'. In S. Granger (Ed.), *Learner English on computer* (pp. 186–198). London: Longman.

**Marina Bondi (University of Modena and Reggio Emilia)**

**A sense of place in history - description and the lexis of space representation across genres**

The paper aims at providing an overview of variation across specialist and non-specialist genres in the discourse of history. The focus is on “general language” rather than specific terminology. Growing attention has been paid to the tools of discourse organization and their evaluative implications, with a view to their discipline specificity (Hyland & Bondi 2006). Disciplines are often characterized by their argumentative strategies as well as by their content, and phraseology can be a helpful signpost to discourse organization. The basically narrative nature of history is widely recognized and temporal notions are often focused on in the applied linguistics literature on historical discourse (Coffin 2006). Setting the scene for the narrative, however, is often equally important. How do historians represent the places where the action takes place? How are these represented in popularising discourse?

The paper explores the functions of descriptive elements in two corpora: a corpus of academic journal articles and a corpus of popularisations (from History today). The corpus allows a double dimension of comparison: across disciplines and across genres (and tenors). The methodology adopted combines a corpus and a discourse perspective (Baker 2006) while focusing on the lexis of space representation (Gerbig 2008).

A preliminary analysis of frequency data (frequency wordlists and statistical keywords) offers an overview of quantitative variation. Attention is paid both to a range of organizational units, from general connectives to other phraseological units involving a representation of space. The study is based on the analysis of concordances and clusters; the co-text of the nodes is analyzed with a view to their textual patterns, so as to bring out the semantic and pragmatic implications of many organizational units. Special attention is paid to the ways in which the generic and argumentative structure of discourse is represented, highlighting for example convergences and divergences between specialist and non-specialist discourse. Frequencies and patterns are interpreted in the light of factors characterizing academic discourse and specific disciplinary values.

The representation of places is shown to contribute to highlighting the significance of the data or conclusions produced, as well as to mapping the territory of current debate with different degrees of explicitness. They thus also become resources by which the author negotiates the his/her position with the reader according to genre-specific orientations.

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

Coffin, C. 2006. *Historical discourse. The language of time, cause and evaluation*, London, Continuum.

Gerbig, A. 2008. 'Travelogues in time and space: a diachronic and intercultural genre study'. In *Language, People, Numbers. Corpus Linguistics and Society*. GERBIG, A. & O. MASON (Eds.)

Amsterdam/New York, Rodopi.

Hyland & Bondi 2006. *Academic discourse across disciplines*. Bern: Peter Lang

**Melanie Borchers (University of Duisburg-Essen, Germany)**

Phraseology Meets Pragmatics *that is to say* Discourse Markers of Reformulation under Phraseological Investigation

The term reformulation markers groups discourse markers such as *that is to say* and *in other words*. By reformulation we basically understand the process of reinterpreting the contents of an utterance through elaboration and/or exemplification. However, just like appositions (cf. Meyer 1992), the reformulations function as paraphrases, reorientations, specialisations and/or corrections of the aforesaid utterance. They are thus used to facilitate understanding.

As the aforementioned discourse markers mostly consist of more than just one single lexeme and have been present throughout the history of the English language, the paper considers them interesting for the diachronic as well as synchronic investigation of phraseology. It thus provides an overview of previous literature and supplies new information on these multifunctional phrases.

Describing the semantic and syntactic behaviour, the diachronic perspective sheds light on the distribution of *that is to say* and its characteristic behaviour as a phraseological unit. Thus, the paper provides a framework for the study of pragmatic markers within the context of phraseology, which, in turn, allows the consideration of all its (historical) variants. An example of such an investigation into different variants is the question whether *that is* is a parallel development to, the origin of or rather the elliptic development of *that is to say*.

With the help of various corpora, this paper presents a contrastive analysis of the Middle English reformulation marker *that is to say* and its Old French pendant *c'est-à-dire*. This way, both diachronic as well as synchronic analyses provide examples of the evolution of the discourse marker itself and its further development within both languages. While the French discourse marker still seems to be frequently used, the *British National Corpus* and other more recent corpora provide proof of its use in contemporary British English. In American English (cf. the *Corpus of Contemporary American English* and the *Corpus of Historical American English*) *that is to say* is clearly on decline since the 1880s and the paper tries to provide reasons for this trend.

It is the combination of comparative (historical) phraseological investigations of discourse markers of reformulation that makes this paper as difficult to describe as it is interesting. This is why the paper focuses on the analysis of the distribution and discourse function(s) of *that is to say* throughout the history of the English language.

Blakemore, Diane. 2007. "Or'-parentheticals, 'that is'-parentheticals and the pragmatics of reformulation." *Journal of Linguistics* 43, 311–339.

Flottum, K. 1994. "A propos de c'est-à-dire et ses correspondants norvégiens." *Cahiers de Linguistique Française* 15, 109-130.

Meyer, Charles F. 1992. *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

**Alex Boulton (CRAPEL – ATILF/CNRS, Nancy-Université)**

Corpora in translation for non-translation students

Aside from their intrinsic value for the purposes of language description, corpora can be used in language courses under the broad umbrella term of "data-driven learning", or DDL (Johns 1990). They can be used as a teaching aid or learning tool, but also as a reference resource, particularly for

writing and revision / error-correction, with an increasing body of empirical evidence (cf. Boulton in press) largely concerning students needing English for academic purposes. Corpora can also be used as a reference resource in translation, and here the DDL-related research focuses largely on translation trainees (e.g. Beeby et al. 2009). However, as Zanettin (2009) points out, translation is also a staple activity in many undergraduate language programmes where learners may also benefit from corpus consultation, and not just to help with the immediate translation assignment. The techniques involved build on many cognitive and metacognitive skills (O'Sullivan 2007), and once mastered, can be applied to other language courses and a variety of future language needs. The main research question in this paper is thus: Can general language corpora be used by non-translation students for translation purposes?

The students in this study are enrolled in a translation course in the third year of a degree in English at the distance education centre of a French university. They have no prior experience of corpus use and are expected only to work only with freely available on-line monolingual corpora of contemporary English. The simple interface provided by Mark Davies (<http://corpus.byu.edu/>) to large corpora of British and American English (100 million and 400+ million words respectively) is widely used for such purposes, with over two thirds of all users declaring their primary interest as language learning, teaching or translation. The sites are suitable for novice users to navigate and are accompanied by tutorials and help features; this is highly desirable in the distance teaching context where face-to-face input is not an option, and means that the course itself can keep the introduction to the basic concepts and techniques to a minimum. After that, the constraints of the distance context play to the strengths of constructivism as the students explore the corpora on their own; though email contact with the teacher is possible, and there are discussion forums to facilitate peer-to-peer collaboration, these are generally under-used.

Following an earlier pilot study, the results presented here are based on on-line examinations where the methodology component requires the students to choose sections of a previously unseen text and demonstrate and explain how they use corpora to solve the problems encountered in context. Comparing data from two sessions shows how the students come to grips with corpora for translation, and allows a qualitative analysis of individual performance on the various techniques used with greater or lesser success. These data are backed up by questionnaires submitted after the examinations to gain feedback from both successful and less successful students. Particular attention is accorded to corpus use beyond the usual concordance lines, including frequencies, register distributions, collocates lists, word comparisons, and so on, which allow the learner to ask not just 'can I say this?', but 'is this appropriate in this translation context?'

Beeby, A., P. Rodríguez Inés & P. Sánchez-Gijón (Eds.). (2009). *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: John Benjamins.

Boulton, A. (2010). 'Learning outcomes from corpus consultation'. In M. Moreno Jaén, F. Serrano Valverde & M. Calzada Pérez (Eds.), *Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching*. London: Equinox, p. 129–144.

Johns, T. (1990). 'From printout to handout: grammar and vocabulary teaching in the context of data-driven learning'. *CALL Austria*, 10, 14–34.

O'Sullivan, I. (2007). 'Enhancing a process-oriented approach to literacy and language learning: the role of corpus consultation literacy'. *ReCALL*, 19/3, 269-286.

Zanettin, F. (2009). 'Corpus-based translation activities for language learners'. *Interpreter and*

<i>Translator Trainer</i> , 3/2, 209–224.
<b>Brian Budgell (Canadian Memorial Chiropractic College)</b>
Titles of biomedical articles
<p><b>Background:</b> Despite the importance of the title to the biomedical article, little quantitative information exists to guide authors and editors in formulating and evaluating titles. Furthermore, modern healthcare practices and policy making require the efficient recovery of clinically relevant information from the mass of electronic resources available. The retrieval of manuscripts for the purposes of informing policy and clinical decision making is strongly influenced by the manuscript title and abstract; hence, the importance of concentrating unambiguous clinically relevant information within these portions of an article. This imperative is especially pronounced in the instance of randomized controlled trials, the most rigorous form of original research into clinical effects of interventions. This article describes the application of the methods of corpus linguistics to the quantitative study of the titles of articles published in leading medical journals, with a special emphasis on titles of randomized controlled trials.</p> <p><b>Methods:</b> Titles were extracted from 2 corpora: the first consisting of 1,000 articles of various genre from the 4 leading medical journals, the second consisting of 310 randomized controlled trials (RCTs). Both corpora were analyzed for commonly occurring words, phrases and formats. The frequencies of the most commonly occurring words and formats in the 2 title corpora were compared. The titles of RCTs were also searched for clinically important information as determined by reference to the CONSORT checklist of content items for randomized controlled trials.</p> <p><b>Results:</b> Titles of biomedical articles are characterized by distinct conventions of word choice, length, recurrent phrases and format. All of these characteristics appear to be influenced by genre of article. For example, titles of RCTs contain approximately twice as many tokens as non-RCT titles. Additionally, format preference varies according to journal. In the leading medical journals, titles of RCTs use a constrained but highly technical vocabulary which makes frequent reference to research methodology, the treatment(s) under investigation, and the target disease.</p> <p><b>Discussion:</b> Quantitative analysis of 2 corpora of biomedical titles has revealed distinctive lexical and syntactical features which could aid in human and machine-based knowledge extraction. Titles of RCTs, in particular, are a dense source of unambiguous clinically important information and so are illuminating in and of themselves. Additionally, titles may act as a Rosetta stone in deciphering abstracts and bodies of full articles.</p>
<b>Miriam Buendía-Castro and Pamela Faber (University of Granada, Spain)</b>
Verbs as key elements in specialized knowledge representation
<p>Corpus linguistic studies in Terminology have focused mainly on the description and analysis of terms that are noun phrases, and have played down the need of the description of other lexical units, such as verbs (Guilbert 1973; Rey 1975; Sager 1990; L'Homme 1998). However, recent studies have highlighted the importance of verbs when they are activated in specialized texts (L'Homme 1998; Lorente 2007). In fact, much of our knowledge is made up of events and states, most of which can be linguistically represented by verbs (Faber 1999).</p> <p>In this paper we study the behaviour of verbs in specialized texts, and provide evidence of how they affect the representation of conceptual information. All the examples used to illustrate our approach were taken from a subcorpus of meteorological texts of one million words, which is part of the larger corpus designed for the Ecosystem research project*, as reflected in the EcoLexicon knowledge base (<a href="http://manila.ugr.es/visual">http://manila.ugr.es/visual</a>). These examples are typical of processes and actions within the EXTREME_EVENT frame in its sense of natural disaster.</p>

Firstly, the conceptual description of this frame was established. The concept of EXTREME\_EVENT was thus linked to its subtypes, such as HURRICANE, TORNADO, EARTHQUAKE, FLOOD, etc., by a closed inventory of conceptual relations. Each subtype has its own set of subordinate concepts and conceptual relations, which encode more specific sub-event knowledge and representation.

Afterwards, the verbs activated by the concepts denoting natural disasters were extracted, based on their frequency and recurrence in texts. Verbs were classified according to their lexical domain membership, based on the parameters of the Lexical Grammar Model\*\* (Martín Mingorance 1984, 1989, 1995; Faber and Mairal 1999). Verb meaning was finally analyzed in terms of the semantic interactions that predicates maintain with their arguments and their semantic roles in real examples of text extracted from the corpus.

Our results show that the basic meaning of each verb profiles the meaning of the different concepts linked to EXTREME\_EVENT in different ways, and offers a way to access the multidimensionality of terms and the concepts they designate. This study also highlights that when nominal terms are studied in conjunction with the verbs that most frequently activate them, this affords valuable information regarding conceptual representation. In fact, verb meaning and argument structure are crucial for the representation of conceptual information connected with the network of semantic relations activated by a specialized knowledge unit.

\*This research is part of the Project Ecosystem: Single Information Space for Frame-based Environmental Data and Thesaurus (FFI2008-06080-C03-01/FILO) funded by the Spanish Ministry for Science and Innovation.

\*\*This model was previously called the Functional Lexematic Model.

Faber, P. 1999. "Conceptual analysis and knowledge acquisition in scientific translation." *Terminologie Et Traduction* 2, 97-123.

Faber, P. and R. Mairal Usón. 1999. *Constructing a Lexicon of English Verbs*. Berlin/New York: Mouton de Gruyter.

Guilbert, L. 1973. "La spécificité du terme scientifique et technique." *Langue Française* 17, 5-17.

L'Homme, M. C. 1998. "Le statut du verbe en langue de spécialité et sa description lexicographique." *Cahiers De Lexicographie* 73(2), 61-84.

Lorente, M. 2007. "Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació." In Lorente, M., R. Estopà, J. Freixa, J. Martí and C. Tebé (eds.). *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*, vol. 2 De deixebles. 365-380. Barcelona: Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.

Martín Mingorance, L. 1984. "Lexical fields and stepwise lexical decomposition in a contrastive english-spanish verb valency dictionary." In Hartmann, R. (ed.). *LEXeter 83: Proceedings of the International Conference on Lexicography*. 226-236. Tübingen (Germany): Max Niemeyer.

Martín Mingorance, L. 1989. "Functional grammar and lexematics." In Tomaszcyk, J. and B. Lewandowska (eds.). *Meaning and Lexicography*. 227-253. Amsterdam/Philadelphia: John Benjamins.

Martín Mingorance, L. 1995. "Lexical logic and structural semantics: Methodological underpinnings in the structuring of a lexical database for a natural language processing." In Hoinkes U. (ed.). *Parorama der Lexikalischens Semantik*. 461-474. Tübingen (Germany): Gunter Narr.

Rey, A. 1975. *La terminologie: Noms et notions*. Paris: Presses Universitaires de France.

Sager, J. C. 1990. *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins.

**Bruno Cartoni and Thomas Meyer (Dept of Linguistics, University of Geneva / IDIAP research institute, Martigny, Switzerland)**

Building "directional corpora" for unbiased contrastive analysis

Large multilingual parallel corpora are easily available and vastly used in Statistical Machine Translation (SMT) and can also constitute an interesting field of investigation for empirical contrastive studies (i.e. the systematic analysis of linguistic phenomena in two (or more) languages in order to highlight differences and similarities see (Granger 2003) for an overview on corpus and contrastive analysis). However, in such corpora, language information (i.e. setting the original language in which the text has actually been written) is scarcely provided.

A few researches do take into account the translation direction of a parallel corpus and analyses the influences it can have. E.g. (Ozdowska 2009) discusses the implications of translation directions used in training language and translation models for SMT. In the field of contrastive analysis, recent studies on large corpora tend to incorporate the directionality of a corpus (like in Johansson 2006), revealing sometimes important discrepancies between analyses performed on translated or original text (and their counterpart) like in (Degand 2005). Making use of multilingual parallel corpora in linguistic investigation would consequently require methodological precaution and some pre-processing.

In this work, we introduce the notion of directional corpora, as parallel corpora where the source language (i.e. the language in which the text and/or speech has been produced) is clearly identified. We present an experiment that has been performed to extract directional corpora out of an existing parallel corpus (namely Europarl (Koehn 2005)). This specific multidirectional parallel corpus contains scarce information about the original language in which each statement was made, and simple extraction of existing language tags would gather only a small amount of directional data. The scarcity of the language information is uneven within the language pairs, so we automatically gathered all the tag information in all the file sets, mutually 'correcting' all the tags and discarding diverging information. Doing so, we significantly increased the unidirectional extraction in terms of number of words (e.g. from an English to French directional corpus of 5,609,994 English token, we result with a new directional corpus of 6'358'597 English token).

Extraction and correction techniques will be presented, together with experiments on specific linguistic phenomena (namely discourse markers) that have been performed on the extended directional corpora extracted in this study, which shows interesting discrepancies in the results in translated languages and in original languages. Further methodological issues are also addressed, such as the "translational origin" of the translated data: In multilingual corpora such as Europarl, while source language is clearly identified as the "original" and target language as the "translated", there is no evidence that the target language has been directly translated from the source language, or through a pivot language. This aspect would require other methodological precautions.

Degand, Liesbeth. (2005), De l'analyse contrastive à la traduction: le cas de la paire puisque-aangezien, in Geoffrey Williams, ed., *La linguistique de corpus*, Presses universitaires de Rennes.

Granger Sylviane. (2003) 'The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies?' In S. Granger, J. Lerot & S. Petch-Tyson (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Rodopi: Amsterdam & New York, 17-29.

Johansson, Stig. (2006), 'How well can well be translated? On the English discourse particle well and its correspondances in Norwegian and German', in Karin Aijmer & Anne-Marie Simon-Vandenberg, ed., *Pragmatic Markers in Contrast*, Elsevier

Koehn Philip. *Europarl: A Parallel Corpus for Statistical Machine Translation*, MT Summit 2005.

Ozdowska, Silvia. (2009). *Données bilingues pour la TAS français-anglais : impact de la langue source et direction de traduction originales sur la qualité de la traduction*. Actes de la conférence Traitement Automatique des Langues Naturelles, TALN'09, Senlis, France, juin 2009.

### **Maggie Charles (Oxford University Language Centre)**

#### **Making concessions in academic writing: A corpus study of patterns and semantic sequences**

The logical relation of contrast/concession is important in the construction of argumentation (Biber et al. 1999), which is a key function of academic discourse and a source of particular difficulty for international students. In this area, however, studies of academic writing have primarily focused on linking adverbials, while little attention has been paid to the role of subordinators. The aim of this paper is to investigate the subordinators of contrast/concession attested by Biber et al. (1999): 'although', 'though', 'while', 'whilst' and 'whereas' and to shed light on the patterns and semantic sequences (Hunston 2008) with which they are associated.

The present study draws on data from the British Academic Written Corpus (BAWE), which contains student assignments written for assessment at three universities in the UK and awarded good grades. The data examined come from four contrasting disciplines, representing each of the knowledge groupings identified by Becher and Trowler (2001): Business Studies (soft-applied); Chemistry (hard-pure); Computer Science (hard-applied); and Politics (soft-pure). This corpus amounts to just under a million words and consists of over 400 assignments.

The total frequencies of the subordinators vary between the disciplines, with the two soft fields showing figures that are at least twice as high per 100,000 words as those found for the hard disciplines (Business Studies: 176.9; Politics: 224.1; Chemistry: 80.5; Computer Science: 77.3). This finding is in the expected direction, since the higher frequencies in the soft disciplines reflect the prevalence of recursive knowledge construction in those fields (Becher and Trowler 2001). Each subordinator is associated with specific patterns of use: thus 'although' predominantly introduces a subordinate clause which occurs before the main clause, while for 'whereas', the pattern is reversed, with the subordinate clause usually following the main clause. Further analysis of 'although' in its most frequent pattern reveals that it tends to introduce a conceded proposition that is evaluated positively. This is followed by the proposition in the main clause, which often contains a negative evaluation:

#### **Example**

*'Although the figures given in Table 4 are useful for obtaining a feel for the results, in the absence of any significance tests, the statistical validity of any comparisons made is not assured.'* (0232b)

This semantic sequence performs several discourse functions, including anticipating and countering a possible criticism of the writer's own work and criticising the work of other researchers. The paper

explores such sequences in more detail and argues that their identification is important for understanding and teaching the construction of arguments in academic writing.

Becher, T., & Trowler, P. (2001). *Academic Tribes and Territories*. (2nd ed.). Buckingham: The Society for Research into Higher Education and Open University Press.

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

Hunston, S. (2008). 'Starting with the small words: Patterns, lexis and semantic sequences'. *International Journal of Corpus Linguistics*, 13(3), 271-295.

**Hao-Jan Howard Chen (National Taiwan Normal University)**

**Constructing a Chinese as Second Language Learner Corpus and a Web-based Concordancer**

Many researchers and language teachers believe that language corpora have great potentials for language learning and teaching. The learner corpora in particular have received much attention recently. Currently, several English learner corpora are available; however, learner corpora for many other languages are not easy to obtain. Recently, because of the rapid economic growth in China, an increasing number of students are learning Chinese as a second language. Although the number of CSL learners is increasing, very few CSL learner corpora are available for teaching and research. For CSL research, the learner corpus can play an important role. Researchers can conduct research on learners' interlanguage development, language assessment, and language pedagogy. In addition, the research findings from the learner corpus can also be used in developing Chinese teaching materials.

This paper will introduce a new Chinese as second language learner corpus and related corpus search tools developed by MTC (Mandarin Teaching Center) and SC-TOP (Steering Committee of Test of Proficiency) in Taiwan. MTC is located at National Taiwan Normal University and it is the largest Chinese learning centers in Taiwan. There are more than 1700 students enrolled in each quarter, and there are more than 200 teachers in this center. Students from more than 70 countries are studying in this center. SC-TOP is a research center sponsored by Ministry of Education for developing various Chinese as a second language tests. Based on the data provided by these two centers, a 3-million-word Chinese as a second language learner corpus has been developed. The learner corpus includes the following three different types of learner data-

1. CSL learners' short essays written in various TOP tests.
2. CSL students' writing assignments at MTC
3. CSL students' writing in the MTC achievement tests at each proficiency level

The learner corpus was further automatically tagged with a Chinese tagger called CKIP (Chinese Knowledge Information Processing) tagger developed by Academia Sinica, Taiwan. The POS-tagged CSL corpus is very useful for research and teaching. In addition to the learner corpus, a web concordancer which has several different search options was also developed. This web concordancer allows users retrieve specific words and phrases from CSL learner corpus. Thus, various CSL learners' errors can be retrieved and studied more easily and systematically. Furthermore, the POS-tagged learner corpus can be used to search for collocates used by learners. When more data are collected, users of the web-based system can also find errors and patterns produced by CSL learners from different native language backgrounds. The availability of this CSL learner corpus and the web concordancer should be able to help more researchers uncover CSL interlanguage patterns. Moreover, many teachers and students can use the learner corpus to enhance their teaching and

learning.
<b>Lucie Chlumska (Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague)</b>
Contextual Fingerprints of Czech and English Verbs
<p>One of the opportunities the corpus linguistics has given us is to explore words and phrases within their contexts. Thanks to large corpora, linguists have, for the first time, a chance to look at the language from the syntagmatic point of view as opposed to the traditional paradigmatic focus of linguistics. Contextual research can be used in different ways. Undoubtedly, the context plays a crucial role in translation and translation-oriented research or contrastive linguistic studies, but it can also shed light on grammatical, lexical and semantic properties of certain word groups or parts of speech. If we paraphrase the famous statement by J. R. Firth, we might as well say “You shall know a word by the context in which it occurs”. The focus of this paper is to apply this corpus-driven contextual approach in the study of frequent Czech and English verbs and to compare and interpret the possible contextual patterns while bearing in mind the typological difference of these two languages. While English is an isolating language with a fixed word order, Czech is largely influenced by its rich inflection which causes flexibility of the word order. The basic assumption is that each and every word has its contextual setting or sum of contexts which can be referred to as “the contextual fingerprint” of a word (see Cvrček, V.: Contextual Approach to Parts of Speech. In InterCorp: Exploring a Multilingual Corpus. NLN Praha, 2010.). Consequently, all semantic, syntactic and formal properties of a word are reflected in this contextual fingerprint. In this paper, I will only focus on the lexical variability, i.e. the number of different units in a certain position around the verb (KWIC). The analysis will be based on the hypothesis that the fewer different word types occur in the given position, the greater is the influence of the KWIC on the position, and thus the more information about the KWIC can be found by examining the particular position. In other words, the difference in variability of certain positions shows us which positions are more important for the KWIC and its meaning, function or valency than others. The assumption is that in both languages there are verbs with similar contextual fingerprints which can be put into groups and analysed. It has already been proved that the immediate context is the most influential; in this paper, I will analyse three positions on both sides of the KWIC. The research will be based on the Czech National Corpus database (the parallel corpus InterCorp and the SYN corpus line) and the BNC, and will comprise several thousand of the most frequent verbs in Czech and English.</p>
<b>Václav Cvrček (Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague)</b>
How Large is the Core of Language?
<p>Corpus research is based on the hypothesis that large and representative collections of texts reflect the language reality truly and precisely. In lexicography we assume that general corpus is a sufficient source for representation of core lexical units. However, the comparison with traditionally elaborated dictionaries shows us that there are still lexemes missing either in dictionaries or less frequently in corpora (peripheral lexemes, scientific terms etc.). Can we determine the range of lexical core exactly by corpus methods?</p> <p>Let us begin with the assumption that the core vocabulary is the part of lexicon which is common to majority of texts and speakers. We can use the proportion of hapax legomena (i.e. words that occur only once) to all word-types in relation to the growing corpus size to identify the frequency range in which core elements occur. In hypothetically small corpus (a few sentences) the hapax-type ratio will be equal to one (each word-type is also a hapax). As we add texts to corpus (up to a few million words) the hapax-type ratio decreases (the number of new words including hapaxes is continuously increasing but the majority of added tokens are new instances of words already present in the corpus) from its maximal value (=1) to the local minimum (between 0.35 and 0.45). This is the turning point (in graph it is represented by a plateau) and from now on with extending the corpus the ratio increases because the amount of hapaxes grows at a faster pace than the number of types</p>

with frequency higher than one. The graph of the hapax-type ratio (which has similar shape in different languages regardless of the types of texts or their order) resembles pipe or chibouque (hence "pipe-graph").

This empirical finding tested on corpora of Czech, English and Italian brings us closer to exact determination of the range of core lexicon (this range differs, of course, in languages with typologically distinct structure). Subsequently, we can deduce the approximate size of a corpus sufficient for compiling a dictionary covering the core lexicon.

Shape of the hapax-type ratio function also suggests that there are still some unknown differences between text and language. Some of the quantitative laws discovered by exploring individual texts might therefore be biased by this phenomenon of qualitative change in data structure when the corpus exceeds certain size. We might thus interpret the pipe-graph (with slight exaggeration) as a special case of the parole-langue distinction: the first part of the graph (decreasing function) reflects the properties of texts (parole), then there is the transient part (plateau) and the third part (increasing function) reflects the whole language or domain (with corpus large enough to eliminate idiosyncrasies of individual texts or speakers).

### **Lyne Da Sylva**

#### **Extracting a vocabulary for structured indexing: comparison among three corpora**

This project aims to develop lexical resources for automatic indexing. We are particularly interested in automatically producing back-of-the-book style indexes, which exhibit structured entries expressing various semantic relationships between main headings and subheadings and whose creation can be quite challenging by automatic methods.

A type of useful vocabulary for this type of indexing is defined, the basic scholarly vocabulary (BSV). It contains words which are used across all disciplines with roughly the same meaning. Examples are: "development", "structure", "onset", "absence", etc. It is akin to Ogden's Basic English (Ogden, 1930) and to Coxhead's Academic Word List (Coxhead, 2000), but with a different purpose and properties. This type of word can be combined with specialized vocabulary items to form evocative, structured index entries such as the following: "combustion engine, structure" or "First World War, onset". An automatic indexing prototype has been developed which combines such specialized terms with BSV terms both occurring within a set window of text (Da Sylva and Doll, 2005). To improve on the system's internal list of manually-compiled BSV words, an experiment of semi-automatic extraction of the basic scholarly vocabulary lexical items from a large English corpus of 14 million words was devised and reported on earlier (Da Sylva, 2009); it consists of abstracts of scholarly articles in pure and applied sciences as well as in the humanities and social sciences.

The present paper describes the results of a further experiment of semi-automatic extraction from two additional corpora of abstracts of scholarly articles. The goal was to extract BSV lists for both English and French; two parallel corpora were used, containing abstracts of articles in pure and applied science only. The abstracts were (human) translations of each other and represented approximately 2 million words (2,5 million for the French one). The new extraction task for French was successful in doubling the size of a previously manually compiled list. Moreover, it has proved more efficient than the equivalent task on its parallel English corpus yielding a greater proportion of BSV in the top-ranking words.

A comparison among the BSV extracted from each of the three corpora has yielded interesting results. Specifically, the extraction task on the two parallel scientific corpora has favoured words typical of the pure and applied sciences (measurements such as "rate", "increase" and "value"), scarcer in social sciences and humanities. This has confirmed our initial choice for a wider-ranging

corpus. Also, the results on the smaller scientific English corpus are inferior to those on the larger, wide-ranging one (where a greater proportion of extracted words belong to the BSV). This suggests that the results on the French corpus could be improved upon, with an appropriate corpus.

Coxhead, Averil. (2000). 'A New Academic Word List'. *TESOL Quarterly*, 34(2), p. 213-238.

Da Sylva L. (2009). 'Corpus-based derivation of a "basic scientific vocabulary" for indexing purposes'. In *Proceedings of the Corpus Linguistics Conference*, Univ. Liverpool, 21-23 July 2009.

Da Sylva, Lyne; Doll, Frédéric. 'A Document Browsing Tool: Using Lexical Classes to Convey Information'. In Lapalme, Guy; Kégl, Balász. *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005 (Proceedings)*, New York : Springer-Verlag, 2005, pp. 307-318.

Ogden, Charles K. (1930) *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber & Co., Ltd.

**Adriana Teresa Damascelli (University of Torino, Italy)**

Discourse and social issues. Changes in the lexis of the British social work

The process of industrialisation and technological innovation brought about changes in many western societies. However, advantages deriving from any kind of development and improvement may be accompanied by a series of negative consequences and, for example, result in social terms in disparities. For this and several other reasons countries usually provide services whose aims are to support the population and their needs. Social workers, for example, are professionals whose profile has been designed to work in social services organisations in order to intervene and mediate in various social contexts and help individuals, families or communities. Social work profession has been evolving: changes in the organisation of activities as well as interdisciplinary approaches have been necessary to make social work a highly specialised job, a profession with bibliographic references including theoretical background and practical applications, and an effective tool to support governmental institutions in the attempt to solve social problems.

A previous study revealed that the field of social work has undergone a process of specialisation since the 1970s, thus reflecting and responding changes in the social condition. The corpus represented by the issues of the British Journal of Social Work (henceforth CBJSW) dating from 1970 to 2008 constituted the source of investigation. The analysis of the data showed that one of the key terms, i.e.: <l>care</l>, has changed in frequency and use. Initially more frequently occurring as a noun accompanied by premodifying items, <l>care</l> has recently become highly productive as a premodifier. This is an example, but further investigation would help draw valid and meaningful conclusions about the way the discourse of social matters is represented by lexis and detecting the changes it has undergone.

This paper aims at providing other examples through the analysis of some items / patterns that have been found exploring the CBJSW corpus. The corpus has been updated and now includes texts appearing in the issues of the years 2009 and 2010 for a total of 4.5 million words. The data considered for present purposes refers to the last ten years, i.e. 2000 to 2010. The use of computer tools performing different tasks, namely the production of wordlists, concordances and key words, were used in order to retrieve the data from the corpus and provide contexts of occurrence. At the same time, some statistical data will be processed in order to show the kind of trend the items found have followed in the lapse of time considered.

Gotti, Maurizio, Ana Maria, Hornero, Maria Jose Luzon, eds. (2006), *Corpus Linguistics. Applications*

*for the study of English*, Bern: Peter Lang

Horner, Nigel, (2003), *What is social work? Contexts and perspectives*, Exeter: Learning Matters

Hundt, Marianne, Nadia Nesselhauf, and Carolin Biewer, (2007), *Corpus Linguistics and the Web*, Amsterdam: Rodopi

Thompson, Neil, (2005), *Understanding Social Work. Preparing for Practice*, Houndsmills: Palgrave Macmillan

Partington, Alan, (1998), *Patterns and Meanings. Using Corpora for Language Research and Teaching*, Amsterdam: John Benjamins

Stubbs, Michael, (2001), *Words and Phrases. Corpus Studies of Lexical Semantics*, Oxford: Blackwell

**Samuel Danso<sup>1</sup>, Eric Atwell<sup>1</sup>, Owen Johnson<sup>1</sup>, Guus ten Asbroek<sup>2</sup>, Seyi Soromekun<sup>2</sup>, Karen Edmond<sup>2</sup>, Chris Hurt<sup>4</sup>, Lisa Hurt<sup>2</sup>, Charles Zandoh<sup>3</sup>, Charlotte Tawiah<sup>3</sup>, Zelee Hill<sup>2</sup>, Justin Fenty<sup>2</sup>, Seeba Amenga Etego<sup>3</sup>, Seth Owusu Agyei<sup>3</sup>, and Betty R Kirkwood<sup>2</sup>.**

<sup>1</sup> Leeds University <sup>2</sup> London School of Hygiene and Tropical Medicine <sup>3</sup> Kintampo Health Research Centre, Ghana

<sup>4</sup> University of Cardiff

A Verbal Autopsy Corpus for Machine Learning of Causes of Death

Verbal Autopsy (VA) is a technique approved by the World Health Organization (WHO) to determine the causes of death (COD) in countries with poor death registration systems [1]. Unregistered deaths account for two-thirds of over 57 million deaths that occur annually [2], which has severe implications for health systems and policy planning. The VA technique involves interviewing of people (such as relatives or caregivers) who were close to the deceased for events that led to the demise of the individual. This information usually contains signs and symptoms of possible illness that caused the death.

The information derived from relatives is then reviewed by medical professionals who then assign the possible cause of death. This approach is very capital intensive and resource hungry especially in settings where medical professionals are in short supply [3]. To address this challenge, there have been attempts to use various computational approaches to achieve the same results by use of closed-ended questions [3, 4]. However, because the close-ended questionnaire is limited in capturing all potentially relevant information, the best option has been the use of a mixture of both closed and open ended questions to avoid any information loss [5].

"WHEN THE CHILD WAS SIXTEEN (16) DAYS OLD SHE FELL SICK WHICH LAUTED FOR THREE (3) DAYS BEFORE SHE DIED. THE CHILD WAS HAVING DIFFICULT BREATHING. ANY TIME, SHE BREATHS, YOU SEE A HOLE IN THE CHEST, AND ALSO MAKING NOISE IN THE CHEST. SHE HAD CONVULSION WHEN SHE WAS SEVENTEEN (17) DAYS OLD BEFORE SHE DIED THE FOLLOWING DAY. SHE ALSO HAD A BULGING FONTENED AND SEVERE HOT BODY WHICH LASTED FOR TWO (2) DAYS BEFORE SHE DIED. THE CHILD ALSO HAD A FIT WHICH SHE COULD NOT OPEN HER MOUTH."

Figure 1.1: A sample of is an extract of a response during VA interview from a mother who lost her child after birth, which illustrates data not captured in the closed-ended part of the questionnaire.

Samples of VA interviews have been secured from over 10,000 individuals (both infants and adults). These were obtained from a large field trial conducted in Ghana: the ObaapaVita study[6]. The VA documents have been reviewed by medical doctors and the cause of death for each individual has been ascertained. The corpus also includes related information such as the instructions for the data collectors on how VA interviews are to be conducted.

The information collected is unstructured and one of the goals of NLP research and development is

to provide a mechanism for indentifying structural elements from unstructured text. Part-of-Speech tagging schemes [7] have been applied for various purposes within the medical domain [8]. However, the VA corpus presents further annotation challenges since the information is collected by non medical professionals. We envisage tagging with a medical semantic tagset such as SNOMED-CT[9]. The research project explores NLP annotation techniques that will result in the development of a tagging scheme for VA corpus. The annotated corpus will in turn be used in training classifiers that will automatically determine cause of death. The figure 1.2 below is a conceptual diagram of the intended approach to tackling the above problem.

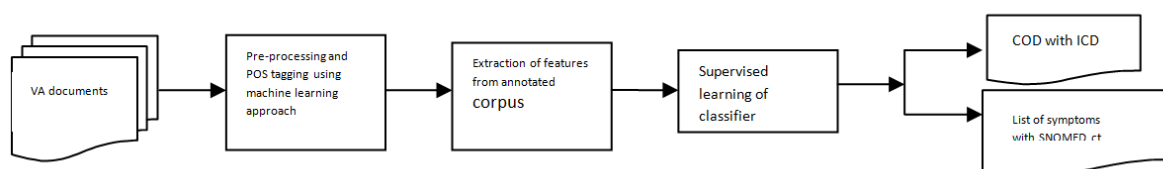


Figure 1.2 Conceptual diagram

1. BAIDEN, F., BAWAH, A., BIAI, S., BINKA, F., BOERMA, T., BYASS, P., CHANDRAMOHAN, D., CHATTERJI, S., ENGMANN, C., GREET, D., JAKOB, R., KAHN, K., KUNII, O., LOPEZ, A. D., MURRAY, C. J. L., NAHLEN, B., RAO, C., SANKOH, O., SETEL, P. W., SHIBUYA, K., SOLEMAN, N., WRIGHT, L. & YANG, G. (2007) 'Setting international standards for verbal autopsy'. *Bulletin of the World Health Organization*, 85, 570-571.
2. BYASS, P., FOTTRELL, E., DAO LAN HUONG, BERHANE, Y., CORRAH, T., KAHN, K., MUHE, L. & DO DUC VAN (2006) 'Refining a probabilistic model for interpreting verbal autopsy data'. *Scandinavian Journal of Public Health*, 34, 26-31.
3. MURRAY, C. J. L., LOPEZ, A. D., FEEHAN, D. M., PETER, S. T. & YANG, G. (2007) 'Validation of the Symptom Pattern Method for Analyzing Verbal Autopsy Data'. *PLoS Med*, 4, e327.
4. MATHERS, C. D., MA FAT, D., INOUE, M., RAO, C. & LOPEZ, A. D. (2005) 'Counting the dead and what they died from: an assessment of the global status of cause of death data'. *Bulletin of the World Health Organization*, 83, 171-177c.
5. MARSH, D. R., SADRUDDIN, S., FIKREE, F. F., KRISHNAN, C. & DARMSTADT, G. L. (2003) 'Validation of verbal autopsy to determine the cause of 137 neonatal deaths' in Karachi, Pakistan. *Paediatric and Perinatal Epidemiology*, 17, 132-142.
6. KIRKWOOD, B. R., HURT, L., AMENGA-ETEGO, S., TAWIAH, C., ZANDOH, C., DANSO, S., HURT, C., EDMOND, K., HILL, Z., TEN ASBROEK, G., FENTY, J., OWUSU-AGYEI, S., CAMPBELL, O. & ARTHUR, P.(2010). 'Effect of vitamin A supplementation in women of reproductive age on maternal survival in Ghana (ObaapaVitA): a cluster-randomised, placebo-controlled trial'. *The Lancet*, 375, 1640-1649.
7. BRANTS, T. (2000) 'TnT: a statistical part-of-speech tagger'. *Proceedings of the sixth conference on Applied natural language processing*. Seattle, Washington, Association for Computational Linguistics.
8. PAKHOMOV, S. V., CODEN, A. & CHUTE, C. G. (2006) 'Developing a corpus of clinical notes manually annotated for part-of-speech'. *International Journal of Medical Informatics*, 75, 418-429.
9. HINA,S. ATWELL E, JOHNSON, O.AND WEST, R( 2010) 'Extracting the concepts in Clinical Documents using SNOMED-CT and GATE', Fourth i2b2/VA Shared-Task and Workshop *Challenges in Natural Language Processing for Clinical Data*, AMIA, Washington DC

and Sandra Aluisio<sup>2,3</sup>

<sup>1</sup> Departamento de Letras Modernas, Universidade de São Paulo, Brazil

<sup>2</sup> Núcleo Interinstitucional de Linguística Computacional (NILC), Brazil

<sup>3</sup> Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil

<sup>4</sup> Universidade Estadual de Maringá, Brazil

Towards a multi-label sentence classifier for automatic identification of rhetorical moves in English abstracts

The relevance of automatically identifying rhetorical moves has been widely acknowledged due to its various applications to the development of Natural Language Processing tools [1,2,4,5,8,9,13,14]. A “move” refers to “a discursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse” [11].

Two approaches have been proposed to automatically detect moves in scientific texts: (i) Argumentative Zoner [12] is language-dependent and makes use of lexical, syntactical and structural features; (ii) the Text Categorization approach is language-independent and uses a bag of clusters with n-grams from 1 to 5 [1,8]. Although rhetorical moves can be realized by a clause, a sentence, or several sentences [11], most current machine-learning classifiers have established a one-to-one relationship between sentences and moves. Another drawback is the limited size of their training corpora.

The present study intends to overcome these limitations and builds on our previous work [5], which introduced AZEA (Argumentative Zoning for English Abstracts), a high-accuracy system which uses a robust set of linguistic features to automatically detect moves in abstracts from pharmaceutical sciences. Our primary aim is to develop a multi-label sentence classifier, using AZEA’s set of features and a list of formulaic-expressions created automatically [7] and two large training corpora from two broad research fields: (i) physical sciences and engineering (PE) and (ii) life and health sciences (LH). The former is made up of 845 abstracts (144,683 tokens) and the latter consists of 690 texts (150,248 tokens), taken from research papers written in English and published by various leading academic journals. Seven moves are considered [10,6]: background, gap, purpose, methodology, results, conclusion, and outline, which is used to classify instances making reference to the structure of the paper.

The Kappa Statistics [3] indicated that the multi-label sentence classification is reproducible, although some disagreements should be settled. Three experienced annotators tagged 38 abstracts from the PE and 34 from the LH corpus, previously parsed with a full syntactic parsing (OpenNLP project) and a scripting code to identify clauses and prepositional phrases. The kappa values were 0.69 (N=529, k=3, n=21) and 0.60 (N=453, k=3, n=22) for the LH and the PE corpora, respectively. The overall kappa was 0.65. The LH corpus was then automatically tagged by AZEA’s current version and manually validated by one single annotator. Full annotation of the PE corpus is in progress.

Since AZEA’s accuracy drops considerably when used in a corpus with research areas different than those from the training phase, we propose to develop two classifiers (for LH and PE). Another critical issue is that multi-labeled sentences represented only 5% of sentences from the manually annotated corpus. This paper discusses the various challenges involved in automatically assigning multi-labels to a given sentence and works towards satisfactory solutions. In addition, we also intend to make the two corpora publicly available so that they may serve as benchmark for the task.

1. Anthony, L.; Lashkia, G. (2003) ‘Mover: A machine learning tool to assist in the reading and writing of technical papers’. *IEEE Transactions on Professional Communication*, 46(3):185–193.

2. Burstein, J.; Marcu, D.; Knight, K. (2003) 'Finding the WRITE stuff: automatic identification of discourse structure in student essays'. *IEEE Intelligent Systems*, 18(1):32–39.
3. Carletta, J. (1996): 'Assessing Agreement on Classification Tasks: The Kappa Statistic'. *Computational Linguistics*, vol. 22, n. 2, pp. 249-254.
4. Feltrim, V. D.; Teufel, S.; Nunes, M. G. V.; Aluísio, S.M. (2005) 'Argumentative Zoning Applied to Critiquing Novices' Scientific Abstracts. Computing Attitude and Affect' in *Text: Theory and Applications*. 1st ed. Dordrecht, The Netherlands: Springer, 1: 159-170.
5. Genovês Jr.; L., Feltrim; V.D., Dayrell C.; Aluísio, S. (2007) 'Automatically detecting schematic structure components of English abstracts. Proceedings of the RANLP 2007', *Workshop on Natural Language Processing for Educational Resources*, Borovets, Bulgaria, pp. 23-29.
6. Hyland, K. (2000). *Disciplinary Discourses*. Harlow, UK: Longman.
7. Machado Jr., D.; Feltrim, V. D. (2009) 'Extração Automática de Expressões Indicativas para a Classificação de Textos Científicos'. *Proceedings of The 7th Brazilian Symposium in Information and Human Language Technology*, I TILIC, 2009 (Poster Presentation in Portuguese).
8. Pendar, N. ; Cotos, E. (2008) 'Automatic Identification of Discourse Moves in Scientific Article Introductions'. *Proceedings of The Third Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, pp. 62-70.
9. Siddharthan, A.; Teufel, S. (2007) 'Whose idea was this and why does it matter? Attributing scientific work to citations'. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 316-323.
10. Swales J. M.; Feak, C. B. (2009) *Abstracts and the Writing of Abstracts*, Michigan: University of Michigan Press.
11. Swales, J. (2004) *Research Genres: Exploration and applications*. Cambridge University Press, Cambridge.
12. Teufel, S. (1999) *Argumentative Zoning: Information Extraction from Scientific Text*, unpublished PhD Thesis, School of Cognitive Science, University of Edinburg, Edinburg, UK.
13. Teufel, S.; Moens, M. (2002) 'Summarising scientific articles experiments with relevance and rhetorical status', *Computational Linguistics* 28 (4), pp. 409-446.
14. Teufel, S. (2005) 'Argumentative Zoning for improved citation indexing'. In James G. Shanahan, Yan Qu, and Janyce Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications*, Springer, Dordrecht, The Netherlands, 2005, pp. 159-170.

**Stefania Degaetano (Saarland University)**

Evaluation across scientific disciplines – a corpus-based analysis

Academic discourse is said to be impersonal and objective. However, a closer look reveals its persuasive character: writers of research articles try to gain readers' attention and convince them to read on by demonstrating that they have something new and worthwhile to say (cf. Hyland 2009:70). It is the research field of evaluation that investigates how writers express their opinions, sentiments and the like. Our main interest lies in commonalities and differences regarding the

expression of evaluation across scientific disciplines. In this regard, we pose the following questions: How do scientific disciplines express evaluation? Do 'young' disciplines show differences in comparison to more established disciplines (e.g., do they promote their work more eagerly? If yes, how is this accomplished?) The data set we employ to investigate these questions is the Darmstadt Scientific Text Corpus (DaSciTex) built out of nine scientific disciplines covering recently emerged disciplines (bioinformatics, computational linguistics, computer-aided design and microelectronics) and their disciplines of origin (computer science, biology, linguistics, mechanical engineering and electrical engineering) (cf. Teich and Holtz 2009; Teich and Fankhauser 2010).

Several investigations have been carried out to classify the phenomenon of evaluation, located according to Halliday (2004) in the interpersonal metafunction (e.g., Hood 2010, Martin 2003, Hunston and Thompson 2003, Conrad and Biber 2003). However, in linguistics only few approaches deal with the identification of evaluation (e.g., Hunston 2004). But, in order to answer the above mentioned questions, we have to identify evaluation first. To identify evaluation in a large corpus, specific preliminary considerations have to be taken into account. Some evaluation is expressed in terms of specific items (e.g., modal adjuncts such as obviously, probably, apparently), which can be easily identified. In addition, lexical items belonging to word classes such as adjective or noun may be inherently evaluative. How can evaluative lexical items be identified reliably in a large corpus? Investigations on patterns (Hunston and Francis 2000) have shown that there are patterns based on adjectives which are primarily used to evaluate. Therefore, we have looked for additional evaluative patterns based on evaluative adjectives and nouns, which we call evaluative lexico-grammatical patterns, in order to reliably identify evaluative meaning in a large corpus.

For the analysis, the Corpus Query Processor (CQP) (Evert 2005) developed at the University of Stuttgart is used as it allows a fast corpus search by means of regular expressions on large annotated corpora. In order to have a basis of comparison, the evaluative modal adjuncts and the evaluative lexico-grammatical patterns are categorized according to meaning groups (obviousness, probability, importance, etc.). For the corpus comparison, the chi-square test and the Fisher's exact test are used to determine significant differences across the subcorpora of DaSciTex.

Conrad, Susan, and Douglas Biber. "Adverbial Marking of Stance in Speech and Writing." In *Evaluation in Text, Authorial Stance and the Construction of Discourse*, edited by Susan Hunston and Geoff Thompson, 56-73. Oxford University Press Inc., New York, 2003.

Evert, Stefan. "The CQP Query Language Tutorial." 2005.

Halliday, M.A.K. *An Introduction to Functional Grammar*. 3. Arnold, 2004.

Hood, Susan. *Appraising Research: Evaluation in Academic Writing*. Palgrave Macmillan, 2010.

Hunston, Susan. "Counting the uncountable: problems of identifying evaluation in a text and in a corpus." In *Corpora and Discourse*, 157-188. Peter Lang, 2004.

Hunston, Susan, and Geoff Thompson. *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford University Press, 2003.

Hunston, Susan, and Gill Francis. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. John Benjamins Publishing, 2000.

Hyland, Ken. *Academic Discourse*. Edited by Ken Hyland. London: Continuum, 2009.

Martin, J.R. "Beyond Exchange: APPRAISAL Systems in English." In *Evaluation in Text, Authorial Stance and the Construction of Discourse*, by Susan Hunston and Geoff Thompson. New York: Oxford University Press Inc., 2003.

Teich, Elke, and Mônica Holtz. "Scientific registers in contact. An exploration of the lexico-grammatical properties of interdisciplinary discourses." *International Journal of Corpus Linguistics* 14(4) (2009): 524-548.

Teich, Elke, and Peter Fankhauser. "Exploring a corpus of scientific texts using data mining." In *Corpus-linguistic applications: Current studies, new directions*, by Gries S., Wulff S. and Davies M., 233-247. Amsterdam and New York: Rodopi, 2010.

**Chiara Degano (Università degli studi di Milano)**

Corpora and argumentation: a case study

This paper explores the possibility of integrating quantitative analysis with argumentation theory in the broader frame of discourse analysis. While qualitative and quantitative approaches to the study of discourse have been profitably integrated with regard to the levels of lexico-grammar and syntax (cf. among others, Stubbs 1996, Partington et al. 2004, Garzone/Santulli 2004, Baker 2006), more rarely has this been the case for higher levels of analysis such as the argumentative structure (Degano 2007, 2010; Mazzi 2007, Mochales/Ieven 2009). Argumentative analysis is generally carried out through close reading, which allows for the reconstruction of the general structure, the identification of schemes, and the subsequent evaluation of argumentation in single communicative events. Such an approach, however does not account for the "incremental effect of discourse" (Baker 2006: 13), i.e. the constitution of recurring patterns that build up cumulatively, often without being noticed, and which can only be observed through the analysis of larger samples of discourse. The application of corpus linguistics tools to argumentative discourse could obviate this limit.

In light of such considerations, this paper sets out to identify ways in which the tools of corpus linguistics can be put to use for the study of argumentation. In order to do so, it will analyse the three televised prime ministerial debates that preceded the 2010 general elections in the UK. The choice of the sample is justified, on the one hand, by the eminently argumentative nature of this type of communicative event and, on the other hand, by its manageable size. As they were an absolute first for the UK, the three debates correspond to the totality of available materials, thus making the corpus maximally representative, and at the same time small enough to make it suitable for qualitative, as well as quantitative investigation. Apart from that, the three debates are part of a unified argumentative effort, in which it is reasonable to suppose that coherent strategies of persuasion were adopted by each candidate in the attempt of conveying a well defined, recognisable message.

A preliminary manual analysis of the debates' scripts, which relied on the model of strategic manoeuvring (van van Eemeren 2010), elaborated within the frame of the pragmadialectical theory of argumentation (van Eemeren / Grootendorst 1992, 2004), has tentatively highlighted patterns of persuasion for each candidate. The study will now proceed to identify possible linguistic indicators of such argumentative preferences with respect to the three levels of strategic manoeuvring: topical selection, adaptation to the audience and presentational choices. The identification of viable indicators would allow to confirm quantitatively the existence of distinguished patterns of argumentation for each candidate, while proving that corpora could profitably be used also for the analysis of argumentation.

Baker, Paul 2006. *Using Corpora in Discourse Analysis*. London/New York: Continuum.

Degano, Chiara 2007. 'Presupposition and dissociation in discourse: a corpus study'. *Argumentation* 21, 361-378.

Degano, Chiara 2010. 'Indicators of argumentation in arbitration awards: a diachronic perspective', 189-205. In Bhatia, V.K./ Candlin, C.N. / Gotti, M. (eds) *The Discourses of Dispute Resolution*. Bern: Peter Lang.

Garzone, Giuliana / Santulli, Francesca 2004. 'What can corpus linguistics do for Critical Discourse Analysis?' In Partington, Alan / Morley, John / Haarman, Louann (eds) *Corpora and discourse*. Bern: Peter Lang, pp. 351-368.

Mazzi, Davide 2007. 'The Construction of Argumentation in Judicial Texts: Combining a Genre and a Corpus Perspective'. *Argumentation*, 21, 21-38.

Mochales, Raquel / Ieven, Aagje 2009. 'Creating an Argumentation Corpus: Do Theories Apply to Real Arguments? A Case Study on the Legal Argumentation of the ECHR'. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law*. New York: ACM.

Partington, Alan / Morley, John / Haarman, Louann (eds.) 2004. *Corpora and Discourse*. Bern: Peter Lang.

Stubbs, Michael 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell Publishers.

van Eemeren, Frans H. / Grootendorst, Rob (1992). *Argumentation, Communication, and Fallacies: A Pragma-Dialectical Perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

van Eemeren, Frans H. / Grootendorst, Rob (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press.

van Eemeren, Frans H. / Houtlosser, Peter (2006). 'Strategic Maneuvering: A Synthetic Recapitulation'. *Argumentation* 20: 381-392.

van Eemeren, Frans H. 2010. *Strategic Maneuvering in Argumentative Discourse. Extending the Pragma-Dialectical Theory of Argumentation*. Amsterdam- Philadelphia: John Benjamins.

**Isabelle Delaere, Koen Plevoets and Gert De Sutter (University College Ghent, Ghent University)**

COMURE: corpus-based, multivariate research investigating register variation between translated and non-translated Belgian Dutch.

Since the 1990s, research in the field of translation studies has not only been focusing on the relation between source and target texts, but also on the relation between translated language versus non-translated language (Baker 1993, 2004). Corpus-based research in translation studies has shown that translated texts differ in a surprisingly systematic manner from non-translated texts resulting in, for example, lexical (Laviosa 1998, Kemppanen 2004, Tirkkonen-Condit 2004) and grammatical differences (Puurtinen 2003, Olohan & Baker 2000). It is, however, remarkable that little research has been carried out to examine the differences between registers in translated versus non-translated language.

We want to investigate (i) to what extent the linguistic choices made in translations are register dependent and (ii) whether the linguistic differences between registers in translated texts are

identical to those between these same registers in non-translated texts. To this end, we use the Dutch Parallel Corpus (DPC), a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French, with Dutch as the central language. It consists of the following registers: fiction, non-fiction, journalistic texts, instructive texts, administrative texts and external communication. (Macken et al. 2011).

We selected a set of linguistic profiles, viz. sets of alternatives with the same meaning and function such as *jij kunt* vs. *jij kan* (you can), *u heeft* vs. *u hebt* (you have) en *heeft gedaan* vs. *gedaan heeft* (have done). On the basis of the frequency of these alternatives in the DPC, we used profile-based correspondence analysis (Plevoets 2008) in order to measure and plot the linguistic distances between the various registers. The general idea behind this technique is that the linguistic distances between registers and translated vs. non-translated texts grow as a function of the extent to which linguistic choices differ.

Preliminary results show that there is a significant distance between translated and non-translated Dutch. More specifically, we observe a tendency towards the use of rather formal language for translated texts and a tendency towards the use of rather informal language for non-translated texts. Furthermore, we see that the linguistic choices made by authors or translators are register dependent, and, for external communication and journalistic texts, these choices are significantly different for translated versus non-translated texts.

Baker, M. (1993). Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology. In honour of John Sinclair*. (pp. 233-250). Philadelphia/Amsterdam: John Benjamins.

Baker, M. (2004). 'A corpus-based view of similarity and difference in translation'. *International Journal of Corpus Linguistics*, 9 (2), 167-193.

Kemppanen, H. (2004). 'Keywords and ideology in translated history texts: A corpus-based analysis'. *Across Languages and Cultures*, 5(1), 89-107.

Laviosa, S. (1998). 'Core patterns of lexical use in a comparable corpus of english narrative prose'. *Meta*, 43, 557-570.

Macken, L., De Clercq, O., & Paulussen, H. (2011). 'Dutch parallel corpus: A balanced copyright-cleared parallel corpus'. *Meta*, 56(2).

Olohan, M., & Baker, M. (2000). 'Reporting that in translated english: Evidence for subconscious processes of explicitation?' *Across Languages and Cultures*, 1(2), 141-158.

Plevoets, K. (2008). *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken belgisch-nederlands*. Doctoral dissertation, Katholieke Universiteit Leuven, Leuven.

Puurtinen, T. (2003). 'Genre-specific features of translationese? Linguistic differences between translated and non-translated finnish children's literature'. *Literary and linguistic computing*, 18(4), 389-406.

Tirkkonen-Condit, S. (2004). 'Keywords and ideology in translated history texts: A corpus-based analysis'. In A. K. Mauranen, Pekka (Ed.), *Translation universals. Do they exist?* (pp. 177-184).

Amsterdam/Philadelphia: John Benjamins.

**Leon Derczynski and Robert Gaizauskas (both University of Sheffield)**

RTMBank: A Corpus Annotated According to Reichenbach's Tense Model

Tense and time have a significant role in language. They allow the reader to locate and relate events in time. Hence the ability to automatically process and interpret tense is a crucial part of the long term programme aiming at computation language understanding.

In his 1947 account, Reichenbach (Reichenbach 1947) offered an analysis of the tenses of verbs, in terms of abstract time points. He posited that each tensed verb can be modelled with a `_speech time_` for when the verb was uttered, an `_event time_` that is the point where the events described by the verb occur, and a `_reference time_` from which events are viewed. For example, in "She will have eaten", if the speech time is the present, we are referring to an event that happens in the future -- that is, speech time is before event time -- and this is described from a viewpoint after the "eating" event, so that the reference time is later than event time. Thus, we can say that for this tensed verb,  $\text{speech time} < \text{event time} < \text{reference time}$ .

RTMML (Derczynski 2011) is a markup language for annotating the tenses of verbs and temporal relations between verbs, that aims to support automated processing of tense and temporal relations in language. RTMML differs from TimeML (Pustejovsky 2004) in that (1) it chiefly only annotates verbs that indicate events, (2) the information annotated about verbs is more nuanced, and (3) inter-verb links are defined using Reichenbach's three abstract points instead of event boundaries. In this paper we describe the annotation scheme, introduce an RTMML annotation tool, and document the creation of an RTMML-annotated corpus, RTMBank.

There are other temporally annotated corpora, the largest and most detailed of which is TimeBank (Pustejovsky 2003) -- a TimeML annotated corpus of 183 newswire articles from US outlets. However, TimeML leaves out some information which is critical to the Reichenbachian model. The three time points are also useful for some tasks that TimeBank was intended to help with, such as the anchoring of temporal expressions on a calendrical timeline. The documents in RTMBank are chosen from those already in TimeBank. Since people using RTMBank may want additional information only contained in TimeBank. Finally, choosing documents for RTMML annotation that are already annotated in TimeML permits partial verification of our annotation effort.

In our full paper, we discuss the composition of RTMBank and include some summary statistics. We also examine temporal context and the temporal relations between verbs, where a temporal context is defined as a time frame shared by one or more events.

Finally, we conclude with potential applications of the corpus, including the training and evaluation of automated systems.

L. Derczynski and R. Gaizauskas. 2011. 'An Annotation Scheme for Reichenbachs Verbal Tense Structure'. In *Sixth Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. ACL.

J. Pustejovsky, P. Hanks, et al. 2003. 'The TimeBank Corpus'. In *Corpus Linguistics*, volume 2003, page 40.

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2004. 'The specification language TimeML'. *The Language of Time: A Reader*. Oxford University Press, Oxford.

H. Reichenbach. 1947. *The Tenses of Verbs. Elements of Symbolic Logic*, pages 287–98.

**Leon Derczynski and Robert Gaizauskas (both University of Sheffield)**

A Corpus-based Study of Temporal Conjunctions

Time and tense are import in natural languages. Human readers are able to read a document and form an ordering of the events reported in it, from cues located in the language of the document -- something that we are not yet able to do automatically (Verhagen 2010). These cues may be things such as tense, the order in which events are presented, or world knowledge. One cue is the presence of a temporal conjunction such as ``before'', ``after'' or ``during''. These conjunctions have a temporal meaning that explicitly states the nature of the relation between two times or events (Lapata 2006).

TimeML (Pustejovsky 2004) is a temporal annotation language. It may be used to annotate (among other things) events, times, temporal relations between events and times (such as ``before'' or ``during''), and ``signal phrases'' - words (such as conjunctions) that provide information about temporal relations. In this paper, we describe the usage of these temporal signals and examine their properties in a TimeML-annotated corpus, TimeBank (Pustejovsky 2003), in both English and Romanian. We consider (among other things) what proportion of temporal relations have an associated signal, and for particular signal phrases, how often this phrase is used in a temporal sense. For example, of all temporal relations (TLINKs) in the English TimeBank, 11.2% use a temporal signal in the original annotation. The most frequent signal word was ``in'', accounting for 24.8% of all signal-using TLINKs. However, only 13% of occurrences of the word ``in'' have a temporal sense. The word ``after'' is far more likely (78% of all occurrences) to have a temporal sense.

Signals are of great help in the difficult task of automatically relating events (Derczynski 2010). Hence we also examine the association between signal phrases and temporal relation type. For example, the phrase ``before'' often (but by no means always) suggests a TLINK of type BEFORE, depending on the textual order of events and signal. We present data showing which temporal relations a signal is most likely to indicate.

In the course of our investigation, we noted the quality of signal annotation in English TimeBank. We found a number of cases where a temporal relation employed a signal word, but was not annotated as such. To remedy this, we identified words that sometimes occur as temporal signals (for example, ``after'' may be used temporally or positionally) and examined them, adding new signal, event and temporal link annotations where necessary.

In an attempt to improve signal annotation accuracy, we added to and clarified the TimeML signal annotation guidelines and then performed a manual re-annotation, which increased the proportion of TLINKs using signals. We will detail revised annotation guidelines in the full paper. Our modifications are freely available from the first author. Finally, we give statistics on English TimeBank after our extra annotation.

L. Derczynski and R. Gaizauskas. 2010. 'Using signals to improve automatic classification of temporal relations'. *Proceedings of the ESLLI StuS*.

M. Lapata and A. Lascarides. 2006. 'Learning sentence-internal temporal relations'. *Journal of Artificial Intelligence Research*, 27(1):85–117.

J. Pustejovsky, P. Hanks, et al. 2003. 'The TimeBank corpus'. In *Corpus Linguistics*, volume 2003, page 40.

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2004. 'The specification language TimeML'. *The Language of Time: A Reader*. Oxford University Press, Oxford.

J. Pustejovsky, L. Littman, R. Sauri, and M. Verhagen. 2006. 'Timebank 1.2 documentation'. <http://timeml.org/site/timebank/documentation-1.2.html>.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. 'SemEval-2010 task 13: TempEval-2'. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. ACL.

**Katrien Deroy (Ghent University, Belgium)**

'The important point is': highlighting information in lectures

Helping students recognise the important points of a lecture is an essential feature of effective lecture delivery. One way in which this can be achieved is by metadiscursive devices signalling relevance, or 'relevance markers' (using a term from Hunston 1994). A better understanding of the use of such lexicogrammatical devices would benefit not only educational research on note-taking and lecture comprehension but also provide information that can be used in the design of English for Academic Purposes courses for non-native speaker lecturers and students. However, there have been very few corpus-informed discussions of relevance markers (Crawford Camiciottoli 2004, 2007 and Swales 2001 being notable exceptions); instead educational studies on discourse organization in lectures include very few (corpus-derived) examples (e.g. Kiewra, 2002), while linguistic studies either remain silent on the source of their examples of discourse organizing expressions (e.g. Chaudron & Richards 1986) or focus on lexical bundles (e.g. Biber 2006).

Using the British Academic Spoken English (BASE) corpus, this study contributes to the mapping of the realizations of relevance markers in lectures. A list of relevance markers was first compiled from a manual search of 40 BASE lectures and from examples provided in Crawford Camiciottoli (2004, 2007). These were then quantified in all 160 lectures using Sketch Engine. The identification and quantification of these markers proved difficult because it often required studying discourse far beyond the concordance line and because some of the most common devices (e.g. remember; the point is) are polysemous. Furthermore, in the case of corpus linguistic research such as this the analysis of evaluation is also hampered by the lack of information about the lecturer's intentions, the students' knowledge, and non-verbal communication.

The investigation revealed a wide variety of patterns based on nouns (e.g. the important point is; that's the bottom line), verbs (e.g. remember; let me just emphasise) and adjectives (e.g. it is important to note; this is absolutely crucial). These could further be classified in terms of their orientation as speaker-oriented (e.g. that's the point i want to make; it's worth mentioning), listener-oriented (e.g. this is important to remember; it's worth knowing), content-oriented (e.g. so what's crucial is; that's the critical point) or as having joint listener-speaker orientation (e.g. we need to bear in mind; these are the things i want you to go home with). Finally, it is interesting to note that findings largely confirm those of previous research on the conversational features of this genre (e.g. Biber 2006), the extremely common use of 'the thing/point/question is' being a case in point.

Biber, D. (2006). 'University language: a corpus-based study of spoken and written registers'. *Studies in Corpus Linguistics* 23. Amsterdam: John Benjamins Publishing Company.

Chaudron, C. & Richards, J. C. (1986). 'The effect of discourse markers on the comprehension of lectures'. *Applied Linguistics*, 7 (2), 113-127.

Crawford Camiciottoli, B. (2004). 'Audience-oriented relevance markers in business studies lectures'.

In Del Lungo Camiciotti, G. and Tognini Bonelli, E. (Eds.). *Academic discourse-new insights into evaluation*. (81-98). Bern: Peter Lang.

Crawford Camiciottoli, B. (2007). *The language of business studies lectures*. Amsterdam: John Benjamins.

Hunston, S. (1994). 'Evaluation and organization in a sample of written academic discourse'. In Coulthard, M. *Advances in Written Text Analysis*. (191-218). London: Routledge.

Kiewra, K. A. (2002). 'How classroom teachers can help students learn and teach them how to learn'. *Theory into Practice*, 41 (2), 71-80.

Swales, J. M. (2001) 'Metatalk in American academic talk: the cases of point and thing'. *Journal of English Linguistics*, 29 (1), 34-54.

**Alison Duguid (University of Siena)**

Control: a semantic feature in evaluative prosody

Our experience of meanings are built up from our experience and encounters in transactional contexts and relationships, from our experience of language events (Hoey 2005). The features of the meaning potential that we choose to respond to will depend very much on such priming. When examining corpus data, we can identify such features, from the situation illustrated in the co-text, which form part of the meaning potential of lexical items, but which are not always made explicit in corpus-based dictionary definition, but which many lexical items have in common. Louw (1993) notes that 'where human beings are in control of their own destiny and are shaping it transitively for themselves, the evaluative prosody is positive, but where people are at the mercy of forces beyond their control, the things which build up intransitively are negative and uniformly threatening'.

In this paper I will consider this highly important functional psychological construct, that of control or, more precisely, of being or not being in control of events and of one's environment. I contend that this is a feature closely bound up with evaluation in language in use (Hunston and Thompson 2000) This insight allows us to make a significant generalisation about evaluation and semantic prosody for a number of frequently discussed items. For example several sets of lexical items which have been observed to carry evaluative meaning and these are cited as examples of semantic or evaluative prosody: (e.g. break out, cause, sit through, set in, end up, budge, orchestrate) (Louw, 1993; Partington, 2004; Sinclair 2004, Stubbs 2001; Hoey 2005; Whitsitt 2005; Hunston 2007; Morley and Partington 2009, Stewart 2010) can be shown to have this semantic feature in common. In discussions of evaluative prosody each item is often considered separately as illustrating the theoretical point: the concept of evaluative prosody. There is however the possibility that a semantic feature, which is discoverable from examining the items in context in real discourse, in this case that of control can be seen to be a component in all the cases.

Firstly I examine the cognitive meaning component of control by examining the lexical item control itself as used in context, by consulting the SiBol newspaper corpus of 290 million words (using Wordsmith 5) to illustrate the range of meaning potential in the term. Then I will illustrate how the notion of control is inseparably bound up with evaluation going on to show how consideration of this aspect of meaning can help resolve some of the questions and difficulties which authors have identified in evaluation theory, in particular, problems in the description of evaluative / semantic prosodies.

**Magali Sanches Duran (Universidade de São Paulo-ICMC-NILC /FAPESP, Brazil) and Carlos Ramisch**

**(LIG-GETALP, University of Grenoble, France, and INF-UFRGS, Brazil)**

## How do you feel? Investigating lexical-syntactic patterns in sentiment expression

This work investigates how sentiments are expressed in Brazilian Portuguese. Two sentiment words were used as seeds to identify recurrent lexical-syntactic patterns. Five of the seven more frequent patterns identified use light verb [1] constructions. This is widely observed in Portuguese [2,3,4], specially in colloquial speech, where sentiment expression is more likely to occur. Patterns revealed two ways of expressing sentiments: focusing on the experiencer (Eu tenho medo de avião=I have fear of airplanes) and focusing on the cause/stimulus (Avião me dá medo=Airplanes give me fear).

To find further sentiment words, we looked for these patterns in the PLN-BR-FULL corpus (<http://www.nilc.icmc.usp.br/plnbr/>), automatically lemmatised and POS-tagged, using the mwetoolkit [5], a computational system for multiword expressions identification. The resulting occurrence lists with 1774 candidates have been analysed by humans to distinguish sentiments from other words, for example ter ódio de (to have hate of) vs. ter camisa de (to have shirt of). From the 173 validated candidates, 138 are expressed by light verb constructions and 35 by content verbs. Due to high polysemy of light verbs, the respective patterns returned the largest amount of noise: precision ranges from 4,5 % to 27,45%. Content verbs are much less ambiguous and their precision ranges from 44.9% to 72.22%.

We merged sentiment nouns detected and combined them with the seven patterns, thus artificially generating 686 collocations that were automatically looked up in the web.

Results showed some collocations with zero occurrences, helping us to distinguish sentiments externally motivated, which present a “theme” as complement, from sentiments internally motivated, which do not need complements. Also, we were able to distinguish more flexible constructions from more fixed (more lexicalised) ones, with zero frequency for all alternative patterns except for the preferred one.

Sentiment words were annotated with the following labels: 1) positive / negative / neutral / context dependent; 2) transitive / intransitive; 3) physiological / psychological-emotional / psychological-rational. This deep analysis of the collocations allows clustering similar expressions.

These results can be fed back into computational systems that try to automatically extract polarity or execute sentiment analysis of textual data [6,7]. They are also relevant for word sense disambiguation: the light verbs ter, ficar, estar and dar are very polysemous, but whenever they combine with a sentiment, they have unambiguous sense, that is, ter=sentir (to feel), dar=provocar (make to feel), ficar=começar a sentir (start to feel), estar=sentir temporariamente (feel temporarily). Furthermore, these collocations may be used to improve bilingual dictionaries with information on how to express sentiments from the point-of-view of a Brazilian speaker.

1. Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. ‘Seeing arguments through transparent structures’. In *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. pp 787-791, Las Palmas, Canary Islands, Spain, May.

2. Inês Duarte et al. 2010. ‘Light verbs features in European Portuguese’. In *Proc. of Verb 2010 : Interdisciplinary Workshop on Verbs. The Identification and Representation of Verb Features*. Pisa, Italy, Nov.

3. Anabela Barreiro. 2006. ‘Extraction and formalization of support verb paraphrases from corpora

— applications in machine translation’. In *Segundo Simpósio Doutoral da Linguatca*, Lisbon, Portugal, Apr.

4. Hilda M. F. Silva. 2006. *Verbo-suporte e expressões cristalizadas. Um enfoque sintático-semântico-discursivo*. Ph.D. thesis, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

5. Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. ‘Multiword expressions in the wild? The mwetoolkit comes in handy’. In *Proc. of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, Aug.

6. Alastair J. Gill, Robert M. French, Darren Gergle, Jon Oberlander. 2008. ‘The Language of Emotion in Short Blog Texts’. In *Proc. of the 2008 ACM Conference on Computer supported cooperative work*.

7. Bo Pang and Lillian Lee. 2008. ‘Opinion Mining and Sentiment Analysis’. *Foundations and Trends in Information Retrieval*. Vol. 2, issues 1-2, pp 1-135.

#### **David Evans (University of Nottingham)**

##### The roles of intensification in spoken discourse

Much of the recent literature on ‘upscaling’ degree modifiers has been influenced by Quirk et al’s (1985) *A Comprehensive Grammar of English*. Quirk et al’s view of intensification was predominantly concerned with the syntactic and semantic restrictions that the modified (usually lexical) item placed on its immediate modifying neighbour. Where they commented on the meaning of these combinations, they focused solely on the ideational.

The focus on the impact of amplifiers on their immediate environments has continued with many corpus linguistic studies. Whilst authors such as Partington (1998, 2004) and Kennedy (2002, 2003) have concerned themselves with amplifiers more in terms of collocation, semantic prosody and semantic preference, the relative restriction of the amplifier’s position in relation to the item it is modifying has meant this area has lent itself well to concordancing. The cumulative effect of this research has been to reinforce a view of intensification based largely on a model from written discourse.

This paper uses data from a corpus of spoken British English (the Cambridge and Nottingham Corpus of Discourse in English) and attempts to move away from looking at the local effects of degree modifiers to consider how they contribute to the flow of face-to-face conversation. Studying amplifiers, both in general and in spoken discourse in particular, poses a number of problems for the corpus linguist. First, the items under investigation do not constitute a closed class, or as Bolinger (1972: 23) notes, ‘... virtually any adverb modifying an adjective tends to have or to develop an intensifying meaning.’ Second, in conversation intensification often appears ‘messy’ as modifying and modified elements are separated across turn boundaries and interpersonal meanings are negotiated between speakers.

In ‘... searching for a method ... to answer questions of a discursal nature.’ (Hunston, 2010: 167) this paper combines frequency lists and plot analyses to show that amplifiers are not distributed evenly or randomly across texts, rather they tend to cluster. Initial scrutiny of these clusters using both concordance lines and qualitative close textual analysis reveals the importance of amplifiers in observation-comment sections of spoken discourse and thus their role in ‘... creating cultural solidarity between speakers and their listeners.’ (McCarthy 1998: 142).

Bolinger, D. (1972) *Degree Words*. The Hague: Mouton.

Hunston, S. (2010) *Corpus Approaches to Evaluation*. Abingdon: Routledge.

Kennedy, G. (1998) 'Absolutely diabolical or relatively straightforward: Modification of adjectives by degree adverbs in the British National Corpus'. In A. Fischer, G. Tottie & H.M. Lehmann (eds.) *Text types and corpora: Studies in honour of Udo Fries*. (pp. 151-163) Tübingen: Gunter Narr.

Kennedy, G. (2003) Amplifier Collocations in the British National Corpus: Implications for English Language Teaching. *TESOL Quarterly*, 37(3), 467-487.

McCarthy, M. (1998) *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Partington, A. (1998) *Patterns and Meanings*. Amsterdam: John Benjamins.

Partington, A. (2004) 'Utterly content in each other's company': Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9(1), 131-156.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985) *A Comprehensive Grammar of English*. Harlow: Longman.

**Stefan Evert (University of Osnabrück) and Andrew Hardie (Lancaster University)**

**Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium**

Corpus Workbench (CWB) is a widely-used architecture for corpus analysis, originally designed at the IMS, University of Stuttgart (Christ 1994). It consists of a set of tools for indexing, managing and querying very large corpora with multiple layers of word-level annotation. CWB's central component is the Corpus Query Processor (CQP), an extremely powerful and efficient concordance system implementing a flexible two-level search language that allows complex query patterns to be specified both at the level of an individual word or annotation, and at the level of a fully- or partially-specified pattern of tokens. CWB and CQP are commonly used as the back-end for web-based corpus interfaces, for example, in the popular BNCweb interface to the British National Corpus (Hoffmann et al. 2008). CWB has influenced other tools, such as the Manatee software used in SketchEngine, which implements the same query language (Kilgariff et al. 2004).

This paper details recent work to update CWB for the new century. Perhaps the most significant development is that CWB version 3 is an open source project, licensed under the GNU General Public Licence. This change has several important consequences: first, enlarging the community of developers and users; second, requiring support for a wider range of OS platforms including Mac OS X, Linux, and Windows; and third, letting us leverage existing open-source libraries in extending CWB's capabilities – including GNU Readline, GLib, and PCRE.

In enhancing the newly-open CWB, one of the priorities has been to implement support for multiple character sets – most especially Unicode (in the form of UTF-8). While the original CWB could only deal with the Latin-1 character set for Western European languages, the UTF-8 support allows all the world's writing systems to be utilised within a CWB-indexed corpus. UTF-8 support for regular expressions is provided by the PCRE library, which has the advantage of embedding a very popular and widely-used dialect of regular expressions (namely, Perl) into CQP.

A key concern in the new CWB is the user-friendliness of the interface. CQP itself can be daunting for beginners. However, it is common for access to CQP queries to be provided via a web-interface. Bespoke interfaces of this kind are supported in CWB version 3 by several Perl modules that allow access to different facets of CWB/CQP functionality. The CQPweb front-end (Hardie forthcoming)

has now been adopted as an integral component of CWB. CQPweb provides analysis options beyond concordancing (such as collocations, frequency lists, and keywords) by using a MySQL database alongside CQP. Available in both the Perl interface and CQPweb is the Common Elementary Query Language (CEQL), a simple-syntax set of search patterns and wildcards which puts much of the power of CQP in a form accessible to beginning students and non-corpus-linguists.

Christ, O. 1994. "A modular and flexible architecture for an integrated corpus query System", in *Proceedings of COMPLEX '94*, pp. 23–32. Budapest.

Hardie, A. Forthcoming. "CQPweb - combining power, flexibility and usability in a corpus analysis tool".

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund Prytz, Y. 2008. *Corpus Linguistics with BNCweb – a Practical Guide*. Frankfurt am Main: Peter Lang.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. "The Sketch Engine", in G. Williams and S. Vessier (eds) *Proceedings of EURALEX 2004*, pp. 105–116. Bretagne, France: Université de Bretagne-Sud.

#### **Stefan Evert (University of Osnabrück)**

##### Quantitative measures of productivity and their significance

The productivity of type-token distributions is an important empirical quantity for corpus-driven approaches to language and many other subfields of linguistics. In addition to recent work on quantitative notions of morphological productivity (e.g. Baayen 1991, 2001; Lüdeling & Evert 2005), applications range from studies of the type-richness e.g. of an author's vocabulary (Efron & Thisted 1976), over stylometrics and authorship attribution (see Juola 2006 for an overview) to patholinguistics (Garrard et al. 2005).

Surprisingly, though, no solid methodological foundation for quantitative studies of productivity has been developed yet. Various measures were suggested in the literature, including the type-token ratio (TTR), Baayen's (1991) productivity index P (hapax legomena / number of tokens), Aronoff's (1976) productivity index I (observed types / possible types), Zipf's law (where the exponent of the rank-frequency law serves as a measure of the "Zipfianness" and hence productivity of the distribution), and even sophisticated statistical models that generalise from finite samples to the type-richness and Zipfianness of the underlying population (so-called LNRE models, Khmaladze 1987, Baayen 2001).

However, there are three fundamental methodological problems shared by all these approaches:

1. Most quantitative measures depend systematically on sample size (i.e. the size of the corpus for which they are computed). This can easily be demonstrated for TTR and P (as argued e.g. by Evert & Lüdeling 2001), but has also been observed with many of the sophisticated LNRE models (Baayen 2001, Fig. 5.12 on p. 182).
2. Usually, no effort is made to assess the uncertainty due to sampling variation. In particular, it is often unclear whether the difference between two observed productivity values can be deemed significant.
3. The interpretation of most productivity measures remains unclear. What exactly is the quantitative phenomenon underlying our intuitive notion of a productive process? And which measure gives the most accurate representation of this phenomenon?

This paper addresses problems 1. and 2. with the help of simulation experiments and empirical corpus data. Special attention is given to sampling variation of the productivity measures, showing how suitable confidence intervals are obtained and significance tests can be carried out. The empirical study is complemented by a discussion of problem 3., which highlights the different intuitive notions of productivity and type-richness encountered in various fields and relates them to the quantitative measures under consideration.

Baayen, R. Harald (1992). 'Quantitative aspects of morphological productivity'. *Yearbook of Morphology* 1991, pages 109–149.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Lüdeling, Anke and Evert, Stefan (2005). 'The emergence of productive non-medical -itis. Corpus evidence and qualitative analysis'. In S. Kepser and M. Reis (eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Mouton de Gruyter, Berlin.

Efron, Bradley and Thisted, Ronald (1976). 'Estimating the number of unseen species: How many words did Shakespeare know?' *Biometrika*, 63(3), 435–447.

Evert, Stefan and Lüdeling, Anke (2001). 'Measuring morphological productivity: Is automatic preprocessing sufficient?' In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, pages 167–175, Lancaster. UCREL.

Garrard, Peter; Maloney, Lisa M.; Hodges, John R.; Patterson, Karalyn (2005). 'The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author'. *Brain*, 128(2), 250–260.

Juola, Patrick (2006). 'Authorship attribution'. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.

Khmaladze, E. V. (1987). 'The statistical analysis of large number of rare events'. *Technical Report MS-R8804*, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

**Vinícius Mourão Alves de Souza and Valéria Delisandra Feltrim (both State University of Maringá)**

An analysis of textual coherence in academic abstracts written in Portuguese

This paper presents a study regarding the analysis of textual coherence in academic texts written in Portuguese, focusing on the abstract section of dissertations in Computer Science. The coherence is analyzed from the semantic relationship observed between certain parts of the abstract that constitute its schematic structure. Such a structure is the same proposed in the writing tool called SciPo, consisting of six rhetorical components (Feltrim et al., 2006): Background, Gap, Purpose, Methodology, Result and Conclusion. Using a corpus of 385 abstracts, the first stage of the analysis was the classification of a total of 2.293 sentences as one of the given rhetorical components. To speed up the process of rhetorical annotation, the sentences were previously submitted to an automatic classifier called AZPort (Feltrim et al., 2004), which assigns a single rhetorical category to each sentence (among the six possible ones), and the results were then corrected by a human annotator. The observed distribution of components was as follows: Background – 34.78% (808 sentences), Gap – 9.26% (215 sentences), Purpose – 19.63% (426 sentences), Methodology – 11.75% (273 sentences), Result – 19.41% (451 sentences) and Conclusion – 5.17% (120 sentences). In the next stage of the annotation process, the annotator was asked to manually annotate three types of relationship for each of the 2.293 sentences: (i) the level of relationship between each sentence of

the abstract with its title, with two possible values, 'high' or 'low'; (ii) the level of relationship between the sentences classified as been part of the Purpose component and the other sentences of the abstract, also with two possible values, 'high' or 'low'; and (iii) if the relationship of the current sentence with its antecedent shows a break in linearity of meaning in the text, with possible values 'yes' or 'no'. Concerning (i), we observed a strong relationship between sentences of the Purpose component and the abstract title, as 83.33% (355) of these sentences were classified as having a high relationship with the title, while in other components we observed an average of 48.79 % of high sentences and 51.21% of low sentences. Regarding (ii), which aims at analyzing the semantic relation between the Purpose component and other components of the abstract, we observed that Conclusion, Methodology and Result are the most related, with 72.55% (74), 67.59% (171) e 66.17% (264) of high sentences, respectively. Unfortunately, no satisfactory result was observed regarding (iii), which leads us to believe that the analysis from the rhetorical structure is not the best way to identify breaks in the sense linearity and that another approach, such as Entity-Grid (Barzilay and Lapata, 2008), could give better results. We believe that these results can be used in the development of an automatic process to identify problems related to coherence in academic texts, aiming at incorporating it in writing tools like SciPo.

Barzilay, R.; Lapata, M. (2008). 'Modeling local coherence: An entity-based approach'. *Computational Linguistics*, 34(1), p. 1–34.

Feltrim, V.D., Pelizzoni, J.M., Teufel, S., Nunes, M.G.V., Aluísio, S.M. (2004). 'Applying Argumentative Zoning in an automatic critiquer of academic writing'. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004)*, São Luis-MA, Brazil. *Lecture Notes in Artificial Intelligence*, 3171, Springer, p. 214-223.

Feltrim, V. D.; Teufel, S.; Nunes, M.; Aluísio, S. (2006). 'Argumentative zoning applied to critiquing novices scientific abstracts'. *Computing Attitude and Affect in Text: Theory and Applications*, Springer, p. 233–246.

**Adriano Ferraresi (University of Naples "Federico II" / University of Bologna, Italy) and Stefan Th. Gries (University of California, Santa Barbara)**

Type and (?) token frequencies in measures of collocational strength: lexical gravity vs. a few classics

Most well-established measures for calculating the collocational strength between words X and Y (e.g. MI, t-score and log-likelihood) are based on their joint frequency  $n$  and their two overall token frequencies  $a$  and  $b$  in a corpus. As such, they do not take into account how many different *types* co-occur with  $x$  in the position of  $y$  (and vice versa). However, this has been suggested to be a relevant criterion, e.g. within so-called "phraseological" approaches (Nesselhauf 2004) where "restricted collocations" are usually defined in terms of the number of different words (i.e. types) a node co-occurs with. Also, recent psycholinguistic studies show that e.g. the acquisition of syntactic patterns and their diachronic change are influenced by the diversity of their linguistic contexts, one operationalization of which are of course type frequencies (cf. Goldberg 2006, Bybee 2010).

Recently, Daudaravičius and Marcinkevičienė (2004) introduced lexical gravity, which incorporates information on frequency of type co-occurrence into its computation. Despite its potential theoretical interest, this measure is still underexplored in the corpus linguistics literature. Exceptions are Gries (2010), who finds that it outperforms t-score when used as a feature in cluster analysis to discriminate between different (sub-)registers in the BNC Baby, and Gries and Mukherjee (2010), who calculate lexical gravities for  $n$ -grams in different components of the ICE corpus revealing differences between varieties of Asian English and British English. However, lexical gravity has never been analysed per se as a measure of collocational strength and compared to better-established ones.

Relying on part-of-speech patterns for the identification of collocation candidates (along the lines of e.g. Evert 2008), this paper takes a first step towards filling this gap. Lexical gravity, MI, log-likelihood and bare frequency are calculated for adjective-noun pairs in two corpora: (i) a relatively small (~ 7m words) specialised corpus of English consisting of webpages of British and Irish universities (Bernardini et al. 2009), and (ii) the British National Corpus. The results are compared by means of rank correlations to explore the degrees to which the new measure returns different results than the others. In a second step we use a "stratified sampling" method, whereby for every measure we split the scored bigram lists into frequency ranges and extract the bigrams with the highest and lowest scores within these ranges, i.e. the bigrams for which, frequency being comparable, the most conflicting results are obtained. This method allows a more thorough scrutiny of the lists, since it does not limit itself to considering the n top-scored bigrams. Results suggest that lexical gravity is strongly correlated with co-occurrence frequency, but at the same time it distinguishes, within one frequency bin, between groups of salient and less salient bigrams.

Bernardini, S., A. Ferraresi and F. Gaspari. 2009. "Institutional English in Italian University websites: the acWaC corpus". Paper presented at *Corpus Linguistics 2009*, University of Liverpool.

Bybee, J. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.  
Daudaravičius, V. and R. Marcinkevičienė. 2004. "Gravity counts for the boundaries of collocations". *International Journal of Corpus Linguistics* 9(2). 321-348.

Evert, S. 2008. "A lexicographic evaluation of German adjective-noun collocations". In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions*. Marrakech, Morocco.

Goldberg, A. E. 2006. *Constructions at work: on the nature of generalization in language*. Oxford: Oxford University Press.

Gries, S. Th. 2010. 'Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora'. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, S. Th. and J. Mukherjee. 2010. 'Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes'. *International Journal of Corpus Linguistics* 15(4). 520-548.

Nesselhauf, N. 2004. *Collocations in a learner corpus*. Amsterdam: Benjamins.

#### **Lynne Flowerdew (Hong Kong University of Science and Technology)**

##### **Corpus-based Critical Discourse Analyses**

The past few years have seen an increasing number of corpus-based studies, either explicitly grounded in or inspired by, the tenets of Critical Discourse Analysis (CDA). Corpus-based CDA studies can be seen as either associated with Fairclough's (2003) approach to CDA or the discourse-historical approach of the Viennese school (Wodak & Meyer 2009). The purpose of this paper is to review these studies, while at the same time raising key issues in the interpretation of corpus data.

Those studies inspired by Fairclough's analytical framework of a discursive event are often grounded in SFL theory, especially the Appraisal system for analyzing evaluative discourse (Bednarek 2006; Coffin & O'Halloran 2006). Work in the area of corpus-assisted discourse studies (CADS) also falls within the Fairclough camp with its focus on dialogic positioning in political discourse (Morley & Bailey 2009). Handford's (2010) work on the language of business meetings can be seen as CDA-inspired with its focus on discursive practices and strategies. Those corpus studies taking a more discourse-historical approach tend to analyse text from a diachronic perspective and employ a multi-

perspective analysis that goes beyond the linguistic elements of the text, encompassing a more contextual perspective (Baker et al. 2008).

However, in spite of the advances in discourse analysis afforded by corpus linguistic methodologies, corpus data do not show what is 'invisible' (de Beaugrande 2002). Moreover, it has also been pointed out that corpus data reflect the product and not the process whereby discourse is created (Widdowson 2004). These two issues will also be discussed with reference to the above studies.

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.

Bednarek, M. (2006) *Evaluation in Media Discourse*. London: Continuum.

Coffin, C. & O'Halloran, K. (2006) 'The role of Appraisal and corpora in detecting covert evaluation'. *Functions of Language*, 13 (1): 77-110.

De Beaugrande, R. (2002) 'Linguistics, sociolinguistics, and corpus linguistics: Ideal language versus real language'. *Journal of Sociolinguistics*, 3 (1): 128-138.

Fairclough, N. (2003) *Analysing Discourse*. London: Routledge.

Handford, M. (2010) *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Morley, J. & Bailey, P. (Eds.) (2009) *Corpus-Assisted Discourse Studies on the Iraq Conflict*. London: Routledge.

Widdowson, H.G. (2004) *Text, Context, Pretext. Critical Issues in Discourse Analysis*. Oxford: Blackwell.

Wodak, R. & Meyer, M. (Eds.) (2009) *Critical Discourse Analysis*. London: Sage.

#### **Richard Forsyth and Serge Sharoff (both University of Leeds)**

From Crawled Collections to Comparable Corpora: an Approach based on Automatic Archetype Identification

With the rise of the "web as corpus" (Kilgarriff & Grefenstette, 2003) many corpora are being compiled by "crawling" the world-wide web using a variety of search strategies (Baroni, et al, 2009). This approach has several advantages, not least the fact that it offers a rapid way for collecting documents in new and rapidly developing fields in a large range of languages, so that we can compare usages across them. Such corpora can be made closer to traditional corpora such as the BNC by applying metadata (Sharoff, 2007).

The present study forms part of a project whose ultimate aim is to facilitate the rapid development of bilingual terminological lexicons in emerging technological fields by compiling comparable corpora in a number of languages (see [www.ttc-project.eu](http://www.ttc-project.eu)). To achieve this goal we first need to ensure monolingual comparability. This paper describes an initial step in that process, the partitioning of the crawled collection into emergent subgroups using unsupervised machine learning, and then the identification of archetypal texts, representative of their subgroups, which can be used as probes in their own right to find additional documents of a similar type.

The procedure has a number of novel aspects. Firstly, the features used to characterize textual

similarity and difference pay more respect to the inescapably sequential nature of language than the more conventional term-vector (or "bag of words") approach. Our feature-finding technique is based what Sinclair (1991) calls the "idiom principle", namely the tendency for speakers and writers, as well as listeners and readers, to work with chunks of language rather than isolated words. The results of such chunking have been referred to by a variety of terms, such as "collocations", "congrams", "lexical bundles", "multi-word expressions", "prefabricated phrases", "skipgrams", among other designations (Cheng et al., 2006). All are generalizations of the basic notion of an n-gram, but different authors have generalized this concept in slightly different ways, and thus the meanings of these terms overlap in a somewhat confusing manner. As the terminology for flexible multi-element linguistic units is not yet standardized, we refer in this paper to "flexigrams" (Min & McCarthy, 2010).

Secondly we use an evolutionary algorithm to find archetypes, by optimizing an objective function which generalizes that used in Ward's method of agglomerative clustering (Ward, 1963).

Thirdly, our software can optionally reduce the dimensionality of the archetype data by finding a subset of informative attributes, using essentially the same algorithm.

We will describe the results of applying these methods to collections gathered from web-crawls designed to gather texts concerning renewable energy in Chinese, English and Russian, among others.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). 'The WaCky wide web: a collection of very large linguistically processed web-crawled corpora'. *Language Resources and Evaluation*, 43(3), 209–226.

Cheng, W., Greaves, C. and Warren, M. (2006). 'From n-gram to skipgram to congram'. *International Journal of Corpus Linguistics*, 11(4), 411-433.

Kilgarriff, A. & Grefenstette, G. (2003). 'Introduction to the special issue on web as corpus'. *Computational Linguistics* 29 (3), 333-347.

Min, H.C. & McCarthy, P.M. (2010). 'Identifying Varietals in the Discourse of American and Korean Scientists: A Contrastive Corpus Analysis Using the Gramulator'. *Proc. 23rd International Florida AI Society Conference (FLAIRS-2010)*, 247-252.

Sharoff, S. (2007). 'Classifying web corpora into domain and genre using automatic feature identification'. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve.

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

Ward, J.H. (1963). 'Hierarchical grouping to optimize an objective function'. *J. American Statistical Association*, 58, 236-244.

**Karèn Fort (INIST-CNRS, France; LIPN, France), Adeline Nazarenko (LIPN and Université Paris-Nord, France), and Claire Ris (INIST-CNRS, France)**

Corpus linguistics for the Annotation manager

Hand crafted annotated corpora are acknowledged as critical elements for the Human Language Technologies but systems have to be trained on domain specific data to achieve a high level of performance. This is the reason why numerous annotation campaigns are launched. The role of the annotation manager consists in designing the annotation protocol, sometimes selecting the source

data, hiring the required number of annotators with the adequate competences and writing the annotation guidelines.

However, for a given task, the complexity of the annotation work seems to be highly dependent on the type of corpus to annotate. Since this affects both the cost and the quality of the annotation, it is an important issue to tackle for the annotation manager. This paper reports on an experiment where an heterogeneous corpus of soccer match reports had to be annotated in order to identify the key actions made by the soccer players during the matches. The goal is then to train a summarising tool to process match reports. The challenge for the annotation manager is to define guidelines that are generic enough for annotators who have to annotate reports that are, for the one part transcriptions of match comments from TV (155 Kwords) and, for the other one, summaries that are written on the spot (i.e. in bad conditions) by journalists (39 Kwords). Beside language modality and conditions of production, the corpus heterogeneity also comes from the variety of commentators, of match levels (international vs. national) and of soccer teams (national vs. club). Of course, this heterogeneity has a strong impact on the quality of the report language, with, among others, a lot of ellipses and aborted sentences.

We show that this corpus heterogeneity affects all aspects of the annotation protocol: the selection of the a sub-corpus for training, the duration of the annotator's training, the complexity of the annotation formalism, the quality of the resulting annotation. From these observations, we propose

- 1) corpus-based indicators that help the annotation manager anticipating the complexity of the annotation task (complexity in target and boundaries identification as well as in annotation formalism and vocabulary),
- 2) recommendations for writing robust guidelines to be used both on written and transcribed reports and controlling the training of annotators,
- 3) Tools for measuring the quality of the resulting annotations (inter-annotator agreement in context of sparse and unbalanced data).

[We want to thank Vincent Claveau (INRIA, France) for the fruitful collaboration on this project, and Alain Zérouki (INIST, France), who annotated part of the corpus. This work was realized as part of the Quaero Programme, funded by Oseo, French State agency for innovation.]

Gut, U. & Bayerl, P. S. 'Measuring the Reliability of Manual Annotations of Speech Corpora', *Proceedings of Speech Prosody*, 2004, 565-568

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

Wynne., M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*

**Ana Frankenberg-Garcia (ISLA Campus Lisboa)**

Sentence boundaries in translation

As predicted by Mona Baker almost two decades ago (Baker 1993), the rise of corpus linguistics has had a great impact on translation studies over the past years. Bilingual comparable corpora have been used in translation training (e.g. Kübler 2011) and to explore functional equivalents between languages (e.g. Tognini Bonelli & Manca 2004). Comparable corpora of translated and non-translated texts have been used to study what sets translated language apart (e.g. Laviosa 1998, Olohan & Baker 2000, Tirkkonen-Condit 2004, Puurtinen 2004, Frankenberg-Garcia 2008). And parallel corpora have been used in various studies comparing source-texts and translations (e.g. Frankenberg-Garcia

2005, Johansson 2007, Frankenberg-Garcia 2009, Pérez Blanco 2009, Lefer 2010, Bernardini 2010, Saldanha 2011). However, most existing corpus-based analyses focusing on the contrasts between source texts and translations are either purely lexical or are constrained by sentence boundaries. Probably the main reason underlying this limitation is the fact that, in parallel corpora, texts are usually segmented at the level of the sentence due to the relative ease with which sentence boundaries can be identified automatically. The alignment of source texts and translations is then usually carried out such that whenever there is not a one-to-one correspondence between source-text and translation segments, they are aligned either on a one-to-many or on a many-to-one basis, blurring out the details of what happens when source-text sentences are not preserved in translation. With the focus of this 6th Corpus Linguistics conference being on discourse, the study of which transcends the level of the sentence, this paper aims to explore shifts in sentence boundaries that occur in the process of translation.

The unique, manually post-edited alignment of COMPARA - a three-million-word parallel corpus of English and Portuguese fiction (Frankenberg-Garcia & Santos 2003) - enables one to retrieve automatically all sentences that were joined together and all sentences that were split apart in translation. Using this free online corpus, the present exploratory study seeks to answer questions such as (1) To what extent do translators preserve sentence boundaries? (2) What is more common in translation: sentence-splitting or sentence-joining? (3) Do English and Portuguese language translators differ in terms of preserving sentence boundaries? (4) Could author or translator style bear an influence on the extent to which sentence boundaries are preserved or not? (5) What kind of sentences do translators split and what kind sentences do they join together? It is believed that the findings observed can have implications not only for translator training, but also for bilingual text alignment and the development of machine translation and translation memory systems.

Baker, M. (1993) 'Corpus linguistics and translation studies. Implications and applications.' In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam & Philadelphia: John Benjamins, 233-250.

Bernardini, S. (2010) 'Parallel corpora and the search for translation norms/universals'. Plenary presented at MATS 2010, Methodological Advances in Corpus-Based Translation Studies, University of Gent, 8-9 January 2010. Slides available online at <http://veto.hogent.be/actua/mats2010/presentations.cfm>.

Frankenberg-Garcia, A. (2005) 'A corpus-based study of loan words in original and translated texts'. In P. Danielsson & M. Wagenmakers (eds.) *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, 14-17 July. Available online at <http://www.corpus.bham.ac.uk/pclc/index.shtml>.

Frankenberg-Garcia, A. (2008) ' "Suggesting rather special facts": a corpus-based study of distinctive lexical distributions in translated texts'. *Corpora*, 3/2, 195-211.

Frankenberg-Garcia, A. (2009) 'Are translations longer than source texts? A corpus-based study of explicitation'. In Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (eds.) *Corpus use and learning to translate (CULT): An Introduction*. Amsterdam & Philadelphia: John Benjamins, 47-58.

Frankenberg-Garcia, A. & D. Santos (2003) 'Introducing COMPARA, the Portuguese-English parallel translation corpus'. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translation Education*. Manchester: St. Jerome Publishing, 71-87.

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive*

*studies*. Amsterdam & Philadelphia. John Benjamins.

Kübler, N. (2011). 'Working with corpora for translation teaching in a French-speaking setting'. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (eds.) *New Trends in Corpora and Language Learning*. London: Continuum, 62-80.

Laviosa, S. (1998) 'Core patterns of lexical use in a comparable corpus of English narrative prose'. *Meta*, 43/4, 557-570. Available online  
<http://www.erudit.org/revue/meta/1998/v43/n4/003425ar.pdf>

Lefer, M-A. (2010) 'Morphological productivity and untranslatability: a corpus-based approach.' Paper presented at *MATS 2010, Methodological Advances in Corpus-Based Translation Studies*, University of Gent, 8-9 January 2010. Slides available online at  
<http://veto.hogent.be/actua/mats2010/presentations.cfm>.

Olohan, M. & M. Baker (2000) 'Reporting that in translated English: Evidence for subconscious processes of explicitation?' *Across Languages and Cultures* 1/2, 141-158.

Pérez Blanco, M. (2009) 'Translating stance adverbials from English into Spanish: a corpus-based study'. *International Journal of Translation*, 21, 41-55.

Puurtinen, T. (2004) 'Clause connectives in Finnish children's literature'. In A. Mauranen and P. Kujamäki (eds.) *Translation Universals, Do They Exist?* Amsterdam & Philadelphia: John Benjamins, 165-176.

Saldanha, G. (2011) 'Translator style: methodological considerations'. *The Translator*, 17(1), 25-50.

Tirkkonen-Condit, S. (2004) 'Unique items – over – or under-represented in translated language?' In A. Mauranen and P. Kujamäki (eds.) *Translation Universals, Do They Exist?* Amsterdam & Philadelphia: John Benjamins, 177-184.

Tognini Bonelli, E. & E. Manca, Elena (2004) 'Welcoming Children, Pets and Guests: Towards a Functional Equivalence in the Languages of "Agriturismo" and "Farmhouse Holidays"'. *TradTerm*, 10, 295-312.

#### **Rachelle Freake (Queen Mary, University of London)**

##### Methodological challenges in bilingual corpus-assisted discourse analysis

This paper presents some of the challenges arising from the use of corpus-assisted discourse studies (CADS) of bilingual, non-parallel corpora (Baker et al., 2008; Partington, 2004; Qian, 2010). Since corpus linguistics and discourse analysis ultimately focus on "real language use" rather than theoretically constructed examples (e.g. Stubbs, 1996; Halliday & Matthiessen, 2004), it follows that data should be as varied as the population, and to an even larger extent when the population is ethnolinguistically diverse. Data for CADS research, then, can pose numerous methodological issues to researchers, particularly if they are using CADS in sociolinguistic research and drawing on multilingual, multicultural data.

Examples for this paper are drawn from two cases in Canada, a country that contains a diverse population indexed by two official languages, English and French (Gal and Irvine, 1995; Heller, 1999). The first bilingual corpus is comprised of texts submitted to the 2007 Bouchard Taylor Commission in Quebec (Freake et al., 2011). The second bilingual corpus consists of Canadian newspaper articles from 2009. In both cases, the unilingual components of the bilingual corpora are of unequal sizes

and index a sector of Canadian society wherein a specific ethnolinguistic population dominates. In the first case, the majority group in Quebec is ethnolinguistically French-Canadian, and so it follows that the French component of the Bouchard Taylor corpus is comparatively larger. In the second instance, there are more English speakers (newspaper producers and readers) in Canada than French speakers, hence the larger English component of the bilingual newspaper corpus. These corpus differences, which arose from the use of naturally-occurring data in a particular but not unusual context, led to a number of challenges which were addressed in the research. The challenges highlighted here are the following:

Studying single-language corpora of texts produced by multiethnic, linguistically-indexed populations;

The essentialization of differences and the reification of pre-existing ethnolinguistic categories;

The semantic preference of “cognates” in corpora of different languages; and

The comparison of keywords derived from comparator corpora of different languages of different sizes.

Overall, CADS is found to be a valuable method and approach to data, providing unique top-down and bottom-up perspectives. However, when CADS is used to examine sociolinguistic data, it is not without potential shortfalls. To address these shortfalls, this paper presents some methodological solutions that were found to be useful in the Canadian context. These solutions may be useful to other researchers employing CADS in other multilingual, multicultural contexts.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. & 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.

Freake, R., Gentil, G. & Sheyholislami, J. (2011). ‘A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec’. *Discourse & Society*, 22(2), 21-47.

Gal, S. & Irvine, J. T. (1995). ‘The boundaries of languages and disciplines: How ideologies construct difference’. *Social Research*, 62 (4), 967-1001.

Halliday, M. A. K. and Matthiessen, C. (2004). *An Introduction to Functional Grammar*. London: Arnold.

Heller, M. (1999a). ‘Heated language in a cold climate’. In J. Blommaert (Ed.), *Language ideological debates* (pp. 143-170). Berlin: Mouton de Gruyter.

Partington, A. (2004). ‘Corpora and discourse, a most congruous beast’. In A. Partington, J. Morley & L. Haarman (Eds), *Corpora and Discourse* (pp. 11–20). Bern: Peter Lang.

Qian, Y. (2010). *Discursive construction around terrorism in the People’s Daily (China) and The Sun (UK) before and after 9.11*. Oxford: Peter Lang.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford, UK & Cambridge, USA: Blackwell Publishers.

**Rachelle Freake (Queen Mary, University of London)****A bilingual corpus-assisted discourse study of language and nationalism in Canadian newspapers**

This paper examines nationalist discourses in Canadian newspapers using the method of bilingual corpus-assisted discourse studies (CADS) (Baker et al., 2008; Freake et al., 2011; Partington, 2004).

Research has suggested that the media discursively isolate the French and English Canadian populations by cultivating cross-cultural misunderstandings (Oakes & Warren, 2007; Taras, 1993). However, remarkably little research has endeavoured to account for differences between Canada's French and English-language media. Media language can have a profound effect on the way populations speak, inform themselves, and even think (van Dijk, 1993), effectively creating "imagined communities" of belonging (Anderson, 1983; Billig, 1995). Canada, while officially bilingual and multicultural, is a "multinational" country, home to multiple national minorities (Kymlicka, 1995). It is proposed here that dominant nationalist discourses circulate in Canada, reflecting and reinforcing the largest and most socioeconomically powerful nations in the country (Bourdieu, 1977; Wodak et al., 2009). There have been remarkably few comparative discourse analyses of French and English Canadian texts (for exceptions, see Gagnon, 2003; Kuhn and Lick, 2009; Robinson, 1998), and research on nationalism in Canadian daily newspapers has "fallen out of fashion" (Roy, 2009: 260-1). Hence, this research addresses this research gap by examining the ways in which official languages (English and French) are represented and used to serve different ideological purposes in the nationalist discourses that are circulated in the English and French Canadian media (Fletcher, 1998).

The data include over 27 000 English and French articles (over 11 million words) from a three-week period in 2009 that are drawn from 18 newspapers (12 English and 6 French) with the highest circulation figures across the five regional divisions of Canada (see Canadian Newspaper Association, 2009). The methodology combines the concepts and tools of corpus linguistics (CL) with discourse analysis (DA). Using this methodology, broad discursive patterns revealed through frequency and statistical significance tests are complemented by considerations of semantic prosody differences between the two languages (Morley & Partington, 2009). These overarching trends are, in turn, complemented with the DA of individual ("concordance") lines of text, and indeed entire newspaper articles. The discourse analysis used here is based on Hallidayan functional grammar (Halliday & Matthieson, 2004), evaluation (Martin & White, 2005), and critical discourse analysis (Wodak et al., 2009; van Dijk, 1993, 2006). Combined, these techniques shed light on the different sociopolitical and historical contexts in which discourses of French and English Canadian nationalism are produced and consumed.

Findings provide unique insight into current nationalism issues in Canada, and more notably, suggest how differences between English and French newspapers may impact and perpetuate the divide between Canada's two majority linguistic groups and impact conceptualizations of belonging within Canada.

Anderson, B. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. & 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.

Billig, M. (1995). *Banal nationalism*. London: Sage Publications.

- Bourdieu, P. (1977). 'The economics of linguistics exchanges'. *Social Sciences Information*, 16 (6), 645-668.
- Canadian Newspaper Association. (2009). *Circulation data report*. Retrieved from [http://www.cna-acj.ca/en/system/files/2009CirculationDataReport\\_1.pdf](http://www.cna-acj.ca/en/system/files/2009CirculationDataReport_1.pdf).
- Fletcher, F. J. (1998). 'Media and political identity: Canada and Quebec in the era of globalization'. *Canadian Journal of Communication*, 23 (3). Retrieved from <http://www.cjc-online.ca/index.php/journal/article/view/1049/955>.
- Freake, R., Gentil, G. & Sheyholislami, J. (2011). 'A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec'. *Discourse & Society*, 22(2), 21-47.
- Gagnon, S. (2003). 'La construction discursive du concept de la souveraineté dans les medias canadiens lors du referendum de 1995'. *Revue québécoise de linguistique*, 32(2): 97-116.
- Halliday, M. A. K. and Matthiessen, C. (2004). *An Introduction to Functional Grammar*. London: Arnold.
- Kuhn, J. & Lick, E. (2009). 'Advertising to Canada's official language groups: A comparative critical discourse analysis'. *Semiotica*, 176(1/4): 165-176.
- Kymlicka, W. (1995). *Multicultural citizenship: A liberal theory of minority rights*. Oxford: Clarendon Press.
- Martin, J. R. & White, P. R. R. (2005). *The language of evaluation. Appraisal in English*. Houndsmills, Basingstoke: Palgrave Macmillan.
- 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.
- Oakes, L. & Warren, J. (2007). *Language, Citizenship and Identity in Quebec*. Basingstoke, England & New York: Palgrave Macmillan.
- Partington, A. (2004). 'Corpora and discourse, a most congruous beast'. In A. Partington, J. Morley & L. Haarman (Eds), *Corpora and Discourse* (pp. 11–20). Bern: Peter Lang.
- Robinson, G. (1998). *Constructing the Quebec referendum: French and English media voices*. Toronto: University of Toronto Press.
- Roy, F. (2009). 'Recent trends in research on the history of the press in Quebec: Towards a cultural history'. In G. Allen and D. J. Robinson (Eds), *Communicating in Canada's Past: essays in media history* (pp. 257-270). Toronto: University of Toronto Press.
- Taras, D. (1993). 'The mass media and political crisis: Reporting Canada's constitutional struggles'. *Canadian Journal of Communication*, 18(2): 131-48.
- Van Dijk, T. A. (1993). *Elite discourse and racism*. Newbury Park, California: Sage Publications.

Van Dijk, T. A. (2006). 'Ideology and discourse analysis'. *Journal of Political Ideologies*, 11(2), 115-140.

Wodak, R., De Cillia, R., Reisigl, M. & Liebhart, K. (2009). *The discursive construction of national identity* (A. Hirsch, R. Mitten Trans.). Edinburgh: Edinburgh University Press.

### **Matteo Fuoli (Università di Trento, Italy)**

Assessing ethical performance: a quantitative analysis of Appraisal in BP and Ikea's Sustainability Reports

Saturated markets, heightened competition and the emergence of new forms of critical consumption compel multinational corporations to invest increasing resources in the implementation and promotion of principles of ethical business. In 'sustainability reports', companies account for and assess their performance across the 'triple bottom line' (environment, society, profit).

Despite the wealth of research on evaluation (see Hunston and Thompson 2000), few studies have concerned the genres of business communication (e.g. Malavasi 2007, 2008). The present work aims at partially filling this gap by applying the Appraisal theory (Martin 1995, 2000; Martin and White 2005; Macken-Horarik and Martin 2003; White 2001) to the analysis of evaluation in a small corpus, comprised of BP and Ikea's 2009 sustainability reports (total word count: approx. 55000 tokens).

Based on the assumption that evaluation plays a fundamental role in the rhetorical 'texturing' of social and institutional identities (Fairclough 2003), the analysis aims to show how these two companies use evaluative resources to represent themselves and to construe their relationship with their stakeholders.

The analysis is quantitative and focuses on the Appraisal systems of Attitude and Engagement, the former concerning the linguistic expression of affect and attitudes, the latter encompassing a wide range of resources that have been studied under the headings of 'evidentiality' (Chafe and Nichols 1986), 'hedging' (Hyland 1996), 'modality' (Hoye 1997, Palmer 1986).

The analysis of Attitude is based on the manual annotation and categorization of instances. In light of the degree of subjectivity which is involved in this process (Hunston 2004), an inter-coder agreement test on a sample from the corpus was carried out. The test yielded a chance-corrected coefficient of  $k = 0,62$  (Cohen 1960), which indicates a substantial level of agreement and can be thus taken as a positive indicator of the reliability of identification and quantification of Attitude in the corpus.

The analysis of Engagement has been performed using an automatic procedure for the quantification of 'markers' of Engagement. Engagement lends itself better than Attitude to software applications, as it is possible to identify in advance a circumscribed set of resources that can be searched for and quantified. For this analysis, I have assembled two collections of potential markers of Engagement, created adapting and integrating the lists of 'stance markers' provided in Biber and Finegan (1989).

The analysis highlights substantial differences in the use of Appraisal in the two reports. BP deploys attitudinal language to strongly foreground its technical capabilities and expertise. Ikea displays affect, emphasizes 'improvements' and hedges propositions more frequently than its counterpart. This different use of evaluation construes two very different institutional and corporate identities, which can be read and decoded in view of the specific features of the contexts in which the two companies operate and of the sustainability challenges they have to face in their daily operations.

Biber, D. and Finegan, E. (1988). 'Adverbial stance types in English'. *Discourse Processes* 11, 1-34.

Biber, D., Finegan, E. (1989). 'Styles of stance in English: Lexical and grammatical marking of evidentiality and affect'. *Text* 9, 93-124.

Cohen, J. (1960). 'A coefficient of agreement for nominal scales'. *Educational and Psychological Measurement* 20, 37-46.

Chafe, W. and Nichols, J. (eds.). (1986). *Evidentiality: the Linguistic Coding of Epistemology*. Norwood, N.J.: Ablex.

Fairclough, N. (2003). *Analysing Discourse*. London and New York, Routledge.

Hunston, S. and Thompson, G. (eds.). (2000). *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford, OUP.

Hunston, S. (2004). 'Counting the uncountable: problems of identifying evaluation in a text and in a corpus'. In Partington, M. and Haarman, L. (eds.), *Corpora and Discourse*. Peter Lang, 157-188.

Hyland, K. (1996). 'Writing Without Conviction: Hedging in Science Research Articles'. *Applied Linguistics* 17, 433-54.

Macken-Horarik, M., and Martin, J.R., (eds.). (2003). Text 23. Special Issue. *Negotiating Heteroglossia: Social Perspectives on Evaluation*. Berlin and New York, Mouton de Gruyter.

Malavasi, D. (2007). 'Lexical analysis of implicit promotional devices in Bank Annual Reports'. *Les Cahiers de ILCEA* numéro 9, 171-184.

Malavasi, D. (2008). 'Banks Annual Reports: an overview of the linguistic means used to express evaluation'. In Garzone, G. and P. Catenaccio (eds.), *Language and Bias in Specialized Discourse*. Milano, CUEM, 139-152.

Martin, J. R. (1995). 'Reading Positions/Positioning Readers: JUDGEMENT in English'. *Prospect: a Journal of Australian TESOL* 10 (2), 27-37.

Martin, J.R. (2000). 'Beyond Exchange: APPRAISAL Systems in English'. In Hunston, S. and Thompson, G. (eds.), *Evaluation in Text*. Oxford, Oxford University Press, 142-75.

Martin, J. R. and White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. New York and London: Palgrave.

Palmer, F. R. (1986). *Mood and Modality*. Cambridge: Cambridge University Press.

White, P. (2001). *An introductory tour through appraisal theory*.  
<http://www.grammatics.com/appraisal/AppraisalGuide/Framed/Appraisal-Overview.htm>

*If*-conditionals and modality: A corpus-based investigation

A number of studies on modality and/or conditionals have presented the claim that conditionals are intimately connected to modality (Comrie, 1986: 89; Dancygier, 1998: 72; Huddleston & Pullum, 2002: 741; Nuyts, 2001: 352; Palmer, 1986: 189; Sweetser, 1990: 141); however, the nature of that connection has not been investigated empirically. This paper reports on parts of a larger study which empirically tested the above claim – namely the corpus-based approach and metrics developed in the study, as well as some significant findings.

More specifically, the paper examines whether, and to what extent, this relationship ...

- a. holds for all conditionals, irrespective of their subordinator (e.g. *if*, *in case*, *unless*);
- b. extends to concessive-conditionals (e.g. *even if*);
- c. is limited to conditionals (and concessive-conditionals), or extends to other constructions sharing subordinators with conditionals (e.g. indirect interrogatives with *if*).
- d. holds for either of the two parts of bi-partite constructions (e.g. protasis and apodosis in conditionals; Fillmore, 1986).

In the case of *if*-conditionals, the paper also examines the extent to which this relationship applies to their two basic types, direct and indirect (Quirk et al., 1985: 1088-1097).

The methodology combines a corpus-based, quantitative approach with close analysis of the data for the purposes of the annotation of modal marking in all corpus samples, and the classification of *if*-conditionals. The study uses eleven random samples from the written BNC, containing the following:

- a. All types of constructions, providing an indication of the average frequency of modal marking in written British English –which was used as the baseline;
- b. Non-conditional constructions, taken collectively;
- c. Conditional constructions (e.g. *assuming*, *if*, *unless*);
- d. Conditional-concessive constructions with *even if* and *whether*;
- e. Indirect interrogative (non-conditional) constructions with *if* and *whether*;
- f. Constructions with *when* and *whenever* (used as conjunctions), as they have been presented as synonymous with unmodalised *if*-conditionals in some studies (e.g. Athanasiadou & Dirven, 1996: 617, 1997: 62; Palmer, 1990: 174-175).

The analysis revealed that the modal load (i.e. the extent of modal marking) in conditionals as a construction family, and *if*-conditionals in particular, is significantly higher than the baseline and non-conditional constructions (taken collectively), as well as most, but not all, non-conditional constructions. More importantly, *if*-conditionals showed a distinctly higher modal load than other conditional constructions. Overall, constructions of the same family tend to have similar modal load; however, this is not consistently the case with individual constructions within a family. Also, constructions across and within bi-partite families show different ratios of modal load in their two parts. More importantly, the protases of *if*-conditionals have a modal load at least equal to that of

the baseline, and, in some cases, significantly higher – despite protases being already modally marked by *if*.

Athanasiadou, A. & Dirven, R. (1996). 'Typology of *if*-clauses'. In Casad, E.H. (Ed.), *Cognitive Linguistics in the Redwoods: The expansion of a new paradigm in linguistics*. Cognitive Linguistics Research 6 (pp. 609-654). Berlin: Mouton de Gruyter.

Athanasiadou, A. & Dirven, R. (1997). 'Conditionality, hypotheticality, counterfactuality'. In Athanasiadou, A. & Dirven, R. (Eds.), *On Conditionals Again* (pp. 61–96). Amsterdam: John Benjamins.

Comrie, B. (1986) 'Conditionals: A typology'. In Traugott, E.C., Meulen, A., Reilly, J.S. & Ferguson, C.A. (Eds.), *On Conditionals* (pp. 77-99). Cambridge: Cambridge University Press.

Dancygier, B. (1998). *Conditionals and Prediction: Time, knowledge and causation in conditional constructions*. Cambridge: Cambridge University Press.

Fillmore, C.J. (1986). 'Varieties of conditional sentences'. *Eastern States Conference on Linguistics*, Vol. 3, 163-182.

Huddleston, R. & Pullum, G.K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Nuyts, J. (2001). *Epistemic Modality, Language, and Conceptualization*. Amsterdam: John Benjamins.

Palmer, F.R. (1986). *Mood and Modality*. Cambridge: Cambridge University Press.

Palmer, F.R. (1990). *Modality and the English Modals* (2nd edn). Cambridge: Cambridge University Press.

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

Sweetser, E.E. (1990). *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.

**Federico Gaspari (University of Bologna, Italy)**

Combining the analysis of lexical bundles and POS-n-grams: a phraseological comparison of the BNC and ukWaC

Several studies have looked into the phraseology of English by focusing on lexical bundles, i.e. sequences of words that occur a certain number of times in a corpus, regardless of their structural completeness or function (e.g. Biber, 2006; Biber & Conrad, 1999; Cortes, 2004; Hyland, 2008; Partington & Morley, 2004; Stubbs, 2007). A complementary strand of research has considered the syntactic-grammatical configuration of English phraseology looking at POS-n-grams, i.e. complexes formed by strings of specific grammatical categories (e.g. William Fletcher's "Phrases in English" database derived from the BNC supports the search for patterns of between 1 and 8 POS tags - see Fletcher, 2007). Surprisingly, however, only very little research has investigated the combination of these two interrelated levels of phraseological analysis as a means of describing and comparing corpora in the same language, one exception being Bernardini et al. (2010:36ff) who examine the differences between native and non-native institutional academic English.

This paper presents an analysis of lexical bundles and POS-n-grams in the BNC (Aston & Burnard,

1998) and ukWaC (Ferraresi et al., 2008) as a basis to contrast the phraseological make-up of the two corpora. ukWaC is a web-derived corpus of English containing more than 2 billion tokens, which has already been compared to the BNC by Baroni et al. (2009) in terms of lexical coverage, text types and subject matters, and by Ferraresi et al. (2008) with a qualitative study of salient nouns, verbs and adjectives. In addition, Sharoff (2006) built web-derived corpora of a similar size to the BNC for a number of languages, using the BNC as a benchmark to evaluate the composition of the English corpus based on text typology and word lists. Finally, Ferraresi et al. (2010) showed that ukWaC and the BNC were similarly helpful in performing lexicographic tasks aimed at dictionary compilation.

Following a discussion of the appropriate length and frequency cut-off point for lexical bundles and POS-n-grams, the paper contrasts these two dimensions of phraseology in the BNC vs. ukWaC. The investigation looks at the most common lexical bundles vis-a-vis the high-frequency POS-n-grams within each of the two corpora, comparing the extent to which the retrieved constructs overlap. The conclusion points out the main phraseological similarities and differences between the BNC and ukWaC, providing new insights into the likeness of these two corpora, which is of interest to researchers using them for a range of theoretical and applied purposes. Finally, we discuss from a methodological perspective the advantages and the potential of combining the analysis of lexical bundles and POS-n-grams both to describe the phraseology of individual corpora and to compare phraseological features across corpora in the same language.

Aston, G. & L. Burnard (1998) 'The BNC Handbook: Exploring the British National Corpus with SARA'. Edinburgh: Edinburgh University Press.

Baroni, M, S. Bernardini, A. Ferraresi & E. Zanchetta (2009) "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". 'Language Resources and Evaluation' 43(3):209-226.

Bernardini, S., A. Ferraresi & F. Gaspari (2010) "Institutional academic English in the European context: a web-as-corpus approach to comparing native and non-native language". L. López, Á. & R. Crespo Jiménez (eds) 'Professional English in the European context: The EHEA challenge'. Bern: Peter Lang. 27-53.

Biber, D. (2006) 'University language: a corpus-based study of spoken and written registers'. Amsterdam: John Benjamins.

Biber, D. & S. Conrad (1999) "Lexical bundles in conversation and academic prose". H. Hasselgård and S. Oksefjell (eds) 'Out of corpora: Studies in honor of Stig Johansson'. Amsterdam: Rodopi. 181-189.

Cortes, V. (2004) "Lexical bundles in published and student disciplinary writing: Examples from history and biology". 'English for Specific Purposes' 23:397-423.

Ferraresi, A., S. Bernardini, G. Picci & M. Baroni (2010) "Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation". R. Xiao (ed) 'Using Corpora in Contrastive and Translation Studies'. Newcastle: Cambridge Scholars Publishing. 337-359.

Ferraresi, A., E. Zanchetta, M. Baroni & S. Bernardini (2008) "Introducing and evaluating ukWaC, a very large Web-derived corpus of English". S. Evert, A. Kilgarriff and S. Sharoff (eds) 'Proceedings of the 4th Web as Corpus Workshop - Can we beat Google? (WAC-4) LREC 2008'. Marrakech, Morocco, 1 May 2008. 47-54.

Fletcher, William H. (2007) "Implementing a BNC-Compare-able Web Corpus". 'Proceedings of the 3rd web as corpus workshop'. Louvain-la-Neuve, Belgium, 15-16 September 2007. 43-56.

Hyland, K. (2008) "As can be seen: Lexical bundles and disciplinary variation". 'English for Specific Purposes' 27: 4-21.

Partington, A. & J. Morley (2004) "From frequency to ideology: Investigating word and cluster/bundle frequency in political debate". B. Lewandowska-Tomaszczyk (ed) 'Practical Applications in Language and Computers – PALC 2003'. Frankfurt am Main: Peter Lang. 170-192.

Sharoff, S. (2006) "Creating general-purpose corpora using automated search engine queries". M. Baroni & S. Bernardini (eds) 'Wacky! Working papers on the Web as Corpus'. Bologna: Gedit. 63–98.

Stubbs, M. (2007) "An Example of Frequent English Phraseology: Distribution, Structures and Functions". Facchinetti, R. (ed) 'Corpus Linguistics Twenty-five Years On'. Amsterdam: Rodopi. 89-105.

**Davide Simone Giannoni (University of Bergamo, Italy)**

From 'Our Methods Apply Equally Well' to 'The Model Does a Very Poor Job': A Corpus-Based Study of Academic Value-Marking

Disciplinary writing reflects and reinforces values whose investigation can yield evidence of the qualities or aspects of reality regarded as variously desirable by scholars, as they (re)negotiate knowledge claims through texts that signal how "the values of academic communities are articulated in discourse meanings" (Hyland 1997: 20). This means that successful academic communication depends on the participants' ability to establish a common ground embracing shared perceptions of what is good or bad, suitable or unsuitable, affective or ineffective, etc. in a given discipline.

While the range of speech acts expressing evaluation in academia has been extensively researched, little is known of the underlying values they encode. This paper illustrates the methods employed in a recent study of academic writing based on a 1 m word corpus of English research articles from ten disciplinary areas (anthropology, biology, computer science, economics, engineering, history, mathematics, medicine, physics and sociology). Using candidate items with 100+ wordlist occurrences, the most prominent value-marking categories were investigated through a combination of automated and manual procedures. Four of these categories (relevance, size, novelty and goodness) were then singled out for further investigation, based on lexical sets and word groups including synonyms, antonyms and different parts of speech. The results show how each disciplinary culture draws on a common repertoire of conventional, largely unqualified axiological meanings instrumental to the advancement of knowledge in its field.

This approach is consistent with earlier efforts (e.g. Biber et al. 2004) combining corpus-linguistic and discourse-analytic methodologies for the analysis of lexical variation in academic genres. While the detailed findings are reported in Giannoni (2010), this paper highlights some of the challenges inherent in its methodology: the use of conventional criteria for selecting and grouping disciplinary fields; the difficult balance between automated text processing and manual investigation; the difference between evaluative speech acts and parameters of value; insights and directions for further corpus-oriented work. Despite its mundane appearance, the significance of such data should not be underestimated, bearing in mind that "every act of evaluation expresses a communal value-system, and [...] this value-system in turn is a component of the ideology which lies behind every text" (Thompson & Hunston 2000: 6).

Biber, Douglas / Csomay, Eniko / Jones, James K. / Keck, Casey 2004. 'A Corpus Linguistic Investigation of Vocabulary-Based Discourse Units in University Registers'. In Connor, Ulla / Upton, Thomas A. (eds) *Applied Corpus Linguistics. A Multidimensional Perspective*. Amsterdam: Rodopi, 53-72.

Giannoni, Davide Simone. 2010. *Mapping Academic Values in the Disciplines: A Corpus-Based Approach*. Bern: Peter Lang.

Hyland, Ken 1997. 'Scientific Claims and Community Values: Articulating an Academic Culture'. *Language and Communication* 17/1, 19-31.

Thompson, Geoff / Hunston, Susan. 2000. 'Evaluation: An Introduction'. In Hunston, Susan / Thompson, Geoff (eds) *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 1-27.

**Gaëtanelle Gilquin (F.N.R.S. – Centre for English Corpus Linguistics, University of Louvain) and Sylviane Granger (Centre for English Corpus Linguistics, University of Louvain)**

The use of discourse markers in corpora of native and learner speech: from aggregate to individual data

Traditionally, corpora have been treated by corpus linguists as 'one big text', in which "the data obtained from (...) different speakers or writers are pooled" (Rietveld et al. 2004: 350). This approach relies on the assumption that well-sampled corpora allow for reliable and valid generalisations about a population as a whole (see Kennedy 1998: 74). However, the fact that "corpora are inherently variable internally" (Gries 2006: 110) suggests that, while generalisations about populations may still be valid and useful, interesting findings are also likely to emerge if we investigate corpus data as a series of individual texts rather than as an aggregate. This is particularly true of learner corpora, because of the "highly heterogeneous nature of learner language" (Granger et al. 2009: 3; see also Durrant & Schmitt 2009: 168).

Following recent studies like Paquot (2010) which take account of the possible variance within a corpus, we will adopt both a global and individual approach to the study of discourse markers in native and non-native speech. Our data will come from the newly published Louvain International Database of Spoken English Interlanguage (LINDSEI, see Gilquin et al. 2010) and its native counterpart, the Louvain Corpus of Native English Conversation (LOCNEC, see De Cock 2004). Starting from the whole of LINDSEI and LOCNEC, we will show how the use and frequency of discourse markers such as 'you know' or 'I mean' differ in native and non-native English. This level of analysis reveals, for instance, an underuse of 'sort of' in non-native English as compared to native English. At the next level of analysis, we will make a distinction between the different learner populations represented in LINDSEI, that is, the groups of learners who share the same mother tongue. This will enable us to highlight features that seem to be transfer-related, such as for example the heavy overuse of 'in fact' in the French component of LINDSEI. Finally, we will consider individual speakers in LINDSEI and LOCNEC in an attempt to identify idiosyncratic features that are limited to just a few speakers. By adopting this threefold level of analysis, we hope to shed new light on the use of discourse markers by native speakers and learners of English and to distinguish between the characteristics that are typical of native or non-native speech in general, those that are limited to certain populations and those that are only found among certain speakers. More generally, we wish to advocate for a combined approach in (learner) corpus research which takes into consideration both the pooled data and the individual texts making up a corpus.

De Cock, S. 2004. 'Preferred sequences of words in NS and NNS speech'. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2: 225-246.

Durrant, P. & N. Schmitt. 2009. 'To what extent do native and non-native writers make use of collocations?' *IRAL* 47: 157-177.

Gilquin, G., S. De Cock & S. Granger. 2010. *The Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (eds). 2009. *International Corpus of Learner English*. Handbook and CD-ROM. Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.

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

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Longman: London & New York.

Paquot, M. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New York: Continuum.

Rietveld, T., R. Van Hout & M. Ernestus. 2004. 'Pitfalls in corpus research'. *Computers and the Humanities* 38: 343-362.

**Gaëtanelle Gilquin (F.N.R.S. – Centre for English Corpus Linguistics, University of Louvain)**

Newspaper editorials and New Englishes: A good match?

Newspapers and magazines have often been used as a source of data in corpus linguistics. In the case of English, corpus linguists have used them to study the 'standard' varieties of English (cf. for example Fries & Lehmann 2006 or Millar 2009), but also, more recently, the new, indigenised varieties of English (e.g. Schmied & Hudson-Ettle 1996, Mukherjee & Hoffmann 2006). Corpora designed to represent a wide cross-section of the language will also, typically, contain a section with samples from newspapers and/or magazines (cf. British National Corpus, International Corpus of English). Besides the pervasiveness of written media in people's everyday life, and hence their obvious impact in linguistic terms, one reason for including this type of data is that usually they can be downloaded very easily from the newspapers' or magazines' websites, which makes it possible to constitute a large corpus in a short time. Using specialised tools such as GlossaNet (<http://glossa.fltr.ucl.ac.be>), it is even possible to automate the daily collection of newspaper articles according to specific criteria (e.g. language, country, newspaper sections, etc). Issues such as editorial policy or copy-editing, however, may make one wonder to what degree the language found in the press reflects the language of the man on the street. This question is all the more important for New Englishes, whose emerging features may not (yet) be recognised to such an extent as to make their way into newspapers or magazines.

In this presentation, I will seek to determine the suitability of newspaper articles, written by professional journalists, to investigate indigenised varieties of English. My analysis will be based on a corpus of editorials and columns in Jamaican English, part of the MULT-ED (Multilingual Editorial) Corpus, and a comparable corpus of readers' comments. The fact that the comments were written in reaction to the editorials and columns included in the MULT-ED Corpus results in similar contents and hence a very high degree of comparability between the two corpora. The phenomenon that will be investigated is that of phrasal verbs, which have been shown to behave differently according to register (Moon 1997: 46), native vs non-native speakers (Siyanova & Schmitt 2007), etc. A comparative analysis of phrasal verbs in the journalists' editorials/columns and in the readers' comments will allow us to identify possible differences between the two and thus judge the reliability of newspaper editorials to shed light on emerging features of New Englishes.

Fries, U. & H. M. Lehmann. 2006. 'The style of 18th century English newspapers: Lexical diversity'. In N. Brownlees (ed.) *News Discourse in Early Modern Britain*, pp. 91-104. Bern: Peter Lang.

Millar, N. 2009. 'Modal verbs in TIME: Frequency changes 1923-2006'. *International Journal of Corpus Linguistics* 14(2): 191-220.

Moon, R. 1997. 'Vocabulary connections: multi-word items in English'. In N. Schmitt & M. McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*, pp. 40-63. Cambridge: Cambridge University Press.

Mukherjee, J. & S. Hoffmann. 2006. 'Describing verb-complementational profiles of New Englishes: A pilot study of Indian English'. *English World-Wide* 27(2): 147-173.

Schmied, J. & D. Hudson-Ettle. 1996. 'Analyzing the style of East African newspapers in English'. *World Englishes* 15(1): 103-113.

Siyanova, A. & N. Schmitt. 2007. 'Native and nonnative use of multi-word vs. one-word verbs'. *IRAL* 45: 119-139.

**Rüdiger Gleim (Goethe University), Armin Hoenen (University of Frankfurt), Nils Diewald, Alexander Mehler and Alexandra Ernst**

Modeling, Building and Maintaining Lexica for Corpus Linguistic Studies by Example of Late Latin

Many corpus linguistic studies depend on powerful lexical resources to support automatic lemmatization, part-of-speech-tagging and related tasks of preprocessing corpus data. This is all the more relevant when long-term historical corpora are analyzed that pose additional challenges because of temporal change of the underlying language. An example of such a corpus is the *Patrologia Latina* (Migne, 1865) that covers documents from the 4th to the beginning 13th century. It unfolds several stages of the development of Late Latin in the direction of Early Romance on various levels of linguistic resolution (Clackson and Horrocks, 2007).

Creating and maintaining adequate lexical resources to preprocess and analyze such historical corpora is a challenging task which relates to linguistic, informational and methodical aspects. From a linguistic perspective, the usage of lexical resources for corpus linguistic studies hinges upon the availability of an underlying data model that is flexible enough in several respects: first of all, the model must reflect morpho-syntactic specifics of a range of different languages. Then, the model has to account for derivations, compositions and other word relations. This is required, among other things, for multi-word units up to the level of collocations as considered in corpus linguistic studies. Furthermore, the data model should account for the change of the expression and content plane of lexical units as well as for the change of their grammatical usage in order to reflect temporal dynamics of the lexicon.

These linguistic requirements give an impression of the complexity the data model has to deal with from an informational perspective. The typology of entities and relations can hardly be cast into a fixed form but need to be captured by an extensible design. Finally, this complexity has to be kept manageable by offering a proper interface and by supporting interoperability with established formats (e.g., TEI P5 TEI Consortium (2010) or RDF W3C (2010)).

Meeting these requirements, we developed the eLexicon, a data model for lexical resources. It is based on a normalized relational data model to represent a lexicon as a typed hypergraph. Using an extensible typology, this model is flexible enough to accurately model linguistic specifics and to be

adjusted to other languages. We provide a description of the data model in connection with a presentation of our work on lexical resources. Furthermore the paper documents means for data access and interchange, namely an RDF-based representation as well as the eLexicon Browser, which has been developed as part of the eHumanities Desktop (Gleim and Mehler, 2010). Using a web interface, it mediates between the complexity of the data model and its usability. We will demonstrate the eLexicon in the context of corpus linguistic studies by means of our work on a lexical resource of Late Latin and the *Patrologia Latina*, which is, one of the largest corpora of this period.

Clackson, J. and Horrocks, G. (2007). *The Blackwell History of the Latin Language*. Blackwell Publishing Ltd, 2 edition.

Gleim, R. and Mehler, A. (2010). 'Computational linguistics for mere mortals - powerful but easy-to-use linguistic processing for scientists in the humanities'. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta/Malta.

Migne, J. P. (1844-1865). *Patrologiae cursus completus: Series latina*. volume 1-221. Chadwyck-Healey, Cambridge.

TEI Consortium, editor (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, <http://www.tei-c.org/Guidelines/P5/>.

W3C (2010). *Resource description framework (rdf)*. <http://www.w3.org/RDF/>. Last access: May 25, 2011.

#### **Svetlana Gorokhova (St Petersburg State University)**

##### Semantic paraphasias in normal speech: A corpus-based analysis

The paper presents a corpus-based analysis of 1191 semantic paraphasias (cases when a semantically related noun is substituted for the target noun) naturally occurring in Russian normal speech. The speech errors were tape recorded or digitally recorded in everyday conversations, telephone conversations, and live TV and radio programs. I used word frequency measures from Russian National Corpus and word association norms from Russian Word Association Thesaurus to analyse the target-error pairs for target and error word frequencies, target-error co-occurrence strength (measured in Mutual Information and T-scores), and word association norms.

Frequencies of target words were found to be significantly lower than frequencies of their substitutes; besides, there is a very significant positive correlation between target and error frequency values, indicating that target word frequency affects the outcome of the error: to successfully compete with a word to be produced, a semantically related word has to be of a higher frequency. Contrary to the view that the frequency effect is located at the stage of phonological encoding (Jescheniak & Levelt 1994; Jurafsky 2003 etc.), the results suggest that frequency is coded at an earlier stage of lexical selection.

Target words appear to have different characteristics of frequency, length, co-occurrence strength, and associative relatedness depending on whether or not the errors that they elicit belong to the same semantic category as the targets, i.e. whether the target and the error are category coordinates (What beautiful **carnations!** → ...**roses!**; The girl is only three **years** old → ...three **hours**...) or not category coordinates (Our soldiers were dying on the eve of the **victory** → ...of the **war**; The cat keeps tearing the **carpets** → ...the **floors**).

On average, targets that elicit category coordinate errors have significantly higher frequencies and

shorter lengths than targets that elicit non-category coordinate errors. At the same time, multiple regression analyses show that word length is a significant variable which determines the outcome of the error for non-category coordinate target-error pairs but not for category coordinate pairs.

While there is a “length match” between category coordinate targets and substitutes, they have much higher measures of co-occurrence strength and of associative relatedness compared to non-category coordinate target-error pairs. Thus, with higher-frequency nouns that elicit category coordinate errors, it is the speaker’s frequent experience of using the target word together with another (higher-frequency) competing item (resulting in stronger associative links between the two words) rather than item length that is likely to cause a substitution. The finding illustrates a special psychological status of category coordinates in a word’s semantic field.

Jescheniak, J.D., & Levelt, W.J.M. 1994. ‘Word frequency effects in speech production: Retrieval of syntactic information and of phonological form’. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(4): 824-843.

Jurafsky, D. 2003. ‘Probabilistic modeling in psycholinguistics: Linguistic comprehension and production’. *Probabilistic Linguistics*, eds. R.Bod, J.Hay, and S.Jannedy, 39–95. Cambridge, MA: MIT Press.

#### **Lukasz Grabowski (Opole University)**

Translation universals revisited: a corpus-driven comparison of translational and non-translational literary Polish

According to Baker (1993, 1995), corpus-based translation studies should focus, among others, on translation as a product by comparing comparable corpora of translational and non-translational texts. This article presents selected results of a corpus-driven comparison of the two custom-designed corpora of translational and non-translational Polish literary texts. The aim of the study was to find stylistic differences between the two corpora as well as identify traces, if any, of translation universals. Therefore, this study aims to examine T-universals (after Chesterman 2004) with emphasis on core patterns of lexical use, as proposed by Laviosa (1998), on the example of literary Polish.

The study was completed with the use of selected methods offered by corpus linguistics, namely descriptive statistics, distribution of n-letter words, frequency profiles, frequency spectra (after Baroni 2009) and computational stylometry (Classical Delta Distance, after Burrows 1986/2002). To that end, the corpora of translational and non-translational literary Polish were compiled. Each corpus includes approx. 500,000 word tokens and contains a selection of fiction novels published in Poland in the years 1937-2001 (non-translational corpus) and translated into Polish in the years 1950-2001 (translational corpus). Finally, with the use of custom-designed R script (Eder – Rybicki 2011), the texts were processed with the use of a multivariate methods [Principal Components Analysis (PCA) and Cluster Analysis (CA)] and presented graphically in order to illustrate relative distances between the translations and typical Polish novels in terms of 1000 the most frequent words (MFW) in each of the novels included in both corpora.

The results of the study showed that translational literary Polish does not conform to core patterns of lexical use found in typical literary Polish, in particular in terms of lexical variety among top-frequency words, word length and sentence length. Furthermore, PCA and CA showed that translations and typical literary Polish cluster together into two separate groups with considerable distance between them, which provides further evidence as to the existence of the levelling-out translation universal. The article concludes with suggestions and problems as regards research on translation universals with the use of language corpora.

Baker, M. (1993). "Corpus linguistics and Translation Studies: Implications and applications". In M. Baker, G. Francis and E. Tognini-Bonelli (eds.), *Text and Technology. In Honor of John Sinclair*. Amsterdam: John Benjamins, 233-250.

Baker, M. (1995). "Corpora in translation studies: An overview and some suggestions for future research". *Target*, 7(2), 223-243.

Baroni, M. (2009). "Distributions in text". In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: An international handbook Volume 2*. Berlin and New York: Walter de Gruyter, 803-821.

Chesterman, A. (2004). "Beyond the particular". In A. Mauranen and P. Kuyamaki (eds.) *Translation Universals: Do they exist?*. Amsterdam: John Benjamins, 33-49.

Eder, M., Rybicki, J. (2011). "Stylometry with R". In: *Digital Humanities 2011: Conference Abstracts*, Stanford, CA (forthcoming).

Laviosa, S. (1998b). "Core patterns of lexical use in a comparable corpus of English narrative prose". In: *Meta*, 43(4), 557-570.

McEnery, T., R. Xiao and Y. Tono (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Xiao, R. (2010). "How different is translated Chinese from native Chinese? A corpus-based study of translation universals". In: *International Journal of Corpus Linguistics* 15(1), 5-35.

**Gintarė Grigonytė (Vytautas Magnus University), Algirdas Avižienis (Vytautas Magnus University, University of California Los Angeles), and Rūta Marcinkevičienė (Vytautas Magnus University)**

Automatic detection of semantically related lexical items in domain corpora

We present our experimental findings of a new unsupervised approach for the detection of semantically related lexical items by aligning paraphrases in monolingual domain corpora. Two methodologies are applied: automatic extraction and alignment of paraphrases (Cordeiro et al 2007), and automatic synonymy detection in paraphrased corpora (Grigonyte et. al 2010). Both methodologies were applied to two domain corpora: ReSIST corpus in English on computer security and dependability (Avizienis et. al 2008) and ŠIMTAI corpus in Lithuanian on education and research policy. In the English corpus semantically related lexical items were identified with a 67.27% precision.

Semantically related lexical items can be identified on the basis of their context (Evert 2005). However, the drawback of distributional similarity approach is its dependency on the specific context where items have to be frequently used. To overcome this bottleneck we propose to align paraphrases from domain corpora and to detect lexical items that could be mutually replaceable within the aligned context, called paraphrase casts patterns (Grigonytė et. al 2010). The difference from the pattern-based approach is that instead of looking for pattern-alike lexical items, the local general environment, or discourse, is used as some sort of pattern. This methodology is language-independent, it also does not depend on linguistic processing, or manual definition of patterns as well as training sets, therefore it guarantees a higher precision when compared to distributional similarity-based approaches.

The process of the extraction of semantically similar word pairs or phrase pairs is as follows: domain texts – paraphrased sentences – aligned paraphrased corpus – paraphrase cast examples — examples of semantically related lexical items. The processing pipeline starts with unstructured domain texts. All the sentences are compared and if similarities are detected, the alignments are established. The method highlights identical or almost identical sentences, it also takes into account pairs of sentences that have a high degree of lexical reordering.

The proposed approach uses Sumo-metric described by Cordeiro et al. (2007) that outperforms simple N-gram overlap, edit-distance and BLEU metric to calculate the semantic similarity of the sentences aligned. When the paraphrases are detected and similar sentences aligned, specific segments, i.e. paraphrase casts, are explored. With the help of paraphrase casts patterns, single word or multi-word lexical items are extracted. Some of extracted pairs appear to be synonymous (advisory vs. expert institution), others have different semantic relationship i.e. that of hyper- and hyponymy (overall education vs. professional training), equanimy (qualification of a psychologist vs. qualification of a teacher) or antonymy (senior vs. junior researcher). A considerable amount of the identified pairs have one word in common without any other specific semantic relation (a roadmap vs. a final version of the strategy).

Avižienis, A., Čulo, O., Grigonytė, G., Marcinkevičienė, R.: 'Building a Thesaurus and Ontology of the Concepts of Dependability and Security'. In proc. of 37th IEEE/IFIP International Conference on Dependable Systems and Networks, 2007, p. 420-421.

Cordeiro, J.; Dias, G. & Cleuziou, G. 2007. 'Biology Based Alignments of Paraphrases for Sentence Compression'. In Proceedings of the Workshop on Textual Entailment and Paraphrasing (ACL-PASCAL / ACL2007).

Evert, S. 2005. 'The Statistics of Word Co-occurrences: Word Pairs and Collocations', Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Grigonytė, G, Cordeiro, J.P, Moraliyski, R. Dias, G. and Brazdil, P. 'Paraphrase Alignment for Synonym Evidence Discovery', Proceedings of the 23rd Int. Conf. on Computational Linguistics, COLING 2010, p.403-411.

#### **Nicholas Groom and Oliver Mason (both University of Birmingham)**

Disciplinary differences in small-group interactions: quantitative and qualitative perspectives on turn-taking in university seminars

The fundamental aim of seminars and other forms of small-group interaction in higher education is to get students to talk, and the underlying assumption shared by educational theorists and university teachers alike is that the more the students talk, the more successful the seminar is. But what does 'more talk' mean - number of words, number of turns, or average length of turn? In this paper we report on a study of the British Academic Spoken English Corpus (BASE), in which we investigate student and teacher contributions to seminars according to each of these three measures. Our main finding is that different knowledge domains perform better according to different measures. Specifically, our quantitative analysis shows that students talk the most in seminars in the social sciences and humanities if we define talk in terms of total words spoken; students in physical sciences talk the most if we quantify talk in terms of number of turns; and students in life sciences talk the most if we measure talk in terms of average turn length. We then argue against the idea that any one of these measures might be inherently better or more desirable than the others. Drawing on qualitative data from the BASE seminars subcorpus, we argue instead that each of these different versions of 'talking more' carries with it a different set of affordances, each of which is more or less well attuned to the particular epistemologies and pedagogic goals of

different academic disciplines. We conclude by considering the implications of our analysis for staff development and training programmes in higher education.

**Katherine Gupta (University of Nottingham)**

Letters to the Editor and the suffrage movement, 1908-1914

This paper explores the terminology used by suffrage organisations in the Letters to the Editor section of The Times newspaper between 190 and 1914. The study is based on a corpus of 7 million words from the Times Digital Archive and combines a corpus linguistic approach with concerns in critical and historical discourse analysis.

The suffrage movement was not a unified one; rather, it was composed of various groups with differing backgrounds, ideologies and aims, and different terminology used to describe different factions of the movement. My research has indicated that the more inclusive term suffragist was used in conjunction with activities more often associated with suffragette groups; however, it was unclear as to whether it reflected self-identification on the part of suffrage organisations usually associated with the term suffragette (especially the Women's Social and Political Union) or whether it was a result of conflation of the two groups by the newspaper. In this paper I focus on the Letters to the Editor section. These pages offered a platform for public discussion of women's suffrage, universal suffrage and equal franchise, both between suffrage supporters and opponents and between different suffrage organisations. These factors make it a useful point of comparison to hard news reporting. I examine collocates of *suffragist*, *suffragists*, *suffragette*, *suffragettes*, *National Union of Women's Suffrage Societies*, *NUWSS*, *Women's Social and Political Union* and *WSPU* to investigate how different identities and organisations within the suffrage movement were discussed. I then analyse these with a particular focus on contested terminology and identities, such as debates over who is and is not a suffragette or suffragist, the activities associated with each group, what counts as direct action, the types of protest or lobbying they should or should not engage in, and condemnation of activities. These will be compared with the results from my earlier research and historical research to offer greater understanding of the tensions between different suffrage organisations and their development in public correspondence.

Through a flexible and interdisciplinary combination of established historical approaches, discourse analysis and corpus linguistic methodologies, this investigation both refines our understanding of the suffrage movement and its socio-historical context and offers an insight into the structures and language of complex political protest organisations.

**Glenn Hadikin (University of Portsmouth)**

Corpus, Concordance, Koreans: a Comparison of the English Spoken in Two Korean Communities

English language programmes have been running in South Korea since 1883 (Shim 1999) but one rarely sees academic papers discussing the use and development of English in that country. Shim (1999) argues that a form of codified Korean English is even taught in schools and, from personal experience, students will often use forms that are not familiar to most British or American listeners such as 'I want to lose my weight' or 'I will marry next year'. Previous studies have tended to focus on written Korean English (see, for example, Jung and Min 1999) but it is clear that one cannot assume results from studies of written English will hold true for spoken English.

In this paper I will report results from what is, to my knowledge, the first PhD level study of spoken Korean English corpora. With a theoretical foundation based on Hoey's Theory of Lexical Priming (Hoey 2005) and Wray's model of the role of formulaic language in second-language acquisition (Wray 2002) I have collected two corpora of spoken Korean English: one from a Korean community in Liverpool, England and one from Seoul, Korea. I hypothesised that there would be measurable, statistically significant differences between the ways certain high frequency forms are used in each community due to the different priming factors affecting their spoken English.

The results show clear differences between the two groups and also suggest that combined factors such as Korean primings and developing skills such as the use of vague language may be giving rise to new completely structures in the Liverpool community that are novel in both Korean and British English. In the paper I also raise corpus theoretical issues and suggest that an unorthodox interpretation of Sinclair's Idiom Principle (Sinclair 1991) is more compatible with Hoey's theory and the data in my study.

Hoey, M. (2005). *Lexical Priming: a New Theory of Words and Language*. London: Routledge.

Jung, K. And Min, S.J. (1999). 'Some Lexico-grammatical Features of Korean English newspapers'. *World Englishes* 18/1:23-37.

Shim, R.J. (1999). 'Codified Korean English'. *World Englishes* 18/2:247-259.

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

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

**Daniel Gallego Hernández (Universidad de Alicante) and Ramesh Krishnamurthy (Aston University)**

COMENEGO (Corpus Multilingüe de Economía y Negocios): design, creation and applications

This paper will describe the initial stages of the COMENEGO project, which is initially creating comparable corpora of Business texts in Spanish and French. The language of business remains a vital field in global activities, and the globalised market requires frequent cross-border, cross-linguistic, and cross-cultural interaction. The need for rapid and accurate translations places increasing demands on the business community, on translation practitioners, and those who train the translators. High-frequency activities also tend to react, innovate, and adapt more quickly to changes in their environment and practices, and therefore there is a constant need to renew outdated resources. However, this pressure often leads to ad hoc collections. The highly competitive nature of commerce also poses problems of data availability, accessibility, and cost. Translated texts may contain traces of source language interference and other non-native-speaker features, hence especially for advanced and specialised translation purposes, comparable corpora may be more suitable. For Spanish and French, these factors and features are evident in previous corpora. For example, MLCC(\*1) is reasonable in size (10m words for each language) but is now outdated (early 1990s), and lacks textual variety (single sources – Expansion for Spanish and Le Monde for French). IULA(\*2) has variety, but is rather small (1m words of Spanish), and contains many translated texts. CLUVI(\*3) is also smaller (0.4m and 1.8m Spanish), and contains translated texts, single-source subcorpora (Consumer Eroski) and some outdated texts (1998 onwards). Vicente(\*4) is diachronic and contrastive (1995 and 2006), contains only press articles (French: Le Monde, Les Echos; Spanish: El Pais, Expansion), and is not publicly available.

COMENEGO focuses on up-to-date texts, textual variety and balance, and the necessary compromise between more idealised corpus design and practical factors, and will constitute a valuable resource for researchers, translators, and translator trainers and trainees. The first cycle of data collection has acquired 7m words of Spanish and 9m words of French. Copyright permissions are being obtained, and pilot analyses are revealing imbalances and deficiencies which are being addressed. This paper will discuss not only the process of corpus design and creation, but also the applications of comparable corpora in translation pedagogy.

MLCC (\*1) - MLCC Multilingual and Parallel Corpora  
<http://www.elda.org/catalogue/en/text/W0023.html>

IULA (\*2) - Corpus técnico del IULA  
<http://bwananet.iula.upf.edu/indexes.htm>

CLUVI (\*3) - Corpus Lingüístico de la Universidad de Vigo  
<http://sli.uvigo.es/CLUVI/index.html>

Vicente (\*4) - VICENTE, Christian (en prensa): «Lingüística de corpus y traducción especializada: aplicaciones a la traducción francés-español de la economía», XXV CILPR: Congrès international de linguistique et de philologie romanes, Innsbruck, 2007.

**Saman Hina (University of Leeds), Eric Atwell (University of Leeds), and Owen Johnson (University of Leeds)**

Enriching a Healthcare Corpus with SNOMED-CT standard medical semantic tags

Clinical patient records include a mix of structured data and narrative text. We have compiled a corpus containing diverse medical narratives from different healthcare partners, and are enriching it with concept-tags from the international standard SNOMED-CT Systematized Nomenclature of Medicine - Clinical Terms ontology. SNOMED CT provides a concept-based classification of clinical terms (Elkin et al. 2006). This paper describes the different partially-structured data sets including doctor's progress notes, hospital discharge summaries, verbal autopsies, and synthetic data from undergraduate and postgraduate medical training. The semi-structured data contains noise and inconsistencies at several levels, such as variable spelling, punctuation and abbreviations of clinical terms, through to wide variations in document structure.

We have developed a prototype medical semantic tagger to annotate the text with semantic tags from the healthcare data standard SNOMED-CT, advocated by the US College of American Pathologists and UK National Health Service to store, retrieve, and standardize clinical information (IHTSDO 2010). The SNOMED-CT ontology includes over 300,000 medical concepts with some very fine-grained distinctions; for example, some of the complex multiword concepts present in SNOMED CT are;

Structure of intervertebral foramen of fifth thoracic vertebra.

Entire pterygoid process of sphenoid bone.

Diagnostic radiography of sacrococcygeal joint.

Structure of venous plexus of the hypoglossal canal.

Structure of posterior temporal diploic vein.

Entire intervertebral disc space of seventh cervical vertebra.

Aliphatic carboxylic acid, C10-C26

Black locks, oculocutaneous albinism, AND deafness of the sensorineural type

The challenge with this ongoing research is to use this data standard in order to extract broader semantic types from the noisy corpus of medical narratives. Previous researchers have investigated biomedical text corpora by parsing it with English language parsers and have improved the accuracy of automatic processing of biomedical texts by simplification of the sentences (Jonnalagadda et al.

2009). Our systems aim to evaluate the complexity of the corpus as well as the SNOMED CT data standard. We have devised a broader medical semantic tag-set, including medical semantic classes such as finding, substance, disorders, sports injury, situation, event etc derived from the SNOMED-CT ontology hierarchies.

For corpus quality assurance, we use GATE - General Architecture for Text Engineering (Cunningham et al. 1997). Our medical tagger will help language and biomedical researchers to annotate clinical texts with authenticated semantic types without relying on annotation guidelines drawn up by domain experts. Our tagger will also help in secure information extraction from clinical documents (Hina et al. 2010) .

Cunningham, H., K. Humphreys, et al. (1997). "GATE - a TIPSTER-based General Architecture for Text Engineering. In the TIPSTER Text Program (Phase III) 6 Month Workshop."

Elkin, P. L., S. H. Brown, et al. (2006). "Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists." *Mayo Clinic Proceedings* 81(6): 741-748.

Hina, S., Atwell. E., et al. (2010). "Secure Information Extraction from Clinical Documents Using SNOMED CT Gazetteer and Natural Language Processing" *Proceedings of The 5th International Conference for Internet Technology and Secured Transactions ICITST-2010*

Jonnalagadda, S., L. Tari, et al. (2009). Towards effective sentence simplification for automatic processing of biomedical text. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Boulder, Colorado, Association for Computational Linguistics.

**Lydia-Mai Ho-Dac, Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle, and Ludovic Tanguy CLLE-ERSS (CNRS & Université de Toulouse)**

High-level discourse structures: topical chains and enumerative structures in a diversified annotated corpus

One of the outcomes of the ANNODIS project (Ho-Dac et al 2009, 2010) is a diversified corpus annotated with two frequent textual motifs: topical chains – TCs – and enumerative structures – ESs. The corpus has been manually annotated with both the motifs and the clues signalling them. These data can now be exploited in a comparative mode in order to examine TCs and ESs in the three sub-corpora: 1) reports in the field of international relations; 2) scientific articles (proceedings of a linguistics conference); 3) encyclopaedia articles (from Wikipedia).

The initial step is to take a quantitative look at each motif: composition, distribution, and match with document structure (Power et al 2003). Though the motifs are common in all three corpora, differences appear in their frequency, in their length and coverage (proportion of text involved), in their composition (for ESs: number of items, presence of a trigger and/or closure). Another important aspect is their granularity: this notion is approximated via a typology in which types correspond to different forms of interaction between the motifs and the document's layout structure (sections and headings, formatted lists and paragraphs).

We then examine the data from several qualitative angles in order to arrive at a functional characterisation of the motifs. Of special interest to us is the link between particular forms of signalling and specific functions: ESs with items introduced by sequencers, for instance, are functionally different from ESs whose items are introduced by circumstantial adverbials. A continuum is proposed from ESs signalled by purely textual cues (e.g. bullet points) to ESs whose cues carry ideational contents (such as adverbials) (Halliday 1977). The different corpora are

compared in terms of the functional classes and their linguistic correlates, as illustrated by the table below:

Corpus	Number of ESs	ESs/text	Coverage (% of text)	Granularity			
				sections	formatted lists	multi-paragraph	intra-paragraph
WIKI	332	13.28	53.77%	19,28%	39,16%	20,78%	20,78%
CMLF	263	11.95	47.91%	9,13%	23,19%	26,62%	41,06%
GEOPO	234	9	28.73%	6,84%	10,26%	20,94%	61,97%
mean		11.36	43.09%	12,55%	25,93%	22,68%	38,84%

Table 1: Frequency and inter-corpus variations for enumerative structures (ESs)

Finally, the two motifs are observed in context and in their interaction. A special case of interaction concerns ESs interacting with themselves via recursivity, a remarkably frequent occurrence in our corpus. This analysis of how the motifs behave in text also leads to cross-corpus comparisons.

Halliday, M.A.K. (1977/2003) 'Text as semantic choice in social contexts'. In *The Collected Works of M.A.K. Halliday (Volume 2): Linguistic Studies of Text and Discourse*, Jonathan Webster (ed.), 23–81. London: Continuum.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J. & Tanguy, L. (2009) 'A top-down approach to discourse-level annotation', *Corpus Linguistics Conference 2009*, 20-23 July, Liverpool, UK.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P. & Rebeyrolle, J. (2010). 'On the signalling of multi-level discourse structures', *MAD 2010 : Multidisciplinary Perspectives on Signalling Text Organisation*, Moissac (France) 17-20 mars 2010, 94-105.

Power, R., Scott, D. & Bouayad-Agha, N. (2003). 'Document Structure'. *Computational Linguistics* 29(2), 211-260.

#### **Lan-fen Huang (University of Birmingham)**

##### The selection of corpora for the investigation of discourse markers in learner English

The availability of learner corpora has attracted research attention to the area of comparative studies between learners and native speakers (NSs) and between learners with different first language backgrounds (e.g. studies in Granger (ed.) (1998), Granger, Hung and Petch-Tyson (eds.) (2002) and Meunier and Granger (eds.) (2008)). Despite these demonstrations of the increasing interest and new approaches to learner language, there is little discussion about the comparability of corpora and difficulties in comparing corpora. This paper will seek to address some questions raised in comparative studies of learner corpora as well as the selection of corpora for the investigation of discourse markers.

Since learners in the environment of English as a foreign language do not talk in English in everyday life, it is very difficult to collect naturally-occurring speech produced by learners. Therefore, most learner corpora of spoken English consist of contrived data, which are elicited in a rather restricted

context. When a learner corpus is used for comparative studies of learners' and NSs' speech, it is difficult to obtain a truly comparable NS corpus. Although it seemed appropriate to recruit NSs to do the same tasks as learners had done for the compilation of a corpus, whether or not NSs are trained to take an oral exam in their first language and whether or not the context is properly duplicated are open to doubt. If the so-called comparable NS corpus can be carefully designed and compiled, this raises another question about the 'un-naturalness' of the elicited NS speech.

The Spoken English Corpus of Chinese Learners (SECCL) is used to investigate discourse markers. I set out some arguments in this paper against compiling a comparable NS corpus. It is difficult to ensure comparable conditions with respect to such factors as exam-oriented and it is also challenging to compile a corpus of NSs' speech with similar size and number of participants. The compromise that is chosen is comparing the uses of discourse markers in the learners' spoken English in SECCL with those in the NS speech in MICASE and ICE-GB. Also, I argue against using the terms 'underuse' and 'overuse', which seem to assume NSs' use of discourse markers is the target norm for learners as well as suggesting learners' lack of competence. Instead, the neutral terms 'under-represent' and 'over-represent' are used.

This paper will discuss if I can legitimately compare the use of discourse markers in SECCL, MICASE and ICE-GB and how I overcome the problem of the issue of comparability of corpora as well as the difficulties in investigating discourse markers across three different corpora.

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

Granger, S. 2002. 'A bird's-eye view of learner corpus research' in Granger, S., J. Hung and S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Meunier, F. and S. Granger. (eds.) 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

#### **Daniel Hunt (University of Nottingham)**

'Remember the E.D. is talking right now. It's not reality': Using corpora to explore discourse and identity in an online anorexia forum

This paper examines the discursive construction of anorexia nervosa in a specialised corpus of postings to an online eating disorders community forum. As well as providing a rich source of data for illuminating experiences of this often fatal condition, the text-based nature of health fora and the volume of language data they produce make them ideally suited to investigation through corpus techniques. Accommodating analytical concerns from both corpus linguistics and critical discourse analysis (CDA), examination of the corpus is conducted initially through quantitative keyword and frequency calculations. I then offer a qualitative analysis of a selection of these keywords, arguing that collocation analysis offers a riposte to criticisms of methodological inconsistency in CDA research (Wooffitt, 2005) by providing a rigorous means for the identification of particular discursive formations across larger datasets (Baker, 2006; Baker *et al.*, 2008).

The corpus-based discourse analysis of these keywords suggests that the forum users' linguistic choices discursively construct a conception of anorexia that broadly reflects established discourses of pathology and patient-hood in the medical sciences. However, corpus analysis techniques also help uncover a wealth of evidence to demonstrate that this medical discourse is inflected with non-scientific notions of morality and an emergent representation of anorexia as a single object that far exceeds purely medical definitions. This linguistic evidence is used to elucidate a number of wider practices for the forum participants, including establishing a sense of shared group identity with

geographically disparate individuals whilst constructing – and sometimes forcefully defending – a medically-validated social position in the face of a widely stigmatised condition. In relating textual data to these broader forum activities, I attempt to demonstrate that corpus linguistics’ methodological precepts make it a valuable vehicle for interrogating online health communication, particularly given the climate of evidence based practice in the British NHS. At the same time, I contend that corpus investigation can be fruitfully supplemented with additional relevant frameworks, such as communities of practice (Wenger, 1998), that further illuminate the textual evidence whilst being less amenable to corpus approaches alone.

Online fora now constitute a significant source of information and emotional support for large numbers of individuals with chronic conditions and their impact on mainstream healthcare practice is increasingly recognised (Bartlett & Coulson, 2011; Cross, 2008). In light of this, and congruent with the interventionist tradition of CDA, I conclude by outlining future directions for this corpus-enabled research and its mediated application in practicing healthcare contexts.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P, Gabrielatos, C., KhosraviNik, M., Kryżanowski, 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 and Society*, 19 (3), 273-306.

Bartlett, Y.K. and Coulson, N. (2011). ‘An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication’. *Patient Education and Counselling*, 83, 113-119.

Cross, M. (2008). ‘How the internet is changing health care’. *British Medical Journal*, 337, 202-203.

Wenger, E. (1998) *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Wooffitt, R. (2005). *Conversational Analysis and Discourse Analysis*. London: Sage.

#### **Sally Hunt (Rhodes University, South Africa)**

Why Hermione is not the Hero: Using Corpus Methods to Analyse the Discoursal Representation of Female Physicality in Children’s Literature

Children’s literature has frequently been analysed in terms of the representation of female characters, typically comparing them quantitatively with the male characters in terms of their relative frequency, centrality and occupation (cf. Diekman and Murnen (2004), Evans and Davies (2000), Gooden and Gooden (2001), Kortenhuis and Demarest (1993), Wharton (2005)). More recent studies have harnessed corpus methods, notably Knowles & Malmkjaer (1996) and Thompson and Sealey (2007) who used concordancers to explore corpora of children’s literature with a view to a more ideological analysis. Although the philosophical underpinnings may vary, research into children’s literature is based on the assumption that it is important in the developing identities of the readers, being “inevitably suffused with ideology” (ibid. 3), and has consistently shown the representation of characters to be strongly gendered (Wharton 2005).

In this paper I will report on my research into the discursive construction of physical femininity and masculinity in the Harry Potter series, via an analysis of body parts, and explicitly combining corpus methods with a critical discourse approach. Jeffries (2007) and Motschenbacher (2009) provide useful analyses of the representation of adult female body parts in magazines, both concluding that

they are linked to dominant gender discourses. Baker (2006, 2008), Baker et al. (2008) and Mautner (2009, 2009a) demonstrate the advantages of the synergistic relationship between corpus linguistics and the critical analysis of discourse, notably that the use of corpus linguistics as a data collection method counters charges of “cherry-picking” frequently levelled against CDA. Based on a corpus-CDA analysis of three of the Harry Potter novels, I will show that the concept of discourse prosody is particularly helpful in revealing how the use of body parts by characters to express emotion and act agentively on the world is gendered in the series, with female characters being far more likely to be associated with emotion, and even being rendered unable to act on the physical world as a result of overwhelming emotion. The females’ interaction with other characters and objects in the world, and, in particular, their response to danger, suggest stereotyped discourses of inequality which see women and girls as requiring protection and being physically incapable. Thus gender is still a particularly salient aspect in this widely-read example of modern children’s literature, despite plots which appear to be fairly positive towards women. The strength of a corpus approach in this study lies in its capacity to reveal objective, and often otherwise fairly covert, trends in language use which enrich the analysis of discourses in these influential texts.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, Paul. 2008. *Sexed Texts: Language, Gender and Sexuality*. London, Oakville: Equinox.

Baker, Paul, Costas Gabrielatos, Majid Khosravini, Michal Krzyzanowski, Tony McEnery and Ruth Wodak. 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* Vol. 19(3) pp. 273-306.

Diekman, Amanda B. and Murnen, Sarah K. 2004. ‘Learning to be Little Women and Little Men: The inequitable Gender Equality of Nonsexist Children’s Literature’. In *Sex Roles* Vol. 50(5/6) pp. 373-385.

Evans, L. and Davies, K. 2000. ‘No Sissy Boys Here: A Content Analysis of the Representation of Masculinity in Elementary School Reading Textbooks’. *Sex Roles* Vol. 42(3/4) pp. 255-270.

Gooden, A.M. and Gooden, M.A. 2001. ‘Gender Representation in Notable Children’s Picture Books: 1995–1999’ *Sex Roles*, Vol. 45(1/2) pp. 89-101.

Jeffries, Lesley. 2007. *Textual Construction of the Female Body: A Critical Discourse Approach*. Palgrave Macmillan, Basingstoke.

Knowles, M, and Malmkjaer, K. 1996. *Language and Control in Children’s Literature*. Routledge.

Kortenhaus, C. and Demarest, J. 1993. ‘Gender role stereotyping in children’s literature: an update’. *Sex Roles* 28(3/4), 219 – 232.

Mautner, Gerlinde. 2009. ‘Checks and Balances: How Corpus Linguistics can Contribute to CDA’ in Wodak, Ruth and Meyer, Michael (Eds) *Methods of Critical Discourse Analysis* (2nd edition) London: Sage pp. 122-143.

Mautner, Gerlinde. 2009a. ‘Corpora and critical discourse analysis’. In *Contemporary corpus linguistics*, ed. P. Baker. London: Continuum, pp. 32-46.

Motschenbacher, Heiko. 2009. ‘Speaking the gendered body: The performative construction of

commercial femininities and masculinities via body-part vocabulary'. *Language in Society* Vol. 38 pp. 1-22.

Thompson, Paul and Sealey, Alison (2007) 'Through children's eyes? Corpus evidence of the features of children's literature'. In *International Journal of Corpus Linguistics* 12:1, 1-23.

Wharton, Sue (2005) 'Invisible Females, Incapable Males: Gender Construction in a Children's Reading Scheme'. In *Language and Education* 19(3) pp. 238-251

**Ersilia Incelli (Sapienza, University of Rome)**

Uncovering metaphorical patterns in legislative texts on immigration: a corpus-assisted approach to a systematic analysis

Metaphor, like the law, is pervasive in almost all branches of discourse, reflecting the attitudes and values of the society that generates them. Following this dictum, this paper examines metaphor within its legal context, and the pragmatic and communicative role it plays within the highly specialized area of immigration law. I pursue the task by exploring a small corpus of legislative texts spanning the period 1999-2009, (400,000 words), subsequently divided into two sub-corpora of EU regulations and UK statutory laws. The work focuses on how metaphorical language is realized at a phraseological level, producing lexico-semantic and grammatical patterns which arguably reveal the underlying discourse of a government's ideology, policy and principles towards immigration issues. In addition, the study is to some extent experimental, in that, firstly metaphor analysis is not a standard methodology of legal discourse; and secondly it integrates specific applied techniques of cognitive linguistic (metaphor) theory with a corpus linguistic approach to legal discourse. The corpus linguistic techniques exploit the software, Wmatrix, (Rayson, 2003), used to retrieve lexical items from semantic (source) domains and the metaphorical expressions associated with those domains, (Hardie et al., 2005). This sort of procedure can reduce the risk of under-representing metaphors in a corpus. The investigation also integrates a metaphor identification procedure (MIP), (Pragglejazz, 2007), which helps decision-making on 'indefinite' metaphors, i.e. those which are not so easily distinguishable.

So far the findings are contrary to expectations, in that this particular form of legal writing, unlike the descriptive evaluative language of political or media discourse, appears neutral on the surface, and metaphorically constructed language seems to be incongruent in such fait accompli documents which do not aim to persuade, or incite emotion or judgement. Instead, this study has revealed that beneath the surface level of legal reasoning lie conceptual metaphorical schemas, the prominent ones being the NATION/EU AS A CONTAINER and the EU AS EDIFICE, which in turn produce smaller sub-level concepts, e.g. CATEGORIES ARE CONTAINERS (expressed through lexical units such as, *first class/second class/prescribed class of person*, etc.), or the EU AS PATH/ DIRECTION, (*with a view to, carry forward*, etc) and the MIGRANT AS OBJECT, (e.g. *automatic deportation*), (Charteris-Black, 2006). Furthermore the legal terminology familiar to the discourse community consists to a large extent of conventional metaphors of etymological (experiential) origin (Lakeoff and Johnson, 1980), which often go unnoticed, e.g. *miscarriage of justice, reflection period, burden of proof*.

Charteris-Black, J., 2006, 'Britain as a container: immigration metaphors in the 2005 election campaign'. *Discourse & Society*, Vol. 17, No. 5, 563-581 Basingstoke: Palgrave Macmillan.

Hardie, A., Koller, V., Rayson, P., Semino, E., 2007, 'Exploiting a semantic annotation tool for metaphor analysis'. In: M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds) *Proceedings of the Corpus Linguistics 2007 Conference*.

Lakoff, G. & Johnson, M (1980), *Metaphors we live by*, Chicago, IL: University of Chicago Press.

Pragglejazz Group, 2007, 'MIP: a method for identifying metaphorically used words in discourse', *Metaphor & Symbol* 22 pp. 1–39.

Rayson P., 2003, *WMatrix/USAS – Semantic Annotation System*. University of Lancaster.

**Eric Hajime Jego (Nihon University School of Medicine – Tokyo, Japan)**

Doctor talk: simple corpus methodologies for EMP teachers

Corpus methodologies have been used to create word lists and identify multi-word units to inform pedagogy for general English, academic English and medical English as well. The value of word lists in ESP education is well documented (Nation, 2001; Chujo and Nishigaki, 2006). From West's General Service List (1953) to Xue and Nation's University Word List (1984) to Coxhead's Academic Word List (2000) and Wang et al's Medical Academic Word List (2008), the road toward more specific word lists for use in EMP education has been established. The lists of the past have had favourable impact on EFL education. Despite this positive momentum, research on corpora of oral discourse is sparse and the need to develop a spoken academic word list for pedagogic purposes is clear (Chujo and Nishigaki, 2006). As Pastizzo et al. assert "the use of spoken word frequency counts is conspicuously absent in the literature" (2007: 1025). This area of EMP is particularly understudied in general and no major studies could be found which have attempted to create a spoken word list from a corpus of doctor discourse used in medical interviews for use in EMP education. Furthermore, the lists of past studies go to great levels of sophistication, for example, lemmatisation, tagging text for various items such as parts of speech, statistical analysis, and ensuring balanced representation in corpus construction all of which require corpus expertise. Although the benefits of identifying useful terms and expressions is established, little research exists which examines spoken corpora for medical English teaching purposes. This study seeks to propose simple procedures for identifying terms and expressions that doctors use with their patients which have potential pedagogical value for educators teaching medical interview English in EFL environments.

The methods described in this study use freely available materials including the following: AntConc corpus software toolkit, authentic video content from the emp-tmu.net website which includes real doctors interviewing real patients, and a stop list of the top spoken words from the British National Corpus. A corpus was compiled of all the doctor discourse from the website. These materials were used to create a word list and extract multi-word units of potential interest to English for medical purposes educators. While similar studies are very sophisticated and comprehensive, this study is distinct in that it describes simple and practical methods that even educators without extensive corpus expertise can apply.

Chujo, K. and Nishigaki, C. (2006) 'Creating Spoken Academic Vocabulary Lists from the British National Corpus', *Practical English Studies*, vol. 12, pp. 19-34.

Coxhead, A. (2000) 'A new academic word list', *TESOL Quarterly*, vol. 34, no. 2, pp. 213-238.

Nation, P. (2001) *Learning vocabulary in another language*, Cambridge: Cambridge University Press.

Pastizzo, M. and Carbone, R. (2007) 'Spoken word frequency counts based on 1.6 million words in American English', *Behavior Research Methods*, vol. 39, no. 4, pp. 1025-1028.

Wang, J., Liang, S. and Ge, G. (2008) 'Establishment of a Medical Academic Word List', *English for Specific Purposes*, vol. 27, pp. 442-458.

West, M. (1953) *A general service list of English words*, London: Longman, Green & Co.

Xue, G. and Nation, P. (1984) 'A university word list', *Language Learning and Communication*, vol. 3, no. 2, pp. 215-219.

**Régis Kawecky (University of Bretagne-Sud, HCTI-LiCoRN Laboratory, Lorient, France)**

Teaching/Learning French with the Help of a Caribbean French Learner Corpus

Thanks to the seminal work of John Sinclair, Corpus Linguistics has now earned its rightful place in the field of language acquisition. Research started using native speaker corpora whether written or oral. It is only quite recently that academics – like Sylviane Granger – began collecting data derived from learners of second or foreign languages in order to build Learner Corpora. This was done primarily for English. Learner Corpora about other languages are now being built. Most of these Learner Corpora target learners with a good proficiency in the language being learned.

Learner Corpora can be of great help in the teaching and learning of foreign languages. They help researchers identify the most recurrent linguistic problems faced by specific populations of students in their endeavour to learn a particular second/foreign language, a set of difficulties generally called their interlanguage or provisional grammar. Identifying such problems is important in adapting the teachers' pedagogy to the characteristic difficulties their students encounter. It makes possible the creation of supplementary material to support textbooks which are often far too generic in their approach. Such corpora with their associated analysis tools can also be made available to learners for independent learning.

This paper will describe the French Learner Corpus which was built at the Centre for Language Learning (CLL) at the University of the West Indies, Trinidad and Tobago. It is original in the sense that it is made only of texts written by students having just started learning French or, at most, by learners with an intermediate proficiency in that language. It is part of a doctoral research currently being completed at the University of Bretagne-Sud in Lorient, France. This corpus was collected during the period Sept. 2007-April 2009 and is being analyzed using a variety of text searching tools: SCP, Xaira, Sketch Engine and Jaguar.

Initial results show that Trinidadian learners of French differ significantly from other English-speaking students of the language due to their location and specific history.

**Dorothy Kenny (Dublin City University)**

A corpus-based study of the discourse of contemporary machine translation

Human translators have co-existed peacefully with automatic translation for decades, but recent advances in statistical machine translation (SMT, as used in systems like Google Translate) are forcing a reappraisal of this relationship, for a number of reasons. One is that SMT is based on the recycling of human translations, so never before has machine translation been so reliant on human translation. Another is that free online machine translation and collaborative translation platforms allow non-professionals to participate in translation projects in ways that were not possible previously.

In this paper I address a number of issues that arise in this context, focusing on the discourse of statistical machine translation. Drawing on ideas from Conceptual Metaphor Theory, lexical semantics, critical discourse analysis and corpus-based studies of metaphor (especially Deignan 2005, Goatly 2007), amongst others, I examine how key participants in SMT conceptualize contemporary machine translation, and attempt to explain some areas of divergence between the view of translation adopted in machine translation and more traditional translation studies. I argue that the "code-breaking metaphor", heavily referenced in machine translation circles, is based on metonymy rather than being a "pure" metaphor, and that it exhibits a number of the features that

Deignan (ibid.) associates with metonymy-based metaphor (eg lack of systematic mapping between source and target domain; limited overlap between collocates in the source and target domain). I ultimately argue that "code-breaking" is not a convincing metaphor with which to explain translation, although it has significant ideological import. I also discuss how other metaphors used by computer scientists to describe translation (eg when translation is understood as "food") entail ideological positions that obscure the input of human beings to the translation process, a fact that raises a number of ethical issues for researchers and teachers of translation technology alike. Finally, I consider how certain sources address non-expert addressees in particular in such a way as to position translation outside the realm of paid work (referring to the addressee's 'friends' rather than his/her colleagues, for example).

The empirical work presented is based on two resources: firstly a multi-modal corpus of high-profile explanations of statistical machine translation (expert-to-lay communication), taken, for example, from Microsoft and Google blogs and video presentations, press reports, and press releases from major industrial sources; and secondly a corpus of journal articles and refereed conference papers (expert-to-expert communication), from the journal *Machine Translation* and the archive of the European Association for Machine Translation.

Deignan, Alice (2005) *Metaphor and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Goatly, Andrew (2007) *Washing the Brain - Metaphor and Hidden Ideology*. Amsterdam/Philadelphia: John Benjamins.

#### **Hannah Kermes (Universität des Saarlandes)**

##### Usage and function of formulaic expressions in scientific texts

Formulaic expressions are commonplace in scientific writing. Scientists use patterns such as 'based on the', 'the/a number of', 'in other words' consciously or unconsciously to convey research interests, the theoretical / data bases of studies, results of experiments, scientific findings, conclusions and as discourse organizers.

Research studies such as Biber et al. (2004); Biber (2006) and Simpson (2004) use frequencies to identify lexical bundles typical for academic language. Biber et al. (2004) differentiate structural types and distinguish among three primary functions (stance expressions, discourse organizers and referential expressions) each with several subcategories. They show the distribution of the different structural types and categories across different types of academic language (conversation, classroom teaching, textbooks and academic prose). Simpson-Vlach and Ellis (2010) in addition argue for the use of statistical measures such as mutual information for the identification of academic formula. Mutual information allows to find typical and salient but low-frequent patterns, which cannot be identified on a simple frequency basis. They classify their list of formula according to the same primary categories used by Biber et al. (2004), adjusting the scheme of subcategories where their data makes it necessary.

In our approach we want to go a step further, investigating the linguistic usage of formulaic expressions in a mostly automatic fashion. The questions we are interested in are: How are formulaic expressions used in scientific text, when are they used and what function do they have? We extract frequency distributions not only with respect to text type but also with regard to the occurrence of formula within the texts. Do they occur in the abstract, the introduction, the main part or to the end of a text. Are they distributed evenly throughout the text? Are there formula dense text areas, with a relative high frequency of formulaic expressions, and areas with a low density of formulas? We are also interested in differences and commonalities

with respect to the frequency distribution of these features across different scientific disciplines. Taking the extracted frequency distribution of the features and statistical measures as basis, we want to group the formulas and find out, whether it is possible to draw conclusions about their function. The results are compared to the lists of Biber et al. (2004) and Simpson-Vlach and Ellis (2010).

As a data basis for our study we use the DaSciTex (Darmstadt Scientific Text) corpus. A corpus with approximately 17 million words (around 2000 texts) from nine scientific disciplines: four interdisciplinary domains (computational linguistics, bioinformatics, computer-aided design, micro-electronics) and the corresponding “pure” disciplines (computer science, linguistics, biology, mechanical engineering, electrical engineering) (Teich and Holtz, 2009; Teich and Fankhauser, 2010).

Biber, Douglas (2006). *University language*. John Benjamins, Amsterdam.

Biber, Douglas, Conrad, Susan, and Cortes, Viviana (2004). “If you look at ...” Lexical Bundles in University Teaching and Textbooks’. *Applied Linguistics*, 25(3), 371–405.

Simpson, Rita (2004). ‘Stylistic features of academic speech: The role of formulaic expressions’. In T. Upton and U. Connor, editors, *Discourse in the professions: Perspectives from corpus linguistics*. John Benjamins, Amsterdam.

Simpson-Vlach, Rita and Ellis, Nick C. (2010). ‘An Academic Formulas List (AFL)’. *Applied Linguistics*, 31(4), 487–512.

Teich, Elke and Fankhauser, Peter (2010). ‘Exploring a corpus of scientific text using data mining’. In S. T. Grief, S. Wulf, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*. Rodopi.

Teich, Elke and Holtz, Mônica (2009). ‘Scientific registers in contact. An exploration of the lexicogrammatical properties of interdisciplinary discourses’. *International Journal of Corpus Linguistics*, 14(4), 524–548.

#### **Mohammad Abid Khan and Fatima Tuz Zuhra (University of Peshawar, Pakistan)**

##### **Role of Corpus in Anaphora Resolution**

Human beings use pronouns for avoiding repetitions but computers need to process text in simplified form for a number of potential applications such as machine translation. A key component of text simplification is anaphora resolution e.g. replacing pronouns by their antecedents. In this research paper, a corpus-based approach for resolving anaphora is worked out. Results show that this approach is much more economical compared to traditional strategies. Pashto corpus is divided into two parts: training part and testing part. Anaphora resolution rules are learnt automatically from the training part and subsequently tested on the testing part of the corpus. Both parts are carefully selected as representative samples from the Pashto corpus of 1.225 million words (Khan and Zuhra, 2009). Discourse boundaries are identified in the training part of the corpus having 14000 words. There are 68 discourse units in the training sample. Anaphora is resolved manually in the training part and the anaphoric and non-anaphoric versions of each discourse unit are stored side by side with each other. This text is then part-of-speech (POS) tagged using the POS tagger for Pashto (Khan, 2010a) including numbering of nouns and noun phrases. Tagged sequences are extracted from the text in such a way that the anaphoric and the non-anaphoric sequences of each discourse unit are saved parallel to each other in a Microsoft Access database table. Later, tagged sequences of sentences are extracted from both versions of the tagged sequences of discourse units using the methodology of Khan (2010b). The tagged sequence of each anaphoric sentence is

compared with the tagged sequence of its non-anaphoric counterpart. The differences are recorded. These differences form the basis for extracting the anaphora resolution rules. A total 492 sentence pairs of the training part are compared and rules are formulated automatically. These rules are then manually normalized using observation of the tagged sequences of discourse units and the corresponding text. A total 142 anaphora resolution rules are bagged in this process. These rules are tested in a semi-automatic way using a testing sample of 50 discourse units containing 371 pronouns, out of which 59 pronouns occurred in a non-anaphoric way i.e. used for pointing purposes. Of the remaining 312 pronouns, 230 pronouns are correctly resolved using the rules learnt by the system. This is the first step in this direction. Now the size of the training part of the corpus can be increased by adding to it the automatically resolved text and new text is added to the testing corpus. Repetition of this iterative process will very quickly lead to the perfection of the technique.

Baker, P. et al. 2008. "Discourse and Society". [Online]. Available from the URL: <http://das.sagepub.com/cgi/content/abstract/19/3/273>

Ali, R., Khan, M.A. and Rabbi, I., "Strong Personal Anaphora Resolution in Pashto Discourse", In proc. IEEE ICET 3rd International Conference on Emerging Technologies, Islamabad, Pakistan. 2007, pp 148-154.

Ali, R., Khan, M.A. and Rabbi, I., "Reflexive Anaphora Resolution in Pashto Discourse", In proc. Conference on Language and Technology, 2008.

Gasperin, C. et al. 2009. 'Learning When to Simplify Sentences for Natural Text Simplification'. In the proc. ENIA 2009 (VII Encontro Nacional de Inteligencia Artificial), Bento Goncalvez, RS, Brazil.

Jurafsky, D. and Martin, J. H. 2002. *Speech and Language Processing*. Pearson Education Series in Artificial Intelligence, Colorado.

Khan, M. A. 2010 a. "A Part of Speech Tagger for Pashto", Working paper, Department of Computer Science, University of Peshawar, 2010.

Khan, M. A. 2010 b. "Extraction of Grammar Rules from the Pashto Corpus", Working paper, Department of Computer Science, University of Peshawar, 2010.

Khan, M. A. and Zuhra. 2007. "A General-Purpose Monitor Corpus of Written Pashto". In proc. Corpus Linguistics 2007. Birmingham, UK.

Khan, M. A. and Zuhra. 2009. "A Corpus-Based Study of Pashto". In proc. Corpus Linguistics 2009. Liverpool, UK.

McEnery, T. et al. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge Applied Linguistics.

Ooi, V. 2001. 'Investigating and Teaching Genres Using the World Wide Web'. In M. Ghadessy, A. Henry and R.L. Roseberry (eds) *Small Corpus Studies and ELT*. Amsterdam: Benjamins.

**Iztok Kosem and Tadeja Rozman (both Trojina, Institute for Applied Slovene Studies), and Mojca Stritar (Faculty of Arts, University of Ljubljana)**

How do Slovenian primary and secondary school students write and what their teachers correct – a corpus of student writing

In the last decade, corpus linguistics has witnessed a large increase in the number of learner

corpora. Most widely known examples are the International Corpus of Learner English (ICLE), and the Cambridge Learner Corpus. Following the steps of these corpora, a corpus of student writing (Šolar) has been compiled within the 'Communication in Slovene' ([www.slovenscina.eu](http://www.slovenscina.eu)) project. The corpus is the first freely available learner corpus of Slovene. Important features of the corpus are that the corpus texts were produced by students as part of the curriculum (rather than solely for project purposes), and were annotated for corrections made by teachers (rather than by researchers). The paper will discuss different stages of corpus compilation, and present the tags used for annotating teacher corrections, and their contribution to the usefulness of the corpus.

The corpus includes texts from different subjects produced by students in primary and secondary schools. Many different types of texts can be found in the corpus, for example essays, tests, invitations, thank you notes, summaries, people profiles, and personal diaries. 39 schools from different Slovenian regions took part in the project, and over 8500 texts were collected. The collection of texts demanded a great amount of effort from teachers-volunteers, as it involved not only copying texts and sending them to us, but also obtaining written permission from students and their parents. Only 2700 texts are included in the first version of the corpus because the processes of manual digitalisation (transcription) and anonymization proved very time-consuming. In addition, the focus was more on achieving a good regional balance of the corpus, and less on its size.

An important added value of the corpus is annotation tags that mark linguistic corrections made by teachers. The tags were divided into four categories (spelling, wording, form and syntax), and were also used as subtypes in cases where two different types of error were identified for the same item in the text. Also annotated were written comments and various graphic symbols, such as underlined text. The annotations will offer an insight into the teachers' interventions, indirectly showing teachers' understanding of the linguistic norm, and the way they teach linguistic norm and conventions of Slovene. Further (corpus) analysis is planned to annotate errors that had not been marked by teachers.

The corpus is also lemmatized, part-of-speech tagged and parsed. Also planned is the design of a user-friendly interface that will, among other search options, offer search by type of error. The corpus will be used within the 'Communication in Slovene' project as one of the resources in designing Pedagogical corpus-based grammar and Manual of Style. It is also expected that the corpus will be used for researching students' writing skills, and in teaching.

**Natalia Judith Laso, Isabel Verdaguer, and Elisabet Comelles (University of Barcelona)**

There is no clear evidence but... what is known about negative polarity in the organisation of scientific research articles?

Much recent research into discourse analysis has focused on the phraseology characteristic of a given genre as well as its textual distribution (Hyland 1997, Tognini-Bonelli 2001, Wray 2002, Cortes 2004, Schmitt 2004, Biber et al. 2007, Meunier & Granger 2008, Granger & Paquot 2008, Hyland 2008). An underlying assumption has been the important role of prefabricated expressions in the textual development of meaning (Gledhill 2000, Kaszubski 2000, Verdaguer *et al.* 2010, Römer and Schulze 2009, 2010). Phraseological empirical studies have also highlighted the need for further research on the phraseological conventions distinctive of specialist genres.

Scientific discourse relies heavily on formulaic constructions which need to be mastered by the members of the scientific community, so as to produce phraseologically competent research articles. A corpus-based approach to the study of phraseology brings to the forward the close relationship between discourse and lexical grammar (Tognini-Bonelli 2008). Thus, a taxonomy of recurrent multiword expressions in scientific language according to their function in the discourse will contribute to a better understanding of the organization of research articles.

This paper explores how discourse meaning is highly dependent on its lexico-grammar, analyzing a corpus of 290 research articles in biology and biochemistry. By linking discourse analysis and corpus linguistics (Charles, Hunston and Pecorari 2009), we will study how negative polarity and the scope of negation in combination with discourse connectors can bring about systematic patterns in scientific discourse. The analysis of the context, position in the clause and text distribution of some adjectives which appear in the negative in the data under analysis will reveal how different rhetorical devices, such as hedging and polarity, contribute to the rhetorical weight of the multiword expressions selected in this study.

Biber, D. Connor, U., Upton, T. A. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam/Philadelphia: John Benjamins.

Cortes, V. 2004. 'Lexical bundles in published and student disciplinary writing: Examples from history and biology'. *English for Specific Purposes*, 23:397-423.

Charles, M., Hunston S. & Pecorari, D. (eds.) 2009. *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum.

Gledhill, C. J. 2000. *Collocations in science writing*. Gunter Narr: Tübingen.

Granger, S. & Paquot, M. 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins.

Hyland, K. 1997. 'Scientific claims and community values: articulating an academic culture'. *Language and Communication*, 17(1), 19-31.

Hyland, K. 2008. "As can be seen: Bundles and disciplinary variation". *English for Specific Purposes*, 27:4-21.

Kaszubski, P. 2000. 'Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-Based Approach'.  
<http://www.staff.amu.edu.pl/~przemka/rsearch.html#PhD>

Meunier, F. & Granger, S. 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/Philadelphia: John Benjamins.

Römer, U. & Schulze, R. (ed.) 2009. *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: John Benjamins.

Römer, U. & Schulze, R. (ed.) 2010. *Patterns, Meaningful Units and Specialized Discourses*. Amsterdam/Philadelphia: John Benjamins.

Schmitt, N. (ed.) 2004. *Formulaic sequences*. Amsterdam: John Benjamins.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Verdaguer, I., Comelles, E., Laso, N. J., Giménez, E. & Salazar, D. 2010. "SciE-Lex: an electronic lexical database for the Spanish medical community" In S- Granger, S & M. Paquot. 2010. *E-Lexicography in the 21st century: New Challenges, New Applications*. Proceedings of E-Lex 2009, Louvain-la-Neuve,

22-24 October 2009, pp. 325-334.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

**Anu Lehto (University of Helsinki)**

Development of subordination in Early Modern English legal discourse

Legal discourse is generally perceived as a highly specialised language variety, especially due to its complex linguistic features including long sentences and generous use of embedded clauses (Danet 1980, Bhatia 1993). In a diachronic perspective, it has been estimated that the language of historical legislation gradually became more complex and that legal documents written in public offices grew in verbosity during the Early Modern period (Hiltunen 1990: 58, Mellinkof 1963: 188). The paper charts this development of historical legal discourse and complexity in more detail by examining subordination by corpus linguistic methods in Early Modern English national statutes (cf. Biber 1992).

The paper assesses the patterns and specific uses of subordination and the possible growth of complexity in a period of over two hundred years. The study is carried out in the Corpus of Early Modern English Statutes (1491–c. 1700) that is compiled specifically for the purposes of diachronically exploring various complexity features in national legislation. Beginning from the first printed statutes in English, the corpus contains printed orders in different subgenres including parliamentary acts, royal orders and proclamations.

Subordination in the statutes can be estimated to partly reflect the general patterns found in other genres of the era: embedded that-clauses, for instance, are most frequently used, and relative that-clauses decrease during the sixteenth century (see Blake 1996: 226). In contrast to other genres, embedded clauses giving explanations, such as those beginning with "because", are rare. The paper shows that subordination is linked to conventional and repeated language use, e.g. that-clauses tend to follow the enacting clause ("Be it enacted that...") and they are typically repeated before each section or subsection. Different types of repeated embedding aid in establishing a regular textual structure. Additionally, an analysis of the if–then structure considers the practice used in law that first describes a possible situation to which particular regulations are then given. Overall, the requirements of exactness and all-inclusiveness that bind contemporary legislation are encountered also in the studied era. Subordination is hence a noticeable feature of sentence complexity and is common in the initial, central and final positions of sentences.

In addition to corpus linguistic methods, the study relies on historical pragmatics by placing the findings in their sociohistorical context. It also analyses legal discourse as a professional genre within institutional setting (see Bhatia 2004: 29). During the period, a set of extra-linguistic changes, such as the introduction of printing, affected the construction of laws. Further, a reshaping of legal thought initiated by Humanism in the sixteenth century placed more emphasis on the role of law in society and made legal professionals more aware of legal drafting (Baker 2003: 17).

Baker, John. 2003. *The Oxford History of the Laws of England, vol. VI 1483–1558*. Oxford: Oxford University Press.

Bhatia, Vijay K. 2004. *Words of Written Discourse. A Genre-based View*. Advances in Applied Linguistics. London: Continuum.

Bhatia, Vijay K. 1993. *Analysing Genre. Language Use in Professional Settings*. London: Longman.

Biber, Douglas. 1992. 'On the complexity of discourse complexity: A multidimensional analysis'.

*Discourse Processes* 15, 133–163.

Blake, N.F. 1996. *A History of the English Language*. New York: New York University Press.

Danet, Brenda. 1980. 'Language in the legal process'. *Law and Society Review* 14.3, 445–564.

Hiltunen, Risto. 1990. 'Chapters on Legal English. Aspects Past and Present of the Language of the Law'. *Academia Scientiarum Fennica*. Jyväskylä: Gummerus.

Mellinkoff, David. 1963. *The Language of the Law*. Boston: Little, Brown and Company.

**Phoebe M.S. Lin (City University of Hong Kong)**

A multimodal analysis of multiword units in university lectures

In the past decades, Sinclair's (1991) idiom principle has inspired many corpus-based investigations into the patterns of use of multiword units. While most studies approach multiword units from the lexical perspective, very few look systematically into the prosody of multiword units.

On the prosody of multiword units, phonologists (e.g. Ashby, 2006; Wells, 2006) suggest that idioms have relatively fixed prosodic patterns. However, these suggestions are based on anecdotal, introspective data and often concern semantically opaque idioms (e.g. to rain cats and dogs).

In corpus linguistics, Aijmer (1996), Altenberg and Olofsson (1990) and Moon (1997) have proposed that multiword units align with tone unit boundaries. However, Aijmer (1996) is the first researcher to use the term 'prosodic fixedness' to describe conversational routines, which are considered a type of multiword units.

This paper reports an empirical study conducted with an aim to establish whether semantically transparent multiword units demonstrate prosodic fixedness. Spontaneous spoken data from the first five minutes of an academic lecture collected in the Nottingham Multi-Modal Corpus (NMMC) were analysed in terms of the division into intonation units, stress placement, the distribution of pauses, articulation rate and the patterns of use of multiword units.

Unlike a previous study by Lin and Adolphs (2009) which used automatic extraction, this study invited 30 non-linguist native speakers to identify multiword units. Although subjectivity of idiomaticity judgement may be an issue, the benefit of using human judges is that the identification of multiword units is based on a range of criteria (e.g. contextual meaning, semantics) rather than form alone. This multiple-criteria approach is particularly important when investigating the use of multiword units in discourse. Besides, supported by the empirical evidence that native speakers can readily judge the relative idiomaticity of multiword units (e.g. Wulff, 2008), this study introduced an idiomaticity scoring system so that each multiword unit can be classified based on how confident the judges are about their idiomaticity judgement of that unit. Only units that were assigned the highest score were used in the next stage of the analysis to map onto the original lecture text which was prosodically transcribed by an independent, professional transcriber.

The results of this study show that multiword units have a tendency to align with intonation units and are markedly less unlikely to receive stress. However, the articulation rate of multiword units is more difficult to model because it is subject to discourse factors. Put simply, some of the multiword units were uttered in a distinctively slow rhythm because they fulfilled specific discourse purposes in the lecture (e.g. signposting). In addition, slow rhythm is found to be a strategy with which the speaker draws listeners' attention to the literal meaning of these multiword units.

Aijmer, K. (1996). *Conversational routines in English*. London: Longman.

Altenberg, B., & Eeg-Olofsson, M. (1990). 'Phraseology in spoken English: Presentation of a project'. In J. Aarts & W. Meijs (Eds.), *Theory and practice in corpus linguistics* (pp. 1-26). Amsterdam: Rodopi.

Ashby, M. (2006). 'Prosody and idioms in English'. *Journal of Pragmatics*, 38(10), 1580-1597.

Lin, P. M. S., & Adolphs, S. (2009). 'Sound evidence: Phraseological units in spoken corpora'. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 34-48). Basingstoke: Palgrave Macmillan.

Moon, R. (1997). 'Vocabulary connections: Multi-word items in English'. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40-63). Cambridge; New York: Cambridge University Press.

Sinclair, J. M. (1991). *Corpus, concordance and collocation*. Oxford: Oxford University Press.

Wells, J. C. (2006). *English intonation: An introduction*. Cambridge: Cambridge University Press.

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach*. London: Continuum.

**Yen Yu Lin and Siaw-Fong Chung (National Chengchi University, Taiwan)**

A Study on the Semantic Preference and Semantic Prosody of "CHALLENGE"

Partington (1998, p.8) once said that it is crucial for ESL/EFL learners to understand not only what is grammatically possible but also what is appropriate. Here, 'appropriate,' refers to the correctness of connotations with the right speaker's attitude. These constitute the definition of 'semantic prosody,' referring to a node word that is typically co-occurring with particular sets of lexical items of the same semantic field, such that this word takes on connotations from that environment (Stubbs, 2001). This paper aims to investigate the word, CHALLENGE, which, according to SYNONYM.COM (<http://www.synonym.com/>), is synonymous to the words like *difficult*, *dispute*, *quest*, and *object*, all of which possess a negative connotation. However, when resorting to dictionary definition, CHALLENGE was found to also possess positive senses such as "stimulating, arousing competitive interest, thought, or action."

Using data from ukWaC, a web-based corpus containing 1,565 million words, the semantic profile of CHALLENGE as verb, noun and adjective was extracted. As a preliminary study, the top forty collocates of CHALLENGE in each grammatical relation (e.g. objects of CHALLENGE, subjects of CHALLENGE, etc.) were coded and categorized into various sense-groups based on the shared semantic features. The percentage calculation of each group upon the accumulated frequency of the top forty collocates clearly informs us of the sense-group ranking. Through interpreting the statistics and the nature of each group, the semantic prosody of CHALLENGE becomes apparent.

The results showed that CHALLENGE bears a mixed prosody. Neutral and positive prosody occur much more frequently than negative prosody. When CHALLENGE serves as a verb, the most frequent sense group of its objects was found to possess a neutral sense (CHALLENGE+notion/perception/thinking) with the percentage frequency of this top-ranking sense group reaching 48% upon all collocates. When serving as a noun, CHALLENGE is more likely to collocate with adjectives with positive prosody (e.g. big/major/key+CHALLENGE, CHALLENGE+is/are+immense/enormous/considerable), which carry the sense of "extremely large in size or degree". Moreover, it was found that CHALLENGE also reveals a strong positive prosody when connecting to other words in the same part of speech by

conjunction:

(1) use interactivity and games to stimulate, challenge, and

absorb learners

(2) focused and looking for a career full of challenges,

and rewards

(3) would like career opportunities which challenging,

exciting, and rather, well very cool

In addition, the verbs preceding CHALLENGE often indicate directionality (e.g., face, pose, encounter).

Comparatively, negative prosody occurs less frequently. For example, when CHALLENGE serves as a verb, its objects which portray an unfavorable meaning (CHALLENGE+stereotype/preconception/racism) account for merely 18.14% of the accumulated frequency. Similarly, when CHALLENGE serves as a noun and when it connects with other words with the same part of speech (cf. (1) (3) above), only about one-fourth of the total occurrences of collocates were labeled with negative sense.

In summary, the corpus data show that, although CHALLENGE is categorized as words possessing negative senses, the prosody of it is found to be positive or neutral. Pedagogic-wise, the findings herein can serve as the base for language instructors to design teaching materials and help EFL/ESL learners to avoid making overgeneralization in their use of semantic prosody.

Stubbs, Michael. (2001). *Words and Phrases*. Oxford: Blackwell.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Philadelphia, PA: John Benjamins.

**Anne Li-E Liu (University of Nottingham)**

Concessives in discourse: the case of multi-word discourse markers

This paper describes a corpus-based investigation of how professional writers, native student writers, and L2 learners create a concessive relation via the use of multi-word discourse markers (MDMs) in different discourse contexts. Altenberg (1986) and Liu (2010) investigate patterns of discourse markers (DMs) and propose a formality scale/continuum on the basis of their corpus findings. Both single-word markers and MDMs in the same semantic category can be positioned on the continuum, depending on the corresponding degree of formality. For example, Liu (ibid.) suggests that despite that is the most frequently used concessive MDM in a formal context, whereas but then again is mainly used in a colloquial context. Despite the fact that both Altenberg and Liu include spoken and written corpora in examining the formality of DMs, a perceived flaw is the lack of a direct comparison made, for example, between formal and less formal written data. As concessive relation seems to be the most complex case and it poses difficulties at different points for both NS (Sanders et al., 1992) and NNS (Godó, 2008; Bell 2010), this study aims at offering a fine-grained description of how language users employ concessive MDMs in various registers by including an E-mail corpus, a formal written corpus and a learner corpus.

Four corpora are included, the British National Corpus (BNC), the British Academic Written English corpus (BAWE), the EnroSent Corpus (a collection of 13 million words of E-mail messages), and the International Corpus of Learners English (ICLE). The EnroSent corpus provides the less formal written samples and will be compared with other corpora. Both qualitative (manual examination) and quantitative methods (raw frequency, normalized frequency and the log-likelihood score) are included in data collection and analysis. At the outset, the concessive MDMs in Liu's study (despite this/that, but at the same time, but (then) again/then again, having said that, and but still) are searched in the above-mentioned corpora. A subtle nuance is perceived between the BNC (academic writing) and the EnroSent data. Among the less formal MDMs (Liu, *ibid.*), but then again, but at the same time, and but still, the shortest MDM, but still, is relatively favoured in the E-mail discourse. That is, a refined formality continuum is formed when the language mode is taken into account. In the second stage, I compare the use of concessive MDMs by professional writers (BNC written texts), by native student writers (BAWE), and by L2 writers (ICLE). The results show that native student writers employ more informal MDMs than the professional writers; yet, they use the formal marker despite that 10 times more frequently than the professional writers. L2 writers, on the other hand, display a similar pattern found in the E-mail corpus.

Altenberg, B. (1986). 'Contrastive linking in spoken and written English', in *English in Speech and Writing*, (ed.) by G. Tottie and I. Backlund. Uppsala: Semqvist & Wiksell International, pp. 13-40.

Bell, D. M. (2010). 'Nevertheless, still and yet: Concessive cancellative discourse markers', *Journal of Pragmatics*, Vol. 42, No. 7, pp. 1912-927.

Godó, A. M. (2008). 'Cross-cultural aspects of academic writing: a study of Hungarian and north American college students L1 argumentative essays', *International Journal of English Studies*, Vol 8, 2, pp. 69-111.

Liu, A. L. E. (2010). 'Multi-word discourse markers in use: the patterns in spoken and written corpora', Paper presented at the 31st ICAME conference, May 2010, Giessen University, Germany.

Sanders, T.J.M., Spooren, W.P.M. & Noordman, L.G.M. (1992). 'Toward a taxonomy of coherence relations', *Discourse Processes*, 15, pp.1-35.

#### **W. E. Louw (University of Zimbabwe)**

##### Re-defining Subtext as Logical Form

The corpus-assisted recovery of subtext lies at the centre of the debate on natural language philosophy and logical form (Wittgenstein, 1922: 7; Russell, 1956: 197; Carnap, 1928). Russell notes that it is possible, crudely, for natural language to mimic logic: '...it is a language that has only syntax and no vocabulary whatsoever' (Pears, 1972: 24). But if this were done, the question would remain concerning the extent to which the full words in ordinary text might be replaced by words, on a cline, that are apparently full upon inspection, but still in the nature of sub-technical terms (Sinclair, 2004: 97), and behave like grammar words. Wittgenstein (1929) declares, uncharacteristically, that this task is difficult. However, his language singles him out (with the notable exception of Malinowski) as the first corpus linguist. He even offers a method involving the use of wildcard searches without knowing that the term 'wildcard' would only be invented 30 years later. Hence, the procedure and the terms recovered by it using Russell's technique can today take the form of wildcard searches for full words as quasi-propositional variables. For example, Yeats's line: 'That is no country for old men' gives rise to a search line of the form: 'that+is+no+\*+for+\*'. Our candidates for the variables represented by the first wildcard turn out to be (from a large reference corpus), REASON and EXCUSE. As they are the variables with the highest frequency, it is correctly hypothesized that their function will be strongly akin to that of grammar words. If we apply them to

the poem from which they come, ‘Sailing to Byzantium’, we not only discover that they falsify the persona’s actions in the poem, but act as a contextually prosodic argument: a logical semantic prosody that is co-extensive with the text’s literary world.

This discovery complements and further systematizes semantic prosodies recovered earlier, confirming that they were never simply part of metaphysics. These techniques will begin to create pressure for the use of collocation to re-organise the dogmas of empiricism and in so doing, improve the quality and accessibility of corpus empiricism (McGinn, 1981: 89). The removal from the Third Dogma of the conceptual scheme by which the given is interpreted will have profound implications for the science of and hence the future of cognitive approaches. This would favour the analysis by discourse communities of the textual duress from which Teubert (2010: 259) foresees no escape. Pessimism must give way to those forms of emancipation (Marx, 1992) already secured by instrumentation for language based on collocation which Firth (1957: 196) referred to as ‘...not directly concerned with the conceptual or idea approach to the meaning of words.’

Carnap, R. (1928). *Der Logische Aufbau der Welt*. Berlin-Schlagentensee: Weltkreis-Verlag.

Firth, J.R. (1957) *Papers in Linguistics 1934-1951*. Oxford: OUP.

Marx, K. (1992) *Capital*. Harmondsworth: Penguin.

McGinn, M. 1981. ‘The third dogma of empiricism’. In *Proceedings of the Aristotelian Society*. New Series, Volume LXXXII. Bristol: Photobooks Press

Pears, D.F. (1972) *Bertrand Russell: A Collection of Critical Essays*. New York: Anchor Books.

Russell, B. (1956) ‘The philosophy of logical atomism’ in R.C. Marsh (Ed.) *Essays in Logic and Knowledge*. London: Allen and Unwin.

Sinclair, J.M. (2004) *Trust the Text*. London: Routledge.

Teubert, W. (2010) *Meaning, Discourse and Society*. Cambridge: CUP.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. Trans. D.F. Pears and D.F.

Wittgenstein, L. (1929) ‘Some remarks on logical form.’ In J. Klagge and A. Nordmann, (Eds.) (1993) *Ludwig Wittgenstein: Philosophical Occasions*. Indianapolis: Hackett Publishing Company.

**Andy Lücking (Goethe-Univ. Frankfurt am Main), Olga Abramov (BMBF project Linguistic Networks, Univ. Bielefeld), Alexander Mehler (BMBF project Linguistic Networks, Univ. Bielefeld and Goethe-Univ. Frankfurt am Main), Peter Menke (CRC 673 Alignment in Communication, Univ. Bielefeld)**

The Bielefeld Jigsaw Map Game (JMG) Corpus

Spoken language still poses a challenge to mechanisms developed for information processing and retrieval. Applications in this area often require a large amount of annotated data, which is hardly obtainable for spoken language. We present a corpus of 64 semi-controlled dialogues (length: approx. 18h) completely transcribed and annotated on various linguistic levels. The dialogues stem from a psycholinguistic, task-oriented coordination game experiment, namely the Jigsaw Map Game (JMG) [2].

In a game round of the JMG two participants had to cooperatively locate predefined objects on a

common interaction space according to a set of instruction cards. Cooperatively means that each participant in turn had the role of the instructor, explaining to his partner where to locate an object according to the arrangement shown on the current instruction card. Each game unit consisted of two rounds: in the first round a naïve participant played the game with a confederate who was trained to use only one of two equally possible names for some objects (control condition). In the second round, the participant played this game again with a new partner (experimental condition). The focus of the study was to observe adaptation behaviour on the use of linguistic constructions (e.g. agreement on particular object names). Since participants were sitting in front of each other, they implicitly needed to agree on a strategy to refer to the common interaction space (e.g. partner- or self-perspective). Here, different kinds of referential strategies were observed. A distinguishing feature of the JMG dialogues is the combination of fixed utterance topics (i.e. the set of objects) and apart from that unconstrained language use (dyadic communication). Primarily developed to study referential aspects of alignment in communication [1], the corpus represents a resource for natural language processing and studies on spoken language of a more general kind. The data is not only fully transcribed into time sequences, words and utterances but also augmented with annotations of recurrent dialogical events, repairs, anaphora, and other kinds of linguistic information. In light of this pre-processing, the JMG corpus contributes an empirical basis to study negotiation processes in dialogue, and dialogue theoretical issues in general, which is hardly available so far. Furthermore, speech processing and parsing algorithms on spoken data can benefit from this resource as a test bed for training and evaluation.

The full paper describes 1. the theoretical motivation of the JMG, 2. its experimental setting, 3. all levels of annotation that were included into the JMG corpus together with their reliability evaluation, and 4. case studies of exploring the JMG corpus in the area of research on lexical alignment.

[1] Martin. J. Pickering and Simon Garrod. 'Toward a mechanistic psychology of dialogue'. *Behavioral and Brain Sciences*, 27:169-226, 2004.

[2] Petra Weiß, Thies Pfeiffer, Gesche Schaffranietz, and Gert Rickheit. 'Coordination in dialog: Alignment of object naming in the jigsaw map game'. In *Proceedings of the 8th Annual Meeting of the Cognitive Science Society of Germany*, April 2008.

#### **Ismail El Maarouf (Université Bretagne Sud, UEB)**

##### Verb-noun collocations at the crossroads of Discourse surface patterns

Collocation has been widely studied in Corpus Linguistics (Firth, 1957; Palmer, 1968). It refers to significant lexical patterns throwing light on word use and meaning. Traditional automatic collocation analysis (Sinclair, 1966, 1991; Sinclair et al., 2004) uses word units, word spans ("windows"), and statistic metrics to retrieve frequent or significant collocations for a given word. Such an approach heavily depends on statistic measures (Mutual Information, Z-score; Church & Hanks, 1989; Clear, 1993) to select and order relevant collocates.

Since then, various techniques have been proposed to classify collocations automatically. One contribution is the use of syntactic data to filter collocates (Kilgarriff et al., 2004). Such an approach classifies collocations according to syntactic relations such as Subject or Object between a noun and a verb. Based on a hand-crafted grammar, the system returns the total number of words satisfying a grammar rule in a given corpus. Thus, the issue of window parameters does not hold explicitly in this approach, but the selection of collocates is dependent on the parser's success i.e. grammar coverage.

In both approaches, text distance between nodes and collocates remains a problem. Regarding the

traditional approach, a significant collocate may well lie outside the fixed span while an “irrelevant” collocate might be counted inside the span, though it does not hold any particular relation with the node. The syntactic approach is not foolproof either, one reason being that grammars are not designed to deal with discourse phenomena. Grammars generally make use of finite-state automata to define rules on a limited context (Evert & Kermes, 2003) but there is little work on how those rules interact with discourse structure (see however Say & Akman, 1997, Bayraktar et al., 1998). For instance, phrases separated by commas are not related to each other, with the outcome that a verb separated from its subject by interpolated material is not retrieved. In this perspective, we may expect precision and recall drops for collocation extraction (Kilgarriff et al., 2010).

Our approach (drawing on Jones, 1996) consists in identifying and organizing sentence blocks according to a set of punctuation and discourse markers before syntactic analysis, in order to deal with discourse variation phenomena, such as interpolated clauses and phrases. In a second step, the parser analyses those pseudo-blocks where material irrelevant for the task of relation detection has been discarded. This method has the benefit of filtering beforehand irrelevant collocates, limiting parser's errors and selecting new candidates for collocation.

The paper presents a quantitative analysis of the types of blocks found by our system in a large Press corpus and investigates the benefits of this method with respect to the Subject relation collocates.

Bayraktar M., Say B. & V. Akman, 1998, “An Analysis of English Punctuation: The Special Case of Coma”. In *IJCL*, 3(1): 33-58.

Church K.W. & P. Hanks, 1989, “Word Association Norms, Mutual Information, And Lexicography”. In *Computational Linguistics*, 16(1) : 22-29.

Clear, J., 1993, “From Firth Principles — Computational Tools for the Study of Collocation”. In Baker M., Francis G. & E. Tognini-Bonelli (eds.), 1993, *Text and Technology*.

Evert S. & H. Kermes, 2003, “Annotation, storage, and retrieval of mildly recursive structures”. In *Proceedings of SProLaC 2003*.

Firth J.R., 1957, *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Jones B., 1996, *What's The Point? A (Computational) Theory of Punctuation*. Phd Thesis, University of Edimburgh.

Kilgarriff A., Rychly P., Smrz P. & D. Tugwell, 2004, “The Sketch Engine”. In *Proceedings of Euralex 2004*.

Kilgarriff A., Kováčik V., Krek S., Srdanović I. & C. Tiberius, 2010, “A Quantitative Evaluation of Word Sketches”. In *Proceedings of Euralex 2010*.

Palmer F. R. (eds), 1968, *Selected Papers of J. R. Firth, 1952-1959*. London: Indiana.

Say B. & V. Akman, 1997, “Current Approaches to Punctuation in Computational Linguistics”. In *Computers and the Humanities*, 30(6). pp 457-469.

Sinclair J McH, 1966, “Beginning the study of Lexis”. In Bazell C. E. , Catford J. C. , Halliday M. A. K., R. H. Robins (eds), 1966, *In Memory of J. R. Firth*. pp. 410-430.

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

Sinclair J. McH, Jones S., R. Daley, 2004 (1970), *English Collocation Studies – The OSTI Report*. R. Krishnamurthy (eds). London: Continuum.

**Markéta Malá (Charles University in Prague, Czech Republic)**

Translation counterparts as markers of function: the case of copular clauses in a parallel corpus

While traditional contrastive approaches to language study focussed on a comparison of language systems, the use of parallel corpora now makes it possible to turn the attention to the study of corresponding ‘units of meaning’ in the languages compared. “If we assume that we may find the meaning of a textual element through its paraphrase, which is also a text, then we may describe parallel corpora as repositories for such paraphrases.” (Teubert, 2001: 151) In other words, multi-lingual corpora “can make meanings visible through translation patterns” (Johansson, 2007: 28).

Our approach relies on the patterns of translation correspondence as devices which enable us to proceed from a particular meaning or function to its realization forms. Assuming that the various English constructions which share the same Czech translation counterpart are functionally equivalent, the counterpart can be used to identify the members of a group of English constructions which share the same function. For instance, copular verbs *seem*, *appear*, *look* and *sound* were found to be frequently translated by the same Czech epistemic adverbials. Searching for the equivalents of these adverbials in English texts, other constructions conveying epistemic modification - such as modal verbs, adverbs and adjectives, or comment clauses – can be grouped together. While there may be parallel paradigms of means of expressing the particular function in the two languages, the actual patterns of choice can be language-specific. Moreover, the pattern of preferences in the source language can leave its mark on the translation through overuse or underuse of particular constructions. This may be illustrated by the higher proportion of copular verbs as means of epistemic modification in English source texts as compared with English translations from Czech.

Since the repertory of copular verbs in English is much broader than that in Czech (with *být* and *stát se*, corresponding to *be* and *become*, only), Czech translations display a variety of equivalents overtly rendering into Czech the various types of ‘modified’ attribution of a quality or value to the subject conveyed by the English copular verbs. Apart from epistemic modification, these equivalents make it possible to search, for example, for the expression of aspectual meanings of ‘remaining’ and ‘becoming’ in English.

The corpus used is InterCorp, a multilingual corpus being put together within the project Czech National Corpus and corpora of other languages. The project aims at compiling a large multilingual corpus with Czech as its pivot language, comprising, at the moment, 21 European languages. Our analysis of translation correspondences of copular clauses is based on the English-Czech bidirectional section of the corpus (about 8 million tokens).

InterCorp: Czech National Corpus and Corpora of Other Languages, <http://www.korpus.cz/intercorp>

Johansson, S. (2007): *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Teubert, W. (2001) ‘Corpus Linguistics and Lexicography’, *International Journal of Corpus Linguistics*, Vol. 6 (Special Issue), 125-153.

**Anna Marchi (Lancaster University)**

'Between journalists and reporters'. A corpus-assisted analysis of occupational representations within journalism

Journalism must be examined as a corpus... (Carey 1997: 148)

This paper outlines findings from a larger project which analyses the ways journalists represent themselves and their trade in their own news-work. The paper focuses on the conceptualizations of journalism emerging from the discourses surrounding different kinds of newswriters, in particular analysing the differences between 'journalists' and 'reporters'. Dictionaries present the two words as near synonyms, but corpus evidence shows that the terms are not exactly interchangeable, and are loaded with rather distinctive meanings. Additionally, a methodological question concerning the bottom-up generation of research questions in corpus-assisted (or based) discourse studies (Baker 2006, McEnery and Gabrielatos 2006) and the serendipitous opportunities offered by CADS (Partington 2009) is addressed.

The main assumption underpinning the research is that journalists create interpretative communities (Zelizer 2004) through the discourses they circulate about their profession, and that the meaning and role of journalism are constituted through daily performance (Matheson 2005) and can be studied by means of the self-reflexive traces newswriters leave in their texts. That is, they can be detected and studied in a newspaper corpus.

The corpus under investigation contains the complete output of the Guardian in 2005 (amounting to approximately 40 million words), a year that was particularly relevant from a journalistic point of view in the newspaper's recent biography (\*). The Guardian was selected in preference to other news outlets because of its interest in media matters, a focus both self-proclaimed by the newspaper itself and also empirically attested by a preliminary Keywords comparison of the Guardian with other British broadsheets (Marchi and Taylor 2009).

The analysis combines two different tools, Wordsmith and Xaira. The concordancers are used in a complementary way in order to optimise the employment of the corpus XML mark-up, which, as it will be reported, proved to be useful in achieving a finer-grained assessment of the data.

Journalism is a segmented and undetermined occupation and journalist is 'a label that people engaged in a diverse range of activities apply to themselves' (Tunstall 1971: 69). The analysis starts by defining candidate items for the investigation and by tracing a lexical profile of the terms *JOURNALIST* and *REPORTER*. By means of a detailed collocation analysis it is shown how the two words are used in different ways and how the patterns associated with them point at some core matters of occupational identity: class, skills, professional ethics, and so on. According to Hampton (2005) there is a difference of status between 'real journalists' and 'mere shorthand reporters'; the focus on the discursive behaviour of the two labels allows us to open a window onto the varied meanings of journalism and the multiple discourses practitioners enact about their profession, in particular with reference to a "good" vs "bad" journalism dialectic.

(\*) In 2005 the Guardian underwent a transformation, changing to the smaller Berliner format. The transformation received extensive coverage and resulted in a broader debate about newspapers and journalism.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Carey, J, (1997). 'The Dark Continent of American Journalism'. In E. Stryker Munson and C. A. Warren (eds.) *James Carey: A Critical Reader*. Minneapolis MN: University of Minnesota Press, pp. 144-88.

Hampton, M. (2005). 'Defining Journalists in Late-Nineteenth Century Britain'. *Critical Studies in Media Communication* 22(2): 138-55.

Marchi, A. and Taylor, C (2009). 'If on a Winter's Night Two Researchers... A Challenge to Assumptions of Soundness of Interpretation'. *Critical Approaches to Discourse Analysis across Disciplines* 3 (1): 1 – 20 ISSN: 1752-3079.

Matheson, D. (2003). 'Scowling at their notebooks. How British journalists understand their writing'. *Journalism* 4(2): 165-83.

McEnery, T. and Gabrielatos, C. (2006). 'English Corpus Linguistics'. In Aarts, B. and McMahon, A. M. S. (eds.) *The Handbook of English Linguistics*. Oxford: Blackwell, pp. 33–71.

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*, pp. 261-303. London: Routledge.

Tunstall, J. (1971). *Journalists at Work*. London: Constable.

Zelizer, B. (2004). *Taking Journalism Seriously. News and the Academy*. Thousand Oaks: Sage.

**Tony McEnery and Andrew Hardie (Lancaster University)**

#### Research ethics in corpus linguistics

While research ethics are as critical for corpus linguistics as for any other branch of linguistics, relatively little consideration has been paid in the literature to ethical issues in corpus construction and exploitation. Although some authors have directly considered their work in relation to ethical issues, for example Hasund (1998), Sampson (2000) and Rock (2001), the central textbooks in the field, including Sinclair (1991), Kennedy (1998), Biber et al. (1998), and McEnery and Wilson (2001), do not treat ethical issues in any depth. This may be because corpus linguists have in many cases 'inherited' their ethical good practices from guidelines developed, for example, for applied linguistics in general. For example, the British Association of Applied Linguistics has a well-developed set of ethical guidelines which are clearly relevant to corpus builders (see [http://www.baal.org.uk/dox/goodpractice\\_full.pdf](http://www.baal.org.uk/dox/goodpractice_full.pdf)).

We will argue, however, that there are questions specific to corpus linguistics which may not be fully addressed by guidelines from outside the field, and thus that research ethics is an area that corpus linguistics should consider in more detail. There are four main groups of such questions. Firstly, in collecting a spoken corpus there are ethical issues relating to the respondents. These relate primarily to privacy – not only of the respondent, but also of the people they are recorded speaking to, and moreover of the people they are recorded talking about. A second set of ethical issues must be addressed in the process of construction of a written corpus. In particular, what is the appropriate attitude to take towards potentially offensive, immoral, or illegal textual data? A third group of questions relate to the sometimes vexed question of the distribution of corpus data. To what extent are corpus distributors ethically obliged to consider whether the purposes to which the data will be put would be approved of by the original donors/collectors of the data? Finally, there are issues that must be faced by any user of corpus data – in particular, the ethical imperatives to take all steps to make sure their analysis is replicable, and to record and preserve aspects of the research method that underlie, but are not contained within, their published results.

While the corpus linguistic literature is mostly silent on ethical issues, it does generally embody good

ethical practice. There are, however, a number of exceptions – instances of relatively poor practice in published corpus research. Some occurred during the infancy of corpus linguistics as a (sub-) discipline, but some are more recent. We will review some examples of such poor practice, and suggest that, as a corrective to these relatively prominent bad examples, it is high time that more explicit regard is given to issues of research ethics in corpus linguistics.

Biber, D, Conrad, S and Reppen, R (1998) *Corpus linguistics: investigating language structure and use*. Cambridge: CUP.

Hasund, K. (1998) 'Protecting the innocent: the issue of informants' anonymity in the COLT corpus', in A. Renouf (ed.) *Explorations in Corpus Linguistics*, Rodopi, Amsterdam, pp 13-28.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. Harlow: Longman.

McEney, T. and Wilson, A. (2001) *Corpus Linguistics* (2nd edition). Edinburgh: EUP.

Rock, F. (2001) 'Policy and Practice in the Anonymization of Linguistic Data', *International Journal of Corpus Linguistics*, Volume 6, Number 1, pp 1-26.

Sampson, G.R. (2000) *CHRISTINE Corpus, Stage I: Documentation*. Available at [www.grsampson.net/ChrisDoc.html](http://www.grsampson.net/ChrisDoc.html)

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

#### **Tony McEney and Amir Salama (Lancaster University)**

##### **De/Victimizing Christian Copts in/outside Contemporary Egypt: A Critical Corpus-Based Study**

This study is critically oriented towards exploring the conflicting discourses that have arisen out of the ideological clash between the majority Muslim community in Egypt and the minority expatriate Christian Copts, who either immigrated to or were born in the US. The study investigates the different discursive representations of Christian Copts in and outside Egypt across two corpora. The first is an Arabic specialized corpus of the Egyptian national Arabic newspaper of Al-Ahram (2000 - 2009) (<http://www.ahram.org.eg/>). This corpus contains 111,323,089 words; it comprises 12 domains: Egyptian local news, Arab world news, international Politics, Opinions, Columns, Al-Ahram Writers, Economy, Sports, Culture and Arts, Women and Children, Al-Ahram Files, science and Medicine. The second is an English corpus with 466,504 words; it includes all the articles archived on the website of the U.S. Copts Association (2000 – 2009) (<http://www.copts.com/english/>).

Drawing on a 'methodological synergy' of corpus linguistics (CL) and critical discourse analysis (CDA) (Baker et al. 2008), we conducted a contrastive collocational analysis which reveals opposing discourses on Christian Copts. We use the corpus software CQPweb (Hardie forthcoming) to identify the 'statistically significant collocates' (Church, Hanks & Moon 1994) of the node words Copts and Coptic in the English Corpus and the corresponding terms in the Arabic corpus. We then use CDA to describe the inter-collocate relations and their respective 'discourse prosodies' on Egypt's Christian Copts across the two corpora. We focus upon nominalization, mitigating vs. intensifying strategies, the use of nationalist inclusive vs. dissenting exclusive metaphors, stance and perception modality, and opposed 'social actors'.

The overall goal is to explore how these conflicting discourses may impact upon the social situations in which the texts in question are produced and read.

**Peter Menke (Universität Bielefeld, Germany) and Alexander Mehler (Goethe-Universität**

**Frankfurt am Main, Germany)**

## From experiments to corpora: The Ariadne Corpus Management System

This paper describes the Ariadne Corpus Management System that assists researchers in all stages of generating multimodal speech corpora. It accompanies users from data acquisition via data analysis to the final publication of corpora. Ariadne is available to registered users in two variants:

1. as a web-based application that can be accessed from any computer connected to the Internet,
2. and as a client application that allows for a seamless integration with the user's file system structure.

Ariadne's field of application is characterized by tasks that scientists recurrently perform in research on multimodal communication:

- ⤴ the encoding of observations elicited in experiments into a machine-readable format (transcriptions or annotations), possibly with the aid of different tools and their disparate data formats,
- ⤴ the performance of various analyses on these data sets,
- ⤴ the publication of subsets of resulting data as corpora in some interoperable format that is readable and further processable by others,
- ⤴ and the compliance to privacy policies (e.g., anonymity of participants).

Ariadne assists at tasks from all these areas. First, it is built on top of a generic data model of communicative events, in combination with an expressive system of types. These types model various conditions and restrictions, and, when combined into special bundles, can accurately express the constraints of data formats from many popular third-party transcription and annotation tools. This mechanism helps to avoid unexpected alteration or fragmentation of data by specifically predicting what changes data will undergo when using a certain processing routine.

Data sets in the central format can then be processed by other modules inside Ariadne that perform part-of-speech tagging and lemmatization, calculation of measures of inter-annotator agreement, correction of values against value rules or vocabularies, and syntactic parsing. In addition, various functions from the field of statistics are provided -- either in the form of data preparation for input into third-party software or by performing analyses directly inside Ariadne.

One of the recent enhancements of Ariadne are components for the preparation and publication of corpora and related linguistic resources:

- ⤴ Metadata modules for the collection of bundles of metadata which conform to different standards for the publication of language resources. These data sets make it easier for specialised search engines and crawlers to access and index resources properly.
- ⤴ Interfaces and mechanisms for a flexible selection and publication of data to the public, in order to achieve compliance to the demands of various funding institutions.
- ⤴ Closely related to these enterprises is the goal of providing RDF functionality for complete corpora or resources and their fundamental components.

Our presentation and full paper will give a more exhaustive introduction to the functionality of the system. Examples of studies currently conducted in a research center on multimodal communication will serve as illustrations.

Gleim, R. & Mehler, A. (2010): 'Computational Linguistics for Mere Mortals --- Powerful but Easy-to-use Linguistic Processing for Scientists in the Humanities'. Proceedings of LREC 2010, ELDA

Klyne, G. & Carroll, J.J. (2004): 'Resource Description Framework (RDF): Concepts and Abstract Syntax'. W3C, <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210>

Menke, P. & Mehler, A. (2010): 'The Ariadne System: A flexible and extensible framework for the modeling and storage of experimental data in the humanities'. Proceedings of LREC 2010, ELDA

**Thomas Meyer (Idiap Research Institute, Martigny, Switzerland), Charlotte Roze (Alpage Group, INRIA and University of Paris VII, France), Bruno Cartoni (Department of Linguistics, University of Geneva, Switzerland), Laurence Danlos (Alpage Group, INRIA and University of Paris VII, France), Sandrine Zufferey (Department of Linguistics, University of Geneva, Switzerland), and Andrei Popescu-Belis (Idiap Research Institute, Martigny, Switzerland)**

Disambiguating discourse connectives using parallel corpora: senses vs. translations

Discourse connectives are words or phrases that indicate senses holding between two spans of text. The theoretical approaches accounting for these senses, such as text coherence, cohesion, or rhetorical structure theory, share at least one common feature: they acknowledge that many connectives can indicate different senses depending on their context. For instance, in English, 'while' can sometimes indicate a temporal sense, but other times a comparison, an opposition, or a concession. Depending on its sense, the translation of a connective into another language can vary greatly, either using an equivalent connective, or using a different construction or even no explicit connective at all.

The objective of this study is to characterize the multifunctionality of a subset of connectives which are both, frequent and ambiguous. We will define the main senses of each connective, describe a reference annotation of connectives with their senses in parallel corpora, and make quantitative observations on the frequencies of senses and their translations. The parallel texts are English/French parliamentary debates (with known source language and its direct translation) from the Europarl (Koehn, 2005) and Hansard (Roukos et al., 1995) corpora.

Two possible approaches to corpus-based studies of connectives have been explored in the past. Our objective is to show that combining the two produces richer and more reliable results.

The first approach provides annotators with descriptions of the possible senses of each connective, and requires them to label each occurrence with one sense, as in the English Penn Discourse Treebank (Prasad et. al., 2008). Similarly, Roze et. al. (2010) have identified possible senses of French connectives in the LexConn database, with 328 connectives totaling 428 possible senses. The senses and their definitions are currently used for annotating English and French texts.

The second approach considers the translations of connectives observed in parallel corpora – e.g. like in the study of causal connectives in French/Dutch novels by Denturck (2010). Our observations on temporal/contrastive or causal connectives show that beyond the large variety of possible translations, there are dominant clusters of translations corresponding to the main senses identified monolingually. For instance, the French connective 'alors que' has four frequent translations into English in the Hansard corpus: ca. 50% 'even though', 10% 'when', 5% 'given that', and 10% of no direct lexical equivalent. These translations reflect its multifunctionality as an indicator of concession or a temporal sense. We will present findings for temporal/contrastive and causal connectives such as 'while', 'since' in English and 'alors que', 'en effet', 'parce que', 'car', and 'puisque' in French, with respect to use in original texts and their translations.

As a result, a multilingual database of connectives will be constructed, including descriptions of their senses and principal translations, augmented with frequency information from parallel corpora. The annotated resource will be used for training and testing an automatic system that disambiguates connectives, as a preliminary stage to their automatic translation.

Denturck, Kathelijne (2010): 'Translation universals: the case of causal connectives in French and Dutch translations. A corpus-based study'. Workshop Connectives across Languages: Explicitation and Grammaticalization of Contiguity Relations.  
<http://www.francais.ugent.be/index.php?id=25&type=file> [16.11.2010].

Koehn, Philipp (2005): *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT Summit 2005.

Roze, Charlotte, Laurence Danlos and Phillippe Muller (2010): 'LEXCONN: a French Lexicon of Discourse Connectives'. Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010).

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008): The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Roukos, Salim, David Graff, and Dan Melamed (1995): *Hansard French/English*. Linguistic Data Consortium, Philadelphia.

#### **Nozomi Miki (Kansai University of International Studies)**

##### **Key Colligation Analysis: Discovering stylistic differences in significant lexico-grammatical units**

Grammatical subjects are not always but very often themes—'a departure of message' in the clause-initial position. Many previous studies have indicated the significant discursual functions of subjects in argumentative prose, including academic writing and editorials (Hyland, 2001, 2002; Gosden, 1993; Francis, 1989; Hawes and Thomas, 1996). However, the way to retrieve subjects with the predicates has mostly been manual, making investigations so time-consuming as to limit the number of texts being investigated and thus making generalisation hazardous. Otherwise, the main targets were pronominal subjects instead of lexical ones, in quantitative research above all. As a solution, this methodology-based research will propose Key Colligation Analysis for the identification of significant lexico-grammatical units in target datasets. In this analysis, a target corpus and the reference corpus were parsed with *Connexor Machine Syntax for English* (Connexor Oy, 2008). The programming scripts were designed to retrieve colligations, such as subject-verb and subject-auxiliary-verb pairs in the main clause, with frequencies from parsed texts and to make lists of them such that frequencies of the same sequences can be compared on the basis of log-likelihood ratios and cut off at  $p < .001$ .

This was demonstrated with editorials from British broadsheet newspapers (*The Times*, *The Guardian*, *The Daily Telegraph*, and *The Independent*). Each dataset contained about 2 million words (2004 to 2010) and the reference corpora were composed of the editorial corpora excluding a target corpus. I investigated sequences of subjects, verbs and auxiliaries, if any, distinguishing active/passive voices (e.g., *it can be argued*, *Blair said*). The results revealed categorical stylistic differences among the newspapers. *The Daily Telegraph* and *The Independent* preferred 'we' but in different ways. *The Daily Telegraph* used this pronoun to introduce to readers the background of arguments in their favour (e.g., *as we can see ...*) and also to make evaluations (e.g., *we hope ...*) in an assertive way, as compared with *The Independent*, which employed 'we' to hedge their strong claims with 'should' and 'must' (e.g., *we must hope ...*) as well as to pass judgement (e.g., *we need ...*). In *The Times*, *The Guardian* and *The Independent*, there were more outstanding nominalisations

including labelling nouns (Francis, 1994) and pronominal subjects (e.g., *it, that, there*) showing variations between them. The four leading newspapers can be classified in a cline of personality and impersonality and also reveal different uses of intertextuality.

This way of identifying characteristics is similar in some ways to Keyword Analysis (Scott, 2000, 2001, 2010) and Key Cluster Analysis (Baker, 2006); but Key Colligation Analysis is unlike these two in targeting phrases as well as words and in capturing abstract, grammatical relations of sequences. This approach can be applied to the investigation of verbs of attribution in other fields, such as academic writing.

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

Connexor Oy (2008). *Connexor Machine Syntax for English* (Ver. 3.9.3.4) [Computer software]. Helsinki: Connexor Oy.

Francis, G. (1989). 'Thematic selection and distribution in written discourse'. *Word*, 40(1-2), 201-221.

Francis, G. (1994). 'Labelling discourse: An aspect of nominal-group lexical cohesion'. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 81-101). London: Routledge.

Gosden, H. (1993). 'Discourse functions of subject in scientific research articles'. *Applied Linguistics*, 14(1), 56-75.

Hawes, T. & Thoms, S. (1996), 'Rhetorical uses of theme in newspaper editorials'. *World Englishes*, 15(2), 159-170.

Hyland, K. (2001). 'Humble servants of the discipline? Self-mention in research articles'. *English for Specific Purposes*, 20, 207-226.

Hyland, K. (2002). 'Authority and invisibility: authorial identity in academic writing'. *Journal of Pragmatics*, 34, 1091-1112.

Scott, M. (2000). 'Focusing on the text and its key words'. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 103-121). Frankfurt: Peter.

Scott, M. (2001). 'Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs'. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus studies and ELT: Theory and practice* (pp. 47-67). Amsterdam: John Benjamins.

Scott, M. (2010). *WordSmith Tools* (Version 5.0) [Computer software]. Oxford: OUP.

#### **Jiří Milička (Charles University in Prague)**

##### **Valency and Information Structure in Arabic Text: A Quantitative Approach**

Managing valency frames is a basic prerequisite for successful communication and a great effort was put in its description either for the purposes of language acquisition or processing. But a sensitive language user feels that the issue of valencies is more complicated than traditional and generative grammar expects. For example one can have the impression that when a verb has two arguments, their order is not random, but it is not fixed as well, and that it is to some extent determined by the intention of the speaker or by information flow or that it may be influenced by phonological patterns of the surrounding text etc. The main purpose of this paper is to convert these subjective

impressions into falsifiable hypotheses and to show which of them are proved false. A suitable way to achieve that is to rely on quantification and to “trust the text”.

The paper is based on research that consists of two parts:

The first one lies in finding verbs and their valencies in an Arabic text. For this purpose, a valency extraction algorithm was designed and applied on a large corpus of modern standard Arabic (about 40 million word tokens) and a diachronic Arabic corpus (about 380 million word tokens). The algorithm does not employ any syntactic formalism and it is based on collocations processing. This approach allows us to exploit untagged corpora and makes the result independent of our assumptions. Advantage is taken of characteristic features of Arabic morphology: there are some forms of Arabic verbs distinct from all forms of other parts of speech and (to some extent in some cases) it is possible to avoid ambiguity quite easily, especially when comparing with English and some other Indo-European languages. This not only enables finding Arabic verbs automatically, it also means that the set of Arabic verbs is combinatorially limited (even though we for some reasons cannot claim that it is a finite set). The phenomena and the algorithm are described in the paper.

In the second part of the research, the gained output is analysed and the algorithm adjusted accordingly. Subsequently the data are interpreted and processed by various statistical techniques. The tested hypotheses and other outcomes introduced in the paper are illustrated by charts and examples from the Arabic corpora.

Possible improvement of the algorithm and techniques used or applying these within other languages are to be left for discussion.

**Neil Millar (University of Birmingham), Brian Budgell (Canadian Memorial Chiropractic College), and Keith Fuller (University of Toronto)**

#### Passive constructions in a corpus of Randomised Controlled Trials

The purpose of the present paper is (1) to provide an overview of the activities of The Centre for Biomedical and Health Linguistics ([www.bmhlinguistics.org](http://www.bmhlinguistics.org)), an organization dedicated to facilitating communications in biomedicine and health; and (2) to present the results of analyses of passive constructions in a corpus of reports of Randomised Controlled Trials (RCTs) – a ‘gold standard’ methodology for testing the efficacy of healthcare interventions.

Writings in biomedicine and health have been criticized for overuse of the passive construction – a structure which critics claim results in verbosity and a lack of clarity (Sheen 1982, Albert 2004). As research in this field is performed with the intention of influencing and improving health care policies and practices, clarity is tantamount. However, authors seeking to publish in biomedical journals are presented with conflicting messages. On the one hand, publishing guidelines for journals and writing guides may discourage use of the passive. On the other hand, in the same journals passive constructions are highly frequent (Amdur et al. 2010) and, therefore, represent an established convention in biomedical writing. It would seem therefore that guidelines which state only, for example, that authors should “[u]se active voice” (Annals of Emergency Medicine 2010) are over-simplistic.

BMH Linguistics, a collaboration between linguists and biomedical/health researchers, aims to enhance understanding of “biomedical English” (the lingua franca of the field). We present here an overview of the activities of BMH Linguistics, focusing on the creation and analysis of corpora of literature from targeted domains in biomedicine and health (e.g. Millar & Budgell 2008) and the development of web-tools to provide access to the corpora. The study of passive constructions in RCTs illustrates the work of BMH Linguistics and how corpus analyses can help facilitate effective

communication.

The RCT corpus comprises all reports of RCTs published in the five top ranking medical journals in 2005 (298 articles; c. 1.2 million words). Using part-of-speech annotation, passive constructions (c. 19,700 in total) were extracted from the corpus and analysed. Results show that long passive constructions, where agent is mentioned, are relatively rare. The distribution of passives within articles indicates that it is used most frequently in methods and results sections – although frequency varies greatly from article to article. Statistical analyses of verbs most strongly associated with the passive construction also indicate that passive constructions are strongly associated with expository functions. The implications of these findings for author guidelines are discussed.

Albert, T. (2004). 'Why are medical journals so badly written?' *Medical Education*, 38(1), 6-8.

Amdur, R., Kirwan, J., & Morris, C. (2010). 'Use of the passive voice in medical journal articles'. *American Medical Writers Association Journal*, 25(3), 98-104.

Annals of Internal Medicine. (2010). 'Instructions for Authors'. Available: <http://www.annemergmed.com/content/instauth>. Last accessed October 31, 2010.

Millar, N., & Budgell, B. (2008). 'The language of public health—a corpus-based analysis'. *Journal of Public Health*, 16(5), 369-374.

Sheen, A. P. (1982). *Breathing life into medical writing: A handbook*. St Louis: Mosby.

#### **Neil Millar (University of Birmingham)**

##### **Collocation and predicative inferencing in reading: Evidence from eye-movement studies**

Probabilistic constraints in language are not uncommon (Bod, Hay, & Jannedy, 2003) and would seem to play an important role in the way language is learnt, processed and used (Seidenberg & MacDonald, 1999). Collocation, the probabilistic tendency of certain words to co-occur, is pervasive, both as a phenomenon observed in natural language use and as the object of study in corpus linguistics. Studies of collocation in corpora (and its extensions – colligation, semantic preference and semantic prosody) have formed the basis for several claims relating to the nature of language in the mind and/or how it is processed (e.g. Hoey 2005, Sinclair 1991, Louw 1993). Although these claims are essentially psychological in nature, few researchers have sought to investigate how corpus derived collocations are actually processed in real-time. The present paper reports two experiments which used the eye-movement paradigm to investigate how native speakers of English process lexical collocation errors and morphological colligation errors extracted from a corpus of learner English.

In experiment 1, a word-by-word self-paced reading procedure was used to compare reading times for sentences containing either a learner collocation error (e.g. heavy crime) or colligation error (e.g. responsibility person) to sentences containing a formulaic native speaker equivalent (e.g. serious crime/responsible person). Using eye-tracking methodology, experiment 2 explored native speaker processing of comparable stimuli under more naturalistic reading conditions. Results show that, in comparison to the native speaker equivalent, (1) learner collocation errors are associated with an increased and sustained processing burden,(2) that the size and duration of the burden is substantially greater for morphological errors (colligation errors) than for lexical errors (collocation errors), and (3) that morphological errors are detected earlier.

The results indicate that colligational 'primings' are stronger than collocational 'primings' – a finding which, it is argued, provides support for Hoey's (2005) theory of Lexical Priming and usage-based

models of language. Drawing on recent theories from cognitive science (in particular, Pickering & Garrod 2007), some explanatory hypotheses are put forward. It is proposed that differences in the size and time-course of participants' response to collocation and colligation errors are indicative of the use of probabilistic knowledge to predict upcoming input. It is suggested that the phenomenon of collocation can support predictive inferences of upcoming input in language comprehension, and, thus, aids fluent language processing. Theoretical and methodological implications for collocation research are discussed.

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Hoey, M. (2005). *Lexical priming: A New Theory of Words and Language*. London: Routledge.

Louw, B. (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. 157-176). Philadelphia, PA: John Benjamins Publishing Company.

Pickering, M. J., & Garrod, S. (2007). 'Do people use language production to make predictions during comprehension?' *Trends in Cognitive Sciences*, 11(3), 105-110.

Seidenberg, M. S., & MacDonald, M. C. (1999). 'A probabilistic constraints approach to language acquisition and processing'. *Cognitive Science*, 23(4), 569-588.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*: Oxford U.P.

**Rosamund Moon (Department of English, University of Birmingham) and Carmen Rosa Caldas-Coulthard (Centre for English Language Studies, University of Birmingham)**

#### Ageing with the corpus

This paper reports on a corpus study of stereotypes of age and ageing, including gender-based stereotypes. Data was drawn from a large reference corpus, and, in order to identify stereotypes, we started by examining collocates of central items: age indicators such as *young* and *old*, and primary gender-marked words such as *man* and *woman*. Our study and findings are in keeping with other discourse-oriented corpus research in these areas: relevant studies include those of *elderly* (Baker 2006, Mautner 2007), *man/woman* (Pearce 2008), *boy/girl* (Holmes and Sigley 2001, Sigley and Holmes 2002), *man/woman/girl/boy* (Caldas-Coulthard and Moon 2010), and *bachelor/spinster* (Baker 2006, Stubbs 2001).

In our paper, we will focus mainly on what we learned from adjective collocates of central items: these seemed particularly revealing in terms of institutionalized evaluations. It was unsurprising that, for example, items clustering with the young were mainly positive (e.g. *beautiful, fresh, attractive, pretty, fit, talented, dynamic*) and those clustering with the old mainly negative (e.g. *sick, tired, infirm, frail, grey, fat, decrepit*), but the extent to which this happened was nonetheless dismaying. Factoring in gender produced a more nuanced picture: here we examined combinations such as *young man, middle-aged women, old lady* – all classifying people according to stage of life – to see what kinds of attribute are associated with men and women at each stage. What emerged were stereotypes that reflected rather traditional social roles/personas and social expectations: "fine young men", suitable as husbands/employees; attractive young women, suitable as mates; predatory older men and sad middle-aged men; old women as gossips or witches; the stubborn old. This is specific linguistic evidence for the age stereotypes referred to widely in the media and in the sociological literature, and it has provided us with benchmarks to use in analysing discourses beyond the confines of a reference corpus.

In the final part of our paper, we look at sets of adjectives which function as signifiers of age and evaluate accordingly. Some are obvious (*fresh, nubile, wrinkled, grey-haired*), but others perhaps less so: for example, *tall, slim, talented* are strongly associated with younger people; *tired, smelly, kindly* with old and older people. Deviant usage ("young" adjectives applied to the old, and vice versa) is likely to be signalled with *still, but, for one's age, prematurely*, etc. Also interesting are contexts where stereotypes are contested or resisted – for example, in dating ads placed by older people, or magazine features aimed at older readers – and here again corpus data provides some useful insights into the lexical strategies adopted. We conclude by discussing wider implications for studies of gender and age, sexism and ageism, in language.

Baker, P. (2006) *Using Corpora in Discourse Analysis*, London: Continuum.

Caldas-Coulthard, C.R., and Moon, R (2010) 'Curvy, hunky, kinky: using corpora as tools in critical analysis', *Discourse and Society* 21/2, 1-35.

Holmes, J. and Sigley, R. (2001) 'What's a word like *girl* doing in a place like this?', in A. Smith and P. Peters (eds.) *New Frontiers of Corpus Linguistics*, Amsterdam: Rodopi, 247-263.

Mautner, G. (2007) 'Mining large corpora for social information: the case of *elderly*', *Language in Society*, 36, 51-72.

Pearce, M. (2008) 'Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine', *Corpora* 3/1, 1-29.

Sigley, R. and Holmes, J. (2002) 'Looking at *girls* in corpora of English', *Journal of English Linguistics* 30/2, 138-157.

Stubbs, M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford: Blackwell.

**Nick Moore (Khalifa University, United Arab Emirates)**

Modelling the Flow of Discourse in a Corpus of Written Academic English

A key feature of discourse is the movement of information from one clause to the next through text. Few linguistic models are able to account for the development of ideas through discourse, and this area presents major challenges for computational and corpus-based approaches to language description. While computational studies have shown that anaphora play a part in the flow of information, they cannot tell the whole story (Beaver, 2004; Botley and McEnery, 2000). Similarly, identifying and following the development of thematic elements can also help to outline the dynamic progression of a text (Ravelli, 1995). However, it is only when these two grammatical systems are combined with an analysis of information structure that we can approach the issue of information flow through a text (Fries, 2002).

A Systemic Functional model of written discourse allows for multiple functions to operate simultaneously on the clause in order to account for meaningful choices. The Textual Metafunction employs the clausal functions of Participant, Theme and Information to achieve its aim of contextualising and instantiating a text. Participant Identification and Tracking can be used to reveal the complex semantic relationships between entities built up in a text through referential signals (Caselli and Prodanof, 2006; Martin, 1992; Moore, 2008). Theme establishes the ground of the discourse from where the current clause originates and reveals the thematic progression, or method of development, of a text (Crompton, 2004; Daneš, 1974), while the writer chooses what is the most important part of the clause and consequently directs the reader's attention to this New Information (Fries, 2002; Halliday and Matthiessen, 2004). Combining these three textual systems produces a

recognisable flow of information.

A model that combines these three textual functions, describes the quantitative and qualitative interactions between them, and makes predictions about the expected effects on readers would offer one step forward in computationally modelling discourse. The proposed model of information flow was evaluated against a corpus of academic texts from engineering disciplines (about 10,000 tokens). Results of the analysis of the three grammatical systems in the Textual Metafunction demonstrate significant patterns, or unmarked choices, where the participant, thematic and information systems combine to powerful effect. Where the systems are not aligned, there is a recognisable effect on the flow of information. Examples from the analysed corpus are used to compare information flow in discourse in texts exhibiting a more or less successful flow of information, and recommendations are offered for implementation in a computational model.

Beaver, D. 2004. 'The optimization of discourse anaphora'. *Linguistics and Philosophy* 27 p.3-56

Botley, S. and McEnery, T. 2000. 'Discourse anaphora: The need for synthesis'. In Botley & McEnery (eds.) *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam: John Benjamins

Caselli T. and Prodanof I. 2006. *Annotating Bridging Anaphors in Italian: in Search of Reliability*. In: LREC 2006 - 5th International Conference on Language Resources and Evaluation (Genoa, Italy, 24-25- 26 May 2006). Proceedings, pp. 1173 - 1176. European Language Resources Association (ELRA)

Crompton, P. 2004. 'Theme in discourse: 'Thematic progression' and 'method of development' re-evaluated'. *Functions of Language* 11/2 p.213-249

Daneš, F. 1974. 'Functional sentence perspective and the organization of the text'. In Daneš (ed.) *Papers on Functional Sentence Perspective*. Prague: Academia

Fries, P.R. 2002. 'The flow of information in a written text'. In Fries, Cummings, Lockwood & Spruiell (eds.) *Relations and Functions within and around Language*. London: Continuum

Halliday, M.A.K. and Matthiessen, C.M.I.M. 2004. *An Introduction to Functional Grammar (Third Edition)*. London: Arnold

Martin, J. R. 1992. *English text: System and structure*. Amsterdam: John Benjamins

Moore, N. 2008. 'Bridging the metafunctions: Tracking participants through taxonomies'. In Jones & Ventola (eds.) *From Language to Multimodality: New Developments in the Study of Ideational Meaning*. London: Equinox

Ravelli, L. 1995. 'A dynamic perspective: Implications for metafunctional interaction and an understanding of Theme'. In Hasan & Fries (eds.) *On Subject and Theme: A Discourse Functional Perspective*. Amsterdam: John Benjamins

**Akira Murakami (University of Cambridge)**

Cross-linguistic influence on the accuracy order of English grammatical morphemes: Insights from a learner corpus

Ever since the study of Dulay and Burt (1973, 1974), it has been accepted in the SLA literature that the L2 acquisition of English grammatical morphemes follows the so-called "natural order" and that L1 has little influence on the order of acquisition. Recently, however, Luk and Shirai (2009)

challenged the view by surveying the studies examining the order of acquisition of English grammatical morphemes, and demonstrated that the order, in fact, may differ depending on L1s.

The purpose of the present study is to directly compare the L2 accuracy orders of English grammatical morphemes across different L1 groups and provide the first-hand evidence of L1 (non-)influence on this theme of SLA. The study exploited the Cambridge Learner Corpus and targeted over 10,000 learners of seven typologically diverse L1s (Japanese, Korean, Spanish, Russian, Turkish, German, and French) across five proficiency levels (roughly corresponding to A2 through C2 in the Common European Framework of Reference). Six target morphemes were chosen from an influential study by Goldschneider and DeKeyser (2001); progressive *-ing*, past tense *-ed*, articles, third person *-s*, plural *-s*, and possessive *'s*.

In order to identify meaningful differences between the target-like use (TLU) scores, morphemes within each L1 and each proficiency level were clustered based on the 95% confidence intervals of the TLU scores obtained through bootstrapping. The clustered orders within each proficiency level were compared across L1 groups. Besides the TLU-based clustering above, the Spearman's rank-order correlations of the order of suppliance in obligatory contexts (SOC) scores were calculated for each pair of observed orders, and the correlations of within-L1 pairs were compared against those of between-L1 pairs. If the former is stronger, L1 influence is likely to be operative.

The two analyses above demonstrated clear influences of L1 on the accuracy order of English grammatical morphemes. Some prominent effects include (i) the accuracy order of articles by Japanese, Korean, Russian, and Turkish learners of English is consistently lower than that by other L1 learners of English, (ii) German learners tend to mark a lower accuracy rank of progressive *-ing* than other L1 learners, and (iii) the accuracy order of Spanish learners of English does not deviate from the natural order, which confirms to a Luk and Shirai's (2009) hypothesis that the natural order is a mere reflection of the acquisition order by Spanish learners of English. All in all, despite the commonly held assumption that L2 learners of English acquire grammatical morphemes in a fixed order, the evidence provided in this study is more than sufficient to cast a strong doubt on the universality of accuracy order.

Dulay, H. C., & Burt, M. K. (1973). 'Should we teach children syntax?' *Language Learning*, 23(2), 245-258.

Dulay, H. C., & Burt, M. K. (1974). 'Natural sequences in child second language acquisition'. *Language Learning*, 24(1), 37-53.

Goldschneider, J., & DeKeyser, D. (2001). 'Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants'. *Language Learning*, 51(1), 1-50. doi: 10.1111/1467-9922.00147

Luk, Z. P., & Shirai, Y. (2009). 'Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural *-s*, articles, and possessive *'s*'. *Language Learning*, 59(4), 721-754. doi: 10.1111/j.1467-9922.2009.00524.x

#### **Rogelio Nazar (Universitat Pompeu Fabra)**

Algorithm qualifies for C1 courses in German exam without previous knowledge of the language: an example of how corpus linguistics can be a new paradigm in Artificial Intelligence

This paper presents an experiment in which a computer program selects the correct answers in a German exam using corpus statistics as the only source of knowledge, that is, without any previous explicit knowledge of the German language. The result obtained with this experiment is the C1

qualification, which is equivalent to the level of German of an educated native speaker. The purpose of this paper is to show how sometimes poor-knowledge approaches based on relatively simple statistics can perform better than fully informed symbolic rule-based systems.

The Goethe Institut offers through its website a free online German exam called “Test your German”, designed for the people who wants to know which course they will best fit in. The exam comprises 30 multiple choice questions, basically filling the blank spaces in a sentence with the correct option from a number of choices. The questions range from very basic grammar knowledge, such as selecting the correct preposition or the correct number and/or gender agreement, to the most difficult part of the language which is to select the correct verb-noun or adjective-noun collocations.

Using the web as a corpus, our program is able to select the correct answers in most of the cases by querying a regular search engine with segments of the example sentence given in the exam filling the blank with each of the different choices. Naturally, the idea is that the query that offers the largest number of hits is selected as the correct answer. For illustration, consider the following example, provided also in Goethe Insitut's exam:

0. Die Familie Müller hat ----- Kinder.

Each question is followed by some choices. In the case of this example, the following four:

A kein B keine C keinen D keiner

The correct answer in this case is B. We can know that without knowing German by simply querying the search engine with the different possibilities to select the one that generates more results. Of course, it is unlikely to find results using the whole sentence as a query, but it suffices to try with one and two positions at each side of the blank. In the case of the above example, the different queries would look like the following: 1) “hat kein Kinder”; 2) “Müller hat kein Kinder”; 3) “hat keine Kinder”; 2) “Müller hat keine Kinder”; and so on. Indeed, the options are not strictly necessary for the solution, since most search engines support the insertion of wildcards (\*) in the query expressions. Thus, one can compose a query such as “hat \* Kinder” or “Müller hat \* Kinder”; etc. The solution in this case is to see which is the most frequent element in the position of the wildcard in the snippets returned by the search engine.

The simplicity of this idea has, of course, no point of comparison with the complexity and the amount of information needed for achieving a similar result with symbolic and linguistically informed rule-based systems.

**Richard Nickalls (EISU, University of Birmingham)**

The use of an error-tagged learner corpus to investigate L1 Mandarin learners' English article interlanguage before and after explicit grammar teaching

This paper presents the findings of a study investigating the changing patterns of overuse and underuse of English articles by 24 L1 Mandarin students using a small error-tagged learner corpus developed from the essays of learners during a three month University preessional course in 2010. In contrast to the common trend of learner corpus research which most commonly compares one L1 group's accuracy with another (Diez-Bedmar and Papp, 2008), this study was designed to longitudinally analyse the same learners' changing accuracy throughout their 3 months of study.

The research presented in this paper is motivated by an interest in one of the most frequent and challenging choices that learners of English have to make: the choice of the, a/an, zero or null

articles in their writing. Whether or not target like use of the English article is perceived as necessary or possible for international students by the research community, there is a clear case for focussing research into learner interlanguage and L1 transfer upon frequently occurring words. Since the definite article is the most frequent word found in language corpora (Sinclair, 1991), it is particularly useful for research into effects of explicit grammar teaching in the classroom (for which this corpus of learner English is a small part of). The only possible challenge for this honour of being the language's most frequent word is the zero article (hence  $\emptyset$ ) which has been argued to be the most frequently occurring free morpheme in the English language (Master, 1997). Even if such zero articles are ignored, it has been pointed out that the/a/an together account for about one in every ten words (Berry, 1991).

Without wishing to claim that the following findings can be generalised before larger scale studies are completed, a profile of the learners' most fossilised errors will be presented to show how an effectively error-tagged corpus can help identify and profile learner errors to the end of testing theories about the effect of the L1 on the L2. Preliminary findings suggest that 1) accuracy improvements in most uses were not sustained when learners' attention stopped being focussed upon article use and that 2) these Upper Intermediate L1 Mandarin learners had little difficulty with  $\emptyset$ , marginally more frequent problems with the definite article and much greater problems with a/an. After some discussion of the difference between this hierarchy of accuracy ( $\emptyset > \text{the} > \text{a}$ ) and Diez-Bedmar and Papp's (2008) corpus based finding ( $\emptyset > \text{a} > \text{the}$ ) the final conclusion made will be that learner corpus research into 'native like' language is inherently difficult to replicate and that a more longitudinal approach using more data from each learner may be a fruitful direction for corpus research.

Berry, R. (1991) 'Re-articulating the article'. *ELT Journal*, 45: (3): 252-259.

Diez-Bedmar, M.B. and Papp, S. (2008) 'The use of the English Article System by Chinese and Spanish learners'. In Gilquin, G.; Papp, S. & Diez-Bedmar, M.B. (Eds.) *Linking up Contrastive and Learner Corpus Research*. Amsterdam, Atlanta, Rodipi 147-175.

Master, P. (1997) 'The English Article System: Acquisition, Function and Pedagogy.' *System*, 25: (2): 215-232.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

**David Oakey (Iowa State University, USA) and Peter Mathias (Bridge Research and Development, UK)**

**Investigating Interdisciplinary Discourse: Corpus-Driven Indicators of Emerging Epistemologies**

Academic research is becoming more interdisciplinary in scope as collaboration between researchers in different disciplines is increasingly relied on as a route to new knowledge. Such collaborators often need to become familiar with the language used in each other's home disciplines (Committee on Science, 2004) as there is recent evidence of the problems caused by poor communication between unprepared researchers (Adam, 2010). However, while funding agencies recommend linguistic familiarization classes for such researchers (Committee on Science, 2004), there is currently little guidance on the linguistic features to be included in the syllabi for such programs. Furthermore, although much has been discovered about the language of discrete disciplines (e.g. Swales, 1990; Hyland, 2000; 2009) in the field of English for Specific Purposes (ESP), the few descriptions of interdisciplinary discourse in the literature (e.g. Samraj, 1995; Teich & Holtz, 2009) do not primarily have a pedagogical focus.

This paper attempts to address this gap by comparing the language frequently used in an

interdisciplinary field with the language frequently used in two "contributory" fields in which the interdisciplinary collaborators normally work. In this study, the interdisciplinary field is that of Interprofessional Care, which contains work by researchers from the contributory fields of Medicine and Social Work. We investigate epistemological differences and similarities between the interdisciplinary and contributory fields through an isotextual comparison (Oakey 2009) of three subcorpora containing 100 research articles from journals in each field.

In order to investigate the epistemological contributions of Medicine and Social Work to Interprofessional Care, we use two corpus-driven phraseological features: frequently occurring lexical bundles (Biber et al, 1999; Cortes 2004, Hyland 2008), and collocations of closed class keywords (Gledhill, 2000; Groom 2007). A comparison using lexical bundles reveals differences in the use of fixed lexico-grammatical patterns in the different fields. Collocations of closed class keywords offer a picture of more flexible lexico-grammatical forms. Much more than single vocabulary items, both forms offer insights into the research goals of these fields and the evaluative stances taken by researchers towards their claims.

We use the results of the investigation to address the question of whether each of these fields has its own discrete epistemological profile, or whether one contributory field 'crowds out' the other in the interdisciplinary discourse. The paper concludes by discussing the application of the findings in English for Specific Purposes (ESP) programs for collaborating scholars.

Adam, D. (2010). 'Climate Scientists Hit Out At 'Sloppy' Melting Glaciers Error'. The Guardian. Accessed 8 February 2010  
[www.guardian.co.uk/environment/2010/feb/08/climate-scientists-melting-glaciers](http://www.guardian.co.uk/environment/2010/feb/08/climate-scientists-melting-glaciers)

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

Committee on Science, E., and Public Policy. (2004). *Facilitating Interdisciplinary Research*. Washington, D.C.: National Academy of Sciences, National Academy of Engineering, Institute of Medicine.

Cortes, V. (2004). 'Lexical bundles in published and student writing in history and biology'. *English for Specific Purposes*, 23(4), 397-423.

Gledhill, C. (2000). 'The discourse function of collocation in research article introductions'. *English for Specific Purposes*, 19, 115-135.

Groom, N. W. (2007). *Phraseology And Epistemology In Humanities Writing: A Corpus-Driven Study*. Unpublished Phd Thesis, University Of Birmingham, Birmingham.

Hyland, K. (2000). *Disciplinary Discourses*. Harlow: Longman.

Hyland, K. (2009). *Academic Discourse*. London: Continuum.

Oakey, D. J. (2009). 'Fixed collocational patterns in isolexical and isotextual versions of a corpus'. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 142-160). London: Continuum.

Samraj, B. T. R. (1995). *The Nature Of Academic Writing In An Interdisciplinary Field*. Unpublished Phd Thesis, University Of Michigan, Ann Arbor.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Teich, E., & Holtz, M. (2009). 'Scientific registers in contact - An exploration of the lexico-grammatical properties of interdisciplinary discourses'. *International Journal of Corpus Linguistics*, 14(4), 524-548.

**Matthew Brook O'Donnell, Ute Römer, and Nick C. Ellis (all University of Michigan)**

Exploring Zipfian Distributions in English Verb Argument Constructions: Corpus and Psycholinguistic Evidence

Faced with novel utterances such as “the ball MANDDOOLS across the ground” and “the teacher SPUGGED the boy the book”, how are we able to arrive at similar interpretations? You know that MANDDOOL is a verb of motion and have some idea of how MANDDOOLING works and similarly that SPUGGING involves some sort of transfer. The larger configuration of words (the construction) has come to carry meaning as a whole (Goldberg, 2003, 2006). MANDDOOL inherits its interpretation from the echoes of the verbs that occupy this Verb Argument Construction (VAC) -- words like ‘come’, ‘walk’, ‘move’, ..., ‘scud’, ‘skitter’ and ‘flit’. Small-scale studies of a handful of constructions in both first- (Goldberg, 2006) and second- (Ellis & Ferreira-Junior, 2009) language acquisition have argued that the Zipfian distributions (Zipf, 1949) of verbs in VACs helps make them learnable. The aim of our research is to explore these factors for VAC acquisition in a larger number of VACs in both a large-scale corpus analysis and through psycholinguistic experiments.

In the corpus analysis, we are building an inventory of specific VACs identified in the COBUILD Grammar Patterns volume (Francis, Hunston, & Manning, 1996) using the BNC. We define searches against part-of-speech categories and dependency relations produced by three different parsers (RASP [Briscoe et al. 2006], Stanford [de Marneffe et al 2006] and C&C [Curran et al. 2007]) and take a consensus of whether a sentence matches the VAC pattern. Then for each VAC, such as V across n, we record the distribution of the verb types, their token frequencies and sentence contexts. We determine the degree to which the distributions are Zipfian (e.g. come 474 ... spread 146 ... throw 17 ... stagger 5). Statistical analyses (MI, Delta-P, Chi-Square) examine the associations between verbs and constructions (e.g. scud, skitter, sprawl, flit have the strongest association with V across n). WordNet and other semantic resources are used to measure the cohesion of the types in each distribution (e.g., semantic fields TRAVEL and MOVE most frequent for V across n). These data allow us to make predictions regarding language users' knowledge of verbs in constructions.

In psycholinguistic experiments we use free association tasks to test these predictions. We have native and non-native speakers of English think of the first word that comes to mind to fill the V slot in a particular VAC frame. The range of the verbs that they generate, and their speed of access, inform us about the representation of these VACs in the human mind. For each VAC, we compare the results from the experiments and the corpus analysis in terms of verb selection preferences. This research demonstrates the productive synergy of corpus linguistic and psycholinguistic methods and findings.

Briscoe, E., Carroll, J., & Watson, R. (2006). 'The Second Release of the RASP System'. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Curran, J.R., Clark, S., & Bos, J. (2007). 'Linguistically Motivated Large-Scale NLP with C&C and Boxer'. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), 33-36.

Ellis, N.C., & Ferreira-Junior, F. (2009). 'Constructions and their acquisition: Islands and the distinctiveness of their occupancy'. *Annual Review of Cognitive Linguistics*, 111-139.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1996). *Grammar Patterns 1: Verbs. The COBUILD Series*. London: Harper Collins.

Goldberg, A. E. (2003). 'Constructions: a new theoretical approach to language'. *Trends in Cognitive Science*, 7, 219-224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

de Marneffe, M., MacCartney, B., & Manning, C.D. (2006). *Generating Typed Dependency Parses from Phrase Structure Parses*. In LREC 2006.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

### **Kieran O'Halloran (Open University, UK)**

Electronic Deconstruction of an argument through its 'supplement': Derrida and corpus linguistic method

A by-product of new social media is an abundant textual record of engagements - billions of words across the world-wide-web in, for example, discussion forums, blogs and wiki discussion tabs. Many such engagements consist of commentary on a particular text and can thus be regarded as electronic supplements to these texts. The purpose of this presentation is to flag the utility value of this electronic supplementarity for corpus-based, critical reading by highlighting the following: how an electronic supplement can reveal particular meanings that the text being responded to can reasonably be said to marginalize and / or repress. In turn, this can show where the text's rhetorical structure can be said to be unstable, in a state of deconstruction. Given the often large size of these supplements, knowing how to mine them with corpus linguistic software is essential. I refer to this new type of corpus-based analysis as Electronic Deconstruction.

Electronic Deconstruction takes part of its theoretical stimulus from the philosopher, Jacques Derrida, and, in particular, his idea of the supplement. We normally understand a supplement as something which is an add-on and thus outside that which is being supplemented. In contrast, for Derrida (1976), any supplement has an undecideable 'inside-outside' relation, e.g., vitamin supplements are both outside the diet in providing additional vitamins and inside the diet in replacing a lack of vitamins.

I report on recent, Derrida-inspired research (O'Halloran, 2010) where I apply the logic of the supplement to an online discussion forum appended to an argument. By employing statistical keyword analysis of this discussion forum supplement via WMatrix software (Rayson, 2008), using the BNC Sampler written corpus as a reference corpus, I reveal that keywords in the forum absent from the argument can be perspectivised as 'lacking' in it. Furthermore, since keywords are generated non-arbitrarily, we have in turn a non-arbitrary basis for intervening in the argument; we can use these keywords to 'add to replace' what can be perspectivised as deficiency in normal discussion of the argument's topic, intervening to replace an absence *inside* the argument with keywords *outside* the argument. Via the logic of the supplement, the border between an argument and its discussion forum supplement is porous.

The next stage is to trace the extent to which this intervention in the argument 'to add to replace' leads to instability in its cohesion. An argument's rhetorical structure is dependent on effective cohesion. If cohesion is disturbed by this intervention, then the rhetorical structure of the argument

is unstable. If the rhetorical structure deconstructs in this way, this can offer insights into repression or marginalisation in the argument *relative* to the particular supplement.

Derrida, J. (1976[1967]). *Of Grammatology* [trans G.C. Spivak], Baltimore: Johns Hopkins University Press.

O'Halloran, K.A. (2010). 'Critical reading of a text through its electronic supplement', *Digital Culture and Education*, 2(2): 210-229.

[http://www.digitalcultureandeducation.com/cms/wp-content/uploads/2011/06/DCE1022\\_ohalloran\\_2010.pdf](http://www.digitalcultureandeducation.com/cms/wp-content/uploads/2011/06/DCE1022_ohalloran_2010.pdf)

Rayson, P. (2008). *Wmatrix: a Web-based Corpus Processing Environment*. Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix>.

### **Magali Paquot (Centre for English Corpus Linguistics, Université Catholique de Louvain)**

#### **Dictionary-cum-corpus: A step towards more customisation in pedagogical lexicography**

Corpus use has become a standard practice in lexicography, most particularly English lexicography. Lexicographers make use of corpus data to select the words they include in a dictionary, describe their meaning and use and illustrate their preferred environment in context (Atkins and Rundell, 2008). With the advent of electronic dictionaries, corpus data is making its way into the dictionary via new components such as example banks or corpus-query systems. In the Longman Dictionary of Contemporary English (5th edition, DVD-ROM, 2009), a sample of 1 million sentences taken from the Longman Corpus Network is used to provide additional examples (via the Example Bank frame of the right menu) and collocations (via the Collocations frame of the right menu) for each lexical entry. In the Collins COBUILD Advanced Learners' English Dictionary on CD-ROM (2006), users can search a five million sample from the Bank of English Corpus via the WordBank tool.

In my presentation I will first investigate the role of corpus-query-tools in pedagogical lexicography. I will focus on electronic learner dictionaries and online tools such as the Base Lexicale du Français (Verlinde et al, 2009). I will argue that, to be useful (and used!), a corpus-query-tool needs to be fully integrated into the dictionary: it has to be available from each lexical entry and point straight at concordance lines for the relevant item. The corpora should also be user-oriented so as to allow learners to visualise senses in a context close to their own working environment (Granger & Paquot, 2010a and in preparation). By way of illustration, I will then introduce the Louvain English for Academic purposes Dictionary (LEAD) (Granger & Paquot, 2010b), a dictionary which contains a rich description of academic words (Paquot, 2010), with particular focus on their phraseology (collocations and recurrent phrases). It is a web-based integrated tool where the actual dictionary is linked up to an open-source corpus-query system, viz. CQPweb (Hardie, 2009). The LEAD innovates by automatically adapting the content to users' needs in terms of discipline and mother tongue background and by giving access to discipline-specific corpora rather than generic corpora. With the example of the LEAD dictionary, I will also argue that the future of specialised pedagogical lexicography lies in more customisation.

Atkins, S. & M. Rundell (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Granger, S. & M. Paquot (2010a) 'Customising a general EAP dictionary to meet learner needs'. In Granger, S. & M. Paquot (eds) (2010) *eLexicography in the 21st century: New challenges, new applications*. Proceedings of ELEX2009. Cahiers du CENTAL. Louvain-la-Neuve, Presses universitaires de Louvain, 87-96.

Granger, S. & M. Paquot (2010b) 'The Louvain EAP Dictionary (LEAD)'. In Proceedings of the XIV EURALEX International Congress, Leeuwarden, The Netherlands, 6-10 July 2010, 321-326.

Granger, S. & M. Paquot (in preparation) 'Automated customization or how to bring electronic dictionaries closer to their users'.

Hardie, A. (2009). 'CQPweb – combining power, flexibility and usability in a corpus analysis tool'. Paper presented at the 30th ICAME conference, Lancaster, 27-31 May 2009.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New-York: Continuum.

Verlinde, S., Leroyer, P. & J. Binon (2009). 'Search and you will find. From stand-alone lexicographic tools to user driven task and problem-oriented multifunctional leximats'. *International Journal of Lexicography* 23(1): 1-17.

### **Laura Louise Paterson (Loughborough University)**

#### A methodological fusion: Problems combining CDA with corpus linguistics

Analyses using corpus linguistics tools, such as concordance analysis, collocation clouds, and frequency counts, tend not to focus on wider social factors involved in text construction. In this paper I consider how models of discourse analysis, specifically the model of Critical Discourse Analysis proposed by Fairclough (1995), can be combined with corpus linguistics in order to do both textual and social analyses of texts. Arguably, the main advantage of using corpus linguistics with CDA is that the resulting method will combine quantitative data (usually associated with corpus linguistics) with qualitative data (usually linked to discourse analysis). Indeed, Mautner describes the combination of these two methods of analysis as "a 'best-of-both-worlds' scenario hardly achievable through the use of purely qualitative... analysis" (2009:125).

However, this combination of methods is problematic. In this paper I use Fairclough's tri-level framework to show that, if one begins from a CDA perspective and attempts to map corpus tools onto the already existing model, the resulting analysis becomes highly complex and specialised to the text type(s) analysed. Using an example involving the analysis of third-person pronouns, I illustrate how each level in Fairclough's model (textual, discursive, and sociocultural) must be tailored to the specific data in the corpus, meaning that what was originally a relatively neat and highly structured model of CDA becomes unsuitable as a general model of analysis.

On the contrary, I argue that if one begins from a corpus linguistics perspective and uses the model of analysis proposed by Mautner (2007) then it is possible to combine both corpus and discourse analysis into a generic model which can be applied across text types. Mautner's method combines both quantitative and qualitative data in order to show how corpus analysis can be well informed by a sharper focus on discourse. Thus, having presented the two models of combined corpus and discourse analysis, I conclude that movement back and forwards between quantitative (corpus) and qualitative (corpus and discourse) analysis, as proposed by Mautner, and Harwood (2006), is an effective way of tackling discourse using tools primarily associated with corpus linguistics.

Fairclough, Norman. 1995. *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.

Harwood, Nigel. 2006. '(In)appropriate personal pronoun use in political science: A qualitative study and a proposed heuristic for future research.' *Written Communication* 23: 424-450.

Mautner, Gerlinde. 2007. 'Mining large corpora for social information: The case of elderly.' *Language in Society* 36(1): 51-72.

**Matthew Peacock (City University of Hong Kong)**

Discipline Variation in High-Frequency Nouns in Research Articles across Eight Disciplines

This paper describes a corpus-based analysis of the form, function, and distribution of high-frequency nouns in academic discourse across eight disciplines: Chemistry, Computer Science, Materials Science, Neuroscience, Economics, Language and Linguistics, Management, and Psychology. It is proposed that high-frequency nouns are an important part of academic discourse including research articles, and worth investigating further, yet little previous research seems to have investigated these areas.

The aims of this research were, within our corpus of 320 research articles, to find and list the most common nouns, investigate frequency and disciplinary variation, and investigate noun function using Francis et al's. functional categories (1998). However, careful and extensive examination of the 16 highest-frequency nouns found that they did not fit these categories, perhaps because they were created for the 'noun + that' pattern rather than nouns. We then devised three new provisional data-driven semantic categories to help clarify our results: 1. Reference to Research, reference to the whole study, or to other research. 2. Affecting Entity, abstract entities which affect something in a study. 3. Research Device, abstract devices used by researchers as part of their research design. We preserved one of Francis et al's. categories, 4. Evidence, signs or evidence that something is the case. This categorization, along with all results, was checked by two evaluators who independently measured inter- and intra-rater agreement. Results for individual nouns are also presented.

Considerable disciplinary variation was found, with many statistically significant differences. For example, 'Reference to Research' was significantly higher in Computer Science and Management, 'Affecting Entity' in Economics and Management, 'Research Device' in Computer Science and Psychology, and 'Evidence' in Computer Science and Economics. Further analysis revealed that Chemistry and Materials Science authors tend to minimize their personal involvement in their findings and present their research in a distinctly narrative and descriptive style. Authors in the other disciplines adopted a different style, relying more on personal presentation and persuasion than on presenting hard facts and letting those facts and data speak for themselves: for example, it appears to be particularly important for Computer Science and Economics authors to explicitly discuss evidence which supports, and explains the meaning of, their results, claims, and arguments.

Conclusions are that abstract noun usage in RAs is discipline specific and that the patterns revealed are disciplinary norms, accepted within disciplines as recognized ways for writers to present their research. This research adds to understanding of discipline variations in the form, function and frequency of high-frequency nouns and also tells us more about academic discourse, how knowledge is constructed, what knowledge is, and research practices across a range of disciplines.

Francis, G., S. Hunston and E. Manning. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

**Gill Philip (University of Macerata)**

Key words, pivotal words and leading words in "Harry Potter"

In one of his less familiar papers, Firth (1956) set out very clear guidelines for the study of collocations in restricted languages. Collocation is "the study of key-words, pivotal words, leading words, by presenting them in the company they usually keep" (ibid: 106), and best investigated in restricted languages rather than the general language. Although restricted languages now tend to be interpreted as being particular genres or types of discourse, Firth specifically included within his definition the works of an individual author, and single books. This paper investigates the

collocations of key words in the "Harry Potter" epic to illustrate how shifts in collocational patternings turn key words into pivotal words, and how these pivotal words and their collocations become leading words in a narrative thread.

This corpus-driven study offers a novel view of keyness in text. Rather than look for words which are key for the text as a whole ("absolute keyness"), the emphasis here is on keyness as a dynamic phenomenon. As the text proceeds, layers of text patterns accumulate, consolidating, elaborating and modifying the patterns that have gone before. The key words in any single chapter ("local keyness") are the ones which reveal how the narrative is proceeding.

It is the collocations of key words which build up and sustain a narrative. This paper will show how major shifts in the collocational patternings signal twists in the tale, the new patternings forming around "Harry", "Dumbledore" and "Snape" generating uncertainty and suspense for the reader.

Firth, J.R. 1956. 'Descriptive linguistics and the study of English'. In F.R: Palmer (ed.) 1968. *Selected Papers of J.R. Firth 1952-59*, 96-113. London and Harlow: Longmans.

**Adam Przepiórkowski (Institute of Computer Science, PAS), Rafał L. Górski (Institute of Polish Language, PAS), Marek Łaziński (University of Warsaw), and Piotr Pęzik (University of Łódź)**

The National Corpus of Polish – benefits of synergy

In this paper we report more than three years of work on the National Corpus of Polish. What makes the National Corpus of Polish project different from a typical YACP (Yet Another Corpus Project)? Before the project started there were four corpora of Polish available. Each had some merits but none of them met all the requirements (large, balanced, annotated, publicly available) of a modern general reference corpus.

The NCP is an effect of a joint effort of four teams which constructed the mentioned corpora: the Institute of Computer Science PAS, Institute of Polish Language PAS, Chair of English (University of Lodz) and Polish Scientific Publishers PWN. Thus – contrary to most national corpora – the work did not start from scratch. Each of the teams brought expertise, but also their resources and tools. The project however went far beyond simply merging the four corpora.

The result of the project is a large corpus of over 1.5 billion tokens, a 300 million balanced subcorpus, and a 1 million manually annotated subcorpus.

In the course of the project a number of tools have been developed, including Anotatornia (an on-line tool for manual annotation of texts), two search engines, a morphosyntactic tagger, tools for word sense disambiguation, named entity recognition and shallow syntactic parsing. Last but not least: a ready solution of XML annotation (based on TEI P5) which is well documented. All these tools are publicly available and can be (or in case of Poliqarp are already) reused.

Still the corpus was compiled not only with a view to Natural Language Processing, but also to purely linguistic applications. Every year a demo version of the corpus was launched, allowing for feedback from the users. This feedback was quite intensive because the corpus, also at the stage of compiling, was an empirical basis of a large dictionary of modern Polish. The corpus was also used for educational purposes, including training translators.

Due to the considerable size of its unbalanced part, the corpus is also treated as a repository of texts which can be rearranged to form a new corpus. Two examples are: a corpus of Polish of the 21st century as a counterpart of a corpus of the sixties for tracking short-term diachronic processes, and a corpus of Polish directly comparable to BNC.

The project has proven that merging various corpora compiled for one language is worth the effort. The effect is much more than simply a sum of the scattered corpora, but provides a corpus of an entirely new quality.

Acedański, S. (2010). 'A morphosyntactic Brill tagger with lexical rules for inflectional languages'. In *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland, Lecture Notes in Artificial Intelligence*, Berlin. Springer-Verlag.

Górski, R.L. (2008): 'Representativeness of a written part of a Polish general-reference corpus. Primary notes'. [w:] Barbara Lewandowska-Tomaszczyk (red.): *Corpus Linguistics, Computer Tools, and Applications - State of the Art*. PALC 2007, Frankfurt/M etc: Peter Lang.

Przepiórkowski, A. and Murzynowski, G. (to appear). 'Manual annotation of the National Corpus of Polish with Anotatornia'.

Pęzik, P. (to appear). 'Providing corpus feedback for translators with the PELCRA search engine for NKJP'.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). 'Towards the annotation of named entities in the National Corpus of Polish'. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta*. ELRA.

#### **Hajar Abdul Rahim (Universiti Sains Malaysia)**

Using Frame Semantics to analyse semantic shift, expansion and divergence in NNS English: the case of 'best' and 'terror' in Malaysian English

Lexical creativity in New Englishes, the literature shows, pushes linguistic borders and in some instances gives rise to debates on whether a form is really creative or merely divergent. This paper reports on a study of the use of the words 'best' and 'terror' in spoken Malaysian English, which in many instances, are rather different from the way they are used in native as well as other varieties of English. A corpus-based method will be employed in the study to generate the data that is needed to analyse the use of the two words in spoken Malaysian English. The spoken sub-corpora of *ICE Malaysia* is the source corpus. The way in which the words are used in the Malaysian context, and the meanings they acquire due to this use will be analysed using Frame Semantics. Frame Semantics provides a semantic and syntactic combinatorial framework that helps to show the system of concepts or frame(s) of the words. To discover the extent to which the two words shift, expand or are divergent in meaning in comparison to native use, the outcome of the analysis will be compared with the information on the two words in FrameNet, an online lexical resource for contemporary English that is based on the principles of Frame Semantics. It is hoped that the findings will provide some new understanding of creativity and linguistic acceptability particularly in the context of New Englishes and other varieties of NNS English.

#### **Esperanza Rama-Martínez (University of Vigo, Spain)**

On defendants (re)initiating talk in criminal trials: The Old Bailey in the Late Modern English period

Defendants in English courtroom interaction have not always had their right to a speaking turn restricted by the lawyers' questioning task. By contrast to present-day courtroom interaction (Atkinson & Drew, 1979; Drew, 1992; Heritage & Clayman, 2010), defendants during the mid-17th to mid-18th century had a prominent role as questioners, addressing questions to witnesses and requests to judicial examiners and lawyers (Archer 2005). Though defendants' active participation seemed to decrease as the EmodE period progressed, Archer attested an increase in defendants' use of (re)initiating questions over the three periods she analysed (1640-1679, 1680-1719 and 1720-

1760).

In an attempt to explore the evolution of defendants' talk-(re)initiating activity from 1760 onwards, a preliminary analysis of the proceedings of the Old Bailey Corpus corresponding to the period 1760-1860 (Rama-Martínez & Martínez-Ínsua, 2010) revealed a considerable decrease in the frequency of turn transitions, including defendants' turns. It was observed that defendants overwhelmingly resorted to the self-selection technique though with seemingly different intentions in each century: to state their cases in a self-defence turn at the end of the trial during the 18th century, but to cross-examine witnesses with the aim of disproving evidence during the first half of the 19th century. Consequently, the types of moves and illocutionary forces varied notably between the two centuries. Notwithstanding the considerable decrease in defendants' contributions to interaction over the two centuries, the 19th century witnessed an increase in defendants' (re)initiating talk throughout the trial. The aim of this paper is to investigate to what extent this trend of decrease in defendants' turns but increase in defendants' questioning talk, initiated in EmodE and maintained until 1780, continues over the latter part of the LmodE period. In line with the achievement of full defence by counsels through the Prisoners' Counsel Act in 1836 (Cairns, 1998), this tendency of a gradual decrease in the production of defendants' (re)initiating moves is expected to hold for the second half of the 19th century, thus establishing a smooth transition into today's courtroom system. For this purpose, I shall adopt both a qualitative and quantitative approach to the study of a random selection of the Old Bailey criminal proceedings corresponding to the period 1860-1899. The analysis will focus on the following parameters: (i) types of turn-allocation techniques used to guarantee defendants' access to a turn at talk; (ii) types of defendants' (re)initiating moves defined in terms of their illocutionary force (eg. question, request, require); (iii) the addressees of those defendants' moves (eg. judge, lawyer, prosecutor, witness) so as to assess the extent to which the (a)symmetrical roles of the addressees imprint on the way in which defendants take (re)initiating moves; and (iv) the frequency of interlocutor (non-)response to defendants' turns.

Archer, Dawn (2005). *Questions and Answers in the English Courtroom (1640-1760)*. Amsterdam & Philadelphia: John Benjamins.

Atkinson, J. Maxwell & Paul Drew (1979). *Order in Court*. London: The Macmillan Press Limited.

Cairns, David J.A. (1998). *Advocacy and the Making of the Adversarial Criminal Trial 1800-1865*. Oxford: Clarendon Press.

Drew, Paul (1992). 'Contested evidence in courtroom cross-examination: the case of a trial for rape'. In Paul Drew & John Heritage (eds.) *Talk at Work*. Cambridge: Cambridge University Press. 163-198.

Heritage, John & Steven Clayman (2010). *Talk in Action. Interaction, Identities, and Institutions*. Malden: Wiley-Blackwell.

Rama-Martínez, Esperanza & Ana-Elina Martínez-Ínsua (2010). 'Courtroom interaction between 1760 and 1860: on defendants taking (re)initiating moves'. 31 ICAME Conference. Giessen: University of Giessen.

**Ana Rita Remígio (Universidade de Aveiro)**

Corpus-based terminography: popularizing discourse and the role of communicative contexts in specialised corpora building

The lack of, or reduced, understanding of the consumer, as far as food items with health claims that are available on the market are concerned has led to the proposal of a terminological database targeted to non-experts, with the aim of making scientifically valid and accurate information on the

so called functional food systematically available. The resource was conceived, built and populated within the framework of a corpus-based approach for terminography, i.e., the applied field of Terminology, which besides the textual, also comprises the conceptual and communicative dimensions.

The singular nature of this project lies in the fact that texts on functional food targeted to the consumer are not only written by researchers, professors or science communicators, but also by actors from the food industry and journalists, who produce texts according to given communicative intentions: if the target public is heterogeneous, so are the text producers. Consequently and upon familiarization with the special subject field, the following question was raised: how to select and organise texts on functional foods produced by different actors, in the corpus, in order to be able to identify and extract term candidates and contexts rich in conceptual information, as a basis for definition writing, to populate the resource with information that suits the needs of the consumer?

The answer to that question was based on the hypothesis that the identification of communicative contexts, i.e the circumstances of discourse production, considering text producers, communicative intention and target public, in which popularising discourse on functional food is produced could be intimately related to text selection criteria and the design of the specialised corpus.

Therefore, texts from different genres were selected, with the aim of obtaining a representative sample of popularizing discourse on functional foods, and organized within the corpus according to the communicative context in which they were produced. Those texts were then separately analysed, using WordSmith Tools, so that terminological information identified and extracted could then be compared: quantitatively and according to relevance. This comparison aimed to evaluate the adequacy of inclusion of texts produced in three different communicative contexts in the corpus, of corpus design and, as a consequence, to assess the need to redesign the corpus, in order to obtain more and better results. A reference corpus of texts targeted at subject field specialists, and thus representative of scientific discourse on the field, was also built, again for comparison purposes and to verify if information not present in the corpus of study and relevant for the resource was identifiable.

Results of this study are going to be disclosed and concluding remarks on the process of designing a specialised corpus, representative of popularizing discourse according to communicative contexts for terminographical purposes, will be discussed.

BIBER, Douglas – ‘Representativeness in corpus design’. *Literary and Linguistic Computing*. ISSN 1477-4615. 8:4 (1993). p. 243-257.

COSTA, Rute; SILVA, Raquel – ‘De la typologie à l’ontologie des textes’. In *Terminologies & Ontologies: théories et applications*. Actes de la 2ème Conférence – Toth Annecy. Annecy : Institut Porphyre, 2008. Savoie et Connaissance.

COSTA, Rute – ‘Terminology, corpus linguistics and ontologies’. In *Contrastive studies and valency: studies in honour of Hans Ulrich Boas*. Frankfurt: Peter Lang. 2006. p. 107-118. ISBN 3-631-54935-0.

JACOBI, Daniel – *La communication scientifique: discours, figures, modèles*. Saint-Martin-d’Hères: PUG, 1999. 277p. ISBN 2-7061-0822-3 (Communication, médias et sociétés).

JACOBI, Daniel – ‘Sémiotique du discours de vulgarisation scientifique’. Semen, De Saussure aux média [Em linha] 02 (1985). [Consultado a 17/06/08] Disponível na WWW:

<http://semen.revues.org/document4291.html>.

L'HOMME, Marie-Claude – *La terminologie: principes et techniques*. Québec: PUM, 2004. 202 p. ISBN 2-7606-1949-4. (Paramètres).

MEYER, Ingrid – 'Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework'. In BOURIGAULT, Didier; JACQUEMIN, Christian; L'HOMME, Maria-Claude, eds. – *Recent advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001. p. 279-302. (Coll. NLP) ISBN 1588110168.

MEYER, Ingrid; MACKINTOSH, Kristen – 'The corpus from a terminographer's viewpoint'. *International Journal of Corpus Linguistics*. ISSN: 1384-6655 1:2 (1996), 257-285.

PEARSON, Jennifer – *Terms in context*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998. 242p. ISBN 1 55619 342 4.

REMÍGIO, Ana Rita – *Processo terminográfico: vertentes conceptual, comunicativa e textual. Proposta de uma base de dados terminológica para o consumidor*. Aveiro: Universidade de Aveiro, 2010. 606 p. Tese de Doutoramento.

REMÍGIO, Ana Rita – *Do processo tradutivo à actividade terminológica: contributo para uma analogia de abordagens*. 3Ts – Revista de Tradução, Terminologia e Tecnologias. ISSN 1646-849X n.º 0 (2008) p. 37-51.

REMÍGIO, Ana Rita; ROBERTO, Maria Teresa; COSTA, Rute – A divulgação científica no sector alimentar: os contextos comunicativos e a informação terminológica veiculada. In XI Simpósio Ibero-Americano de Terminologia (RITerm 2008): "A terminologia no terceiro milénio: pela adopção de boas práticas terminológicas" [Em linha] Lima, 2008. Disponível na WWW: [http://www.riterm.net/actes/11simposio/Remigio\\_Ana-Roberto\\_MT.htm](http://www.riterm.net/actes/11simposio/Remigio_Ana-Roberto_MT.htm)

REMÍGIO, Ana Rita; ROBERTO, Teresa; COSTA, Rute – 'Organização conceptual e diversidade de contextos de comunicação na construção de um corpus: o caso das Ciências da Nutrição'. In *Jornada REALITER sobre Metodologia para a Recolha e Sistematização de Corpora para fins Dicionarísticos*. [Em linha] Rio de Janeiro, 2006. Disponível na WWW: <http://www.realiter.net/spip.php?article552>.

SCOTT, Mike – *Oxford WordSmith Tools version 4.0*. [Em linha] Oxford: Oxford University Press, 2006a. 259 p. [Consultado a 11/10/07] Disponível na WWW: <http://www.lexically.net/wordsmith/>.

SCOTT, Mike – 'Word-lists: approaching texts'. In SCOTT, Mike; TRIBBLE, Christopher – *Textual patterns: key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2006b. p. 11-32. ISBN 90 272 2294 0.

SINCLAIR, John – 'Introduction'. In HOEY, Michael [et al.] – *Text, discourse and corpora: theory and analysis*. London/New York: Continuum, 2007. p. 1-5. ISBN 978-08264-9171-8. (Studies in Corpus and Discourse).

SINCLAIR, John – 'Current issues in Corpus Linguistics'. In SINCLAIR, John – *Trust the text: language corpus and discourse*. London: Routledge, 2004. p. 185-193 ISBN-10 0415317681.

**Irene Renau and Arceli Alonso (Institut Universitari de Lingüística Aplicada, Universitat Pompeu**

**Fabra, Barcelona)**

Using Corpus Pattern Analysis for the Spanish Learner's Dictionary DAELE (Diccionario de aprendizaje del español como lengua extranjera)

Learner's dictionaries for non-native speakers are based on the assumption that users look up words in the dictionary not only for decoding texts, but also for encoding. In this sense, more grammatical information is included as the main target users, who do not have full command of the grammar of the language, need information about the use of words in a specific context. Corpus linguistics has brought a light for the building-up of dictionaries in the English lexicographical tradition, and specifically, in the elaboration of learner's dictionaries. Spanish lexicographical tradition, on the contrary, has dedicated itself mainly to elaborating monolingual dictionaries for native speakers, and as a result, currently there are neither many learners' dictionaries nor many Spanish learners' dictionaries which take into account the use of words in texts.

Corpus-driven methodologies have been applied for English learners' dictionaries for the selection of headwords, for the organisation of data and for establishing the uses of a given word that are to be included in the dictionary (Atkins & Rundell 2008; Rundell 2002). One of these methodologies has been developed by Patrick Hanks for the creation of a Pattern Dictionary of English Verbs (Hanks 2004, 2005). Corpus Pattern Analysis (CPA) is a corpus-driven method to establish syntactic and semantic verbal patterns. It is based on the Theory of Norms and Exploitations - TNE (Hanks 2004, forthcoming) which establishes that there are norms for a given word which can be exploited to create specific uses in a given context. TNE and CPA establish a theoretical and practical framework for "mapping meaning onto words in text". Although CPA has been developed for English language, it is also being applied to other languages such as English, Italian or Spanish.

In previous studies (Alonso 2009; Alonso & Renau, forthcoming), it has been stated that CPA is not only a useful technique for studying the main common patterns of a word, but also for determining terminological uses of a given word in a specific context. Actually, CPA is being analysed for being applied in the case of the Spanish Learners' Dictionary DAELE (Bernal & Renau 2010; <http://www.iula.upf.edu/rec/daele/>).

In the present study, the use of CPA applied to Spanish language is shown by analysing different Spanish verbs. Several entries of Spanish verbs created by using CPA are compared to the same entries created without using this methodology. As it is well known, corpus analysis implies a thoroughly study of occurrences of a word. If the process of going through the occurrences is not systematized, the lexicographer may make mistakes at the time of interpreting the data. Our main objectives are, in one hand, to illustrate in which way CPA contributes to the systematization of this process and improves the way of writing dictionary entries and, on the other hand, to show the potential of CPA as a unique method for being applied to several languages, specifically, to Romance languages.

Alonso, A. (unpublished, 2009) *Características del léxico del medio ambiente en español y pautas de representación en el diccionario general*. [PhD dissertation]. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Alonso, A. & Renau, I. (forthcoming). "Los verbos españoles y el Corpus Pattern Analysis de Hanks: primera propuesta de adaptación". IV Congreso Internacional de Lexicografía Hispánica. Tarragona (Spain), 20-22 September 2010. Tarragona: Universitat Rovira i Virgili.

Atkins, B. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bernal, E. & Renau, I. (2010). "¿Lo que necesitan es lo que encuentran? Reflexiones a propósito de la representación de los verbos en los diccionarios de aprendizaje de español". In A. Dykstra, Schoonheim, T., (eds.). Proceedings of the XIV Euralex International Congress. Ljowert (Nederlands): Fryske Akademy, pp. 484-496.

Hanks, P. (2004). "The Syntagmatics of Metaphor and Idioms". *International Journal of Lexicography*, Volume 17, Number 3. Oxford: Oxford University Press. 245-274.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hanks, P. & Pustejovsky, J. (2005). "A Pattern Dictionary for Natural Language Processing". *Revue Francaise de linguistique appliquée*, 10:2. 63-82.

Rundell, M. (2002). "Good old fashioned lexicography: Human judgement and the limits of automation". In Corréard, M. H. (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Grenoble, France: EURALEX. 138-155.

**Suhaila Saeed, Ranaivo-Malançon Bali, and Tang Enya Kong (Multimedia University, Malaysia)**

A Construction of Computational Morphology Resources for U-RL

Computational Morphology Resources are resources needed for morphology analysis and generation in Computational Morphology field. In Asia, only a small number of listed languages has started to do research on morphology analysis and generation. In fact, Sarawak languages are Asian languages excluded from the list. To our knowledge, there does not exist any applications in the Natural Language Processing context which involve Sarawak languages. For example, morphology analysis and generation that is known as a first process in text processing area. In addition, Sarawak ethnic languages could be categorized as Under-Resourced Languages due to the lack or no digitized resources yet in the Information and Communications Technology context, in general and traditional technology as well.

Morphological data acquisition is a crucial and hard step in pre-processing stage of automatic morphology analysis and generation due to lack of linguistic resources in terms of lexicon, corpus, and grammar. Therefore, in this paper, we highlighted our main problem which is no resources at all in term of unavailability of internal structure when word is analysed in the case of Under-Resourced Languages. The issue related to this is how to get the required resources? There are two types of most required resources in automatic morphological system which are i) corpus and ii) list of stems and affixes. Both resources play an important role to the next steps in morphology analysis and generation, indeed.

On top of that, two research questions have been encountered from the mentioned problem and there are: i) What is the best work flow for corpus acquisition by considering the Under-Resourced Languages issues? and ii) How many data sets are needed to acquire morphology information in morphology induction for Under-Resourced Languages?

In this paper, a work flow for corpus acquisition in the context of Under-Resourced Languages has been proposed. The workflow consists of three main stages which are: i) Stage 1: data collection [three types of sources are dictionaries, grammar book(s), written text], ii) Stage 2: text formation [two possible processes would involve either digitisation (mainly for hardcopy version of sources) or conversion (mainly for softcopy version of sources)] and iii) Stage 3: compilation [compiling three sources that are in the text format into one text file to produce unannotated corpus]. In fact, the three stages are depending on each other in order to construct the corpus. Besides, a result gained

from the morphology induction that would be a list of stems and affixes from a very small quantity of Under-Resourced Languages resources also be a part of the proposed solution in this research.

Furthermore, the contributions from this research would be: i) An unannotated corpus that can be used in other Natural Language Processing applications, mainly for Sarawak languages and ii) Morphology information of Sarawak languages that induced from unsupervised machine learning. Last but not least, the challenges from this research would be discussed in the last section of the paper.

1. Karagol-Ayan, B. (2007). *Resource generation from structured documents for low-density languages*. (Doctoral dissertation, University of Maryland, College Park). Retrieved from <http://www.lib.umd.edu/drum/handle/1903/7580>

2. Feldman, A. (2006). *Portable language technology: A resource-light approach to morpho-syntactic tagging*. (Doctoral dissertation, The Ohio State University). Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.615&rep=rep1&type=pdf>

3. Cucerzan, S., & Yarowsky, D. (2002). Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th conference on Natural language learning - COLING-02*, pp. 1-7. Morristown, NJ, USA: Association for Computational Linguistics. doi: 10.3115/1118853.1118859

4. Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2), 153–198. MIT Press. Retrieved from <http://www.mitpressjournals.org/doi/abs/10.1162/089120101750300490>

**Silke Scheible, Paul Bennett, Martin Durrell, and Richard J. Whitt (all University of Manchester)**

The GerManC project: Creating an annotated historical corpus of German

This paper describes the results of an AHRC/ESRC-funded project on the creation of a corpus of Early Modern German scheduled for completion in September 2011. GerManC aims to be a representative corpus of written German between 1650 and 1800, and includes a total of eight different genres (representing both print-oriented and orally-oriented registers). It is further subdivided into five dialect regions and three 50-year time periods. The corpus consists of sample texts of 2,000 words, equally distributed across the 'genre', 'period', and 'region' categories, yielding a total of nearly a million words. The structure of the corpus resembles the design of the ARCHER corpus for English. Thus, GerManC is not only of interest for historical German linguists, but it also promises to be an important resource for comparative studies of the development of the two languages.

Our paper reports on the challenges encountered in compiling the corpus, which involves identifying suitable texts and digitising texts printed in Fraktur ('black letter', 'Gothic'). It further provides a detailed account of the annotation of the corpus in terms of structural mark-up (TEI) and linguistic mark-up (sentence boundaries, normalised word forms, lemmas, and POS-tags). The presentation will focus on the following major goals of the annotation process: a.) Creating an automatic linguistic annotation pipeline which is especially suited to historical German in this period, and b.) identifying strategies for a speedy manual correction of the errors produced by the automatic tools. Both points are of vital importance for maximising the overall accuracy of the annotations in the corpus.

To address a.), a novel tokenizer and sentence boundary detector have been created which can deal with multi-genre input such as found in our corpus. Furthermore, we will describe experiments carried out on a manually-annotated subcorpus of GerManC (ca. 50,000 tokens), whose aim is to maximise the performance of state-of-the-art POS-taggers such as the TreeTagger and the TnT

Tagger for German. We will report the results of running the taggers on 'raw' vs. 'normalised' data, and compare these findings to the performance of the taggers when re-trained on our gold-standard subcorpus. The results of these experiments are utilised for creating a historical text processing pipeline optimised for historical input, which minimises the amount of manual correction necessary for a gold standard annotation of the corpus. Finally, addressing point b.), we will introduce a novel web-based annotation platform which is currently being developed as part of the project. Its purpose is to facilitate fast and easy manual correction of token-based annotations such as lemmas and POS tags. The platform will be freely available, and we plan to give a short demonstration of its functionalities at the end of our presentation.

**Elena Semino (Department of Linguistics and English Language, Lancaster University)**

Using corpus linguistic methods to compare a simplified version of *Romeo and Juliet* with the original play

Shakespeare's plays are among the canonical literary works that have been simplified in a variety of editions aimed at readers who cannot easily access the original versions, due to their (young) age, lack of expertise and/or insufficient mastery of English as a foreign or second language. Previous studies on simplified or 'graded' readers have been concerned with their role in the teaching of English as a second or foreign language, particularly with respect to the acquisition of vocabulary. In this paper I report the results of a study that combined corpus-based methods with traditional 'manual' analysis in order to arrive at a systematic account of the differences between the original version of *Romeo and Juliet* and a particular simplified version, published in a series entitled *Shakespeare Made Easy* (Nelson Thornes). The series includes twelve parallel-text editions of Shakespeare's plays, aimed primarily at older school-children (e.g. those preparing for GCSE-level tests in the UK). The software tool Wmatrix (<http://ucrel.lancs.ac.uk/wmatrix/>) was used to compare the original text of the play with the simplified version. The key word tool in Wmatrix identified approximately 100 overused items, while the semantic annotation tool identified approximately 10 overused items (in both cases the cut-off point was a log-likelihood value of 6.63, which corresponds to 99% significance). A detailed analysis of both lists of overused items makes it possible to identify a variety of systematic patterns of simplification, including particularly: the modernisation of Early Modern English spellings, morphology, and lexis; a reduction in the use of personification and body-part metonymies; a reduction in the use of particular metaphors. In other words, a comparison between the two versions of the play by means of two of the tools within Wmatrix does not simply reveal predictable differences such as the replacement of 'thou/thee' with 'you', but also helps to identify less predictable differences that can be explained, for example, as strategies to reduce the density of figurative language in the play, and to allow for differences in attitudes and background knowledge between Elizabethan audiences and the main readership of the simplified version (e.g. a reduction in the use of commercial metaphors for women). Overall, the employment of corpus methods helped to test the validity of hypotheses made on the basis of a textual analysis of the two versions of the play, and additionally produced further insights that could not easily have been reached by other means.

**Cinzia Spinzi, Giulia Riccio, and Marco Venuti (University of Naples, Federico II)**

Mapping the stance clusters: a diachronic corpus-based study of the White House Press Briefings

Literature on genre analysis mainly focuses on the description of language use in the different professional and institutional domains (Bhatia 2004). Despite the different directions of the studies on genre (Bhatia 1993; Martin and Christie 1997; Swales 1990), a common orientation may be seen in their tendency to describe homogeneous concepts, such as communicative situation, register and function.

Nevertheless, genre-specific features are subject to changes due to the ongoing processes of internationalisation and globalisation (Candlin and Gotti 2004; Cortese and Duszak 2005; Crystal 1997). In particular, political and institutional communication genres have been experiencing in-

depth transformation in the last few decades, mainly due to evolutions in the media market, fuelled by technological developments and by the economic globalisation (Blumler and Kavanagh 1999).

This paper looks in particular at the stance clusters or, to put it differently, it deals with the speaker-positioning phrases in the discourse structure of White House press briefings. The data come from a monolingual corpus which includes all the Press briefings across three presidencies (January 1993 – October 2010). The addition of XML mark-up to the whole corpus, including information about individual speakers and their role, as well as providing a chronological subdivision within the corpus, allows us to compare different discourse strategies adopted by different speakers in the briefings at different points in time. This leads us to determine the extent of the differences in the patterns found as well as the nature of the variation from one podium to the next one and to establish if these patterns are semantically primed (Hoey 2005).

The perspective of the analysis is phraseological in that, as Hopper argues (1987: 150), linguistic form, often in prefabricated chunks, is shaped by discourse use.

The analysis relies on two pieces of software: Wordsmith Tools (Scott 2007) to retrieve the clusters and Xaira to study their distribution across the years.

What we aim to demonstrate from a methodological point of view is that clusters, which are here categorized according to functional criteria (see Mahlberg 2007), can be revealing for identifying specificity of this particular genre and the changes stance clusters in particular are experiencing from Clinton to Obama.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Bhatia, V. K. (2004). *Worlds of Written Discourse: A Genre-Based View*. London: Continuum International Publishing Group Ltd.

Blumler J.G. and Kavanagh, D. (1999) "The third age of political communication: influences and features". *Political Communication*, 16, (3): 209–230.

Candlin, C and Gotti, M. (2004) (eds). *Intercultural Aspects of Specialized Communication*. Bern: Peter Lang.

Cortese, G. and Duszak, A. (2005) (Eds). *Identity, Community, Discourse*. Bern: Peter Lang. 139-166.

Hoey; M. 2005. *Lexical Priming*. Abingdon: Routledge.

Hopper, P.J. 1987. "Emergent grammar". *Berkeley Linguistics Society* 13:139-157.

Kumar, M. J. (2007). *Managing the President's Message. The White House Communications Operation*. Baltimore: The Johns Hopkins University Press.

Kurtz, H. (1998). *Spin Cycle. Inside the Clinton propaganda machine*. New York: The Free Press.

Mahlberg, M. (2007). "Clusters, key clusters and local textual functions in Dickens in Corpora" in *Corpora* Vol. 2 (1) 1-31.

Partington, A. (2003). *The Linguistics of Political Argumentation: The Spin-doctor and the Wolf-pack*

*at the White House*. London: Routledge.

Scott, M. (2007). *WordSmith Tools. Version 5.0*. Oxford: Oxford University Press.

Sinclair, J. (2005). 'The phrase, the whole phrase, and nothing but the phrase'. Paper presented at the Phraseology 2005 conference: The many faces of Phraseology. Université catholique de Louvain, 13-15 October.

Swales, J. (1990). *Genre analysis*. Cambridge: Cambridge University Press.

**Dina Strong (University of Latvia)**

Representation of 'Other' in Erasmus Exchange Students' Discourse

In the course of the stay abroad, Erasmus exchange students are confronted and have to find a way to deal with their new environment as well as the many "strangers" they meet. Previous studies argued that while abroad Erasmus students tend to withdraw from the local community, forming "a cocoon" (Papatsiba, 2006; Dervin, 2007) exclusive to exchange students, which remains closed to the local community throughout the time spent in the host country. Thus, the exchange students' marginal status (Dervin & Dirba, 2006) has been held accountable for the frequent proliferation of negative representations that appear in their discourses on the locals. However, what the previous studies of student mobility seem to have overlooked are the personal histories of Erasmus exchange students and the ambivalent nature of the representations of Other (i.e., not only the locals but also their compatriots and other exchange students), which is characteristic of their discourses. By drawing on Critical Discourse Analysis (i.e., van Leeuwen's (1996) theory of Social Actor Representation and Reisigl & Wodak's (2001) conceptualisation of discursive strategies for self and other presentation) and Corpus Linguistics (Stubbs, 1996; Baker, 2006), our paper offers insight into the affect of previous experience of extensive travel and living abroad on the representations of Other produced in discourses of Erasmus exchange students. For the purposes of this study, two small corpora obtained from the semi-structured interviews with two "groups" of Erasmus exchange students: the inexperienced travellers (INTs) and the experienced travellers (EXTs) were recorded and fully transcribed. Contrary to our expectations, the findings point out that the main difference between the EXTs and the INTs was the stance taken towards the statements (assertive/ hesitant; distant/ involved; expert/novice) they made and their stance towards the Other (positive/negative) present in their discourse. EXTs emphasised their advantageous status as experienced travellers and asserted the "expert stance", as they strived for generalisations and the use of intensification strategies when drawing "portraits" (Papatsiba, 2006) of national groups they were familiar with, which served as a vantage point for the comparison. In contrast, INTs avoided making comparisons between the national groups, though sometimes resorted to nationality as heavily mitigated explanation for the behavioural patterns of Other. Moreover, INTs tended to avoid the proliferation of "national portraits" by demonstrating their awareness of heterogeneity of national communities. Despite the sound theoretical and methodological foundation for this study, we realize that discourse produced in the course of the interviews is never impartial and that it is inevitable that the situational context of the interview and the national and institutional affiliations of the interviewer had an effect on the outcome of the interview and have to be taken into consideration.

**Caroline Tagg and Oliver Mason (University of Birmingham)**

Orthographic creativity in Twitter: tweeting about the World Cup 2010

This paper looks at the extent to which corpus methods can be used to explore Internet interactions on the social network site (SNS), Twitter.

Twitter users post messages or 'Tweets' which are restricted to 140 characters and which form a noisy and disjointed online space. It is also fast-paced, with 95 million Tweets posted each day (according to information on the Twitter website). Users have responded to this seemingly

incoherent discourse through their appropriation of various online facilities, whereby a Tweeter's posts can be commented on by their 'followers', and links to external websites can be posted and discussed. Although much has now been written on the use of '@' to address a particular interlocutor, the hash key '#' to indicate topic and 'RT' to show that a message has been forwarded or 'Retweeted' (e.g. Honeycutt and Herring 2009), little has yet been documented on how language is used to indicate stance or how Tweeters mark their engagement with each other linguistically. Assuming that communicative constraints such as those which characterise Twitter often encourage creativity in the fulfilling of interpersonal and evaluative functions, we would expect to see creative, playful language use in Twitter.

In this presentation, we report on an initial study into orthographic creativity, that is, the respelling of words, often to capture spoken pronunciation such as with <wanna> or to create eye-catching forms such as <ur> for 'your' (c.f. Sebba, 2007). The study draws on a corpus of 1902 Tweets collected in July 2010 on the subject of the World Cup. Respellings are identified and grouped together using a list of wordforms ordered by frequency, and are then explored using concordances. It emerges that respelling in Twitter does not serve to abbreviate or shorten a posting, as might be expected due to the similar constraints and functions of SMS text messaging where such practices have been observed (Tagg, 2009). Instead, respellings often lengthen words (as in <ghanaaaaa>) and reflect local pronunciations (such as <dis> for 'this'), and thus appear to contribute to personal expression and the formation of community on Twitter.

Whilst giving insights into the discourse of Twitter, this study raises questions regarding the extent to which Tweets sent via Twitter can be regarded as forming a coherent discourse; how linguists can find a way into the vast, rapidly changing 'Twittersphere'; and the role which corpus methods can play in this.

Honeycutt, C. & Herring, S.C. (2009) 'Beyond Microblogging: Conversation and Collaboration via Twitter'. Proceedings of the Forty-second Hawai'i International Conference on System Sciences (HICSS-42).

Sebba, M. (2007) *Spelling and Society: the culture and politics of orthography around the world*. Cambridge: Cambridge University Press.

Tagg, C. (2009) *A Corpus Linguistics Analysis of SMS Text Messaging*. Unpublished PhD thesis, University of Birmingham.

#### **Charlotte Taylor (University of Portsmouth)**

##### **Searching for similarity: The representation of boy/s and girl/s in the UK press in 1993, 2005, 2010**

This aim of this paper is to raise the methodological importance of searching for similarity and stasis as well as difference and change in corpus-assisted discourse analysis. To date, most research in corpus approaches to discourse studies has largely focussed on difference, which carries the risk of neglecting the similarities (as noted in Baker 2006: 182) and leading the researcher to finding the changes or differences that s/he set out to look for.

This tendency to focus on difference in comparative work is frequently reflected in the corpus linguistic software and in the paper I look at what automated searches for similarity are possible and how these may be employed, for example Sketch Engine's Sketch Difference also includes collocates that are shared between two lexical items.

One notion which has been developed to cover the gap is consistent collocates (c-collocates), introduced in the Lancaster RASIM project to describe the lexical items which collocated with

refugees/asylumseekers/immigrants/migrants in at least seven out of the ten annual subcorpora (described in Gabrielatos and Baker 2008). Baker (2009) further addressed the issue by introducing the concept of lockwords ie items that have a similar frequency across corpora as a parallel to keywords, that is those items which are significantly different between corpora. Another source for approaches to similarity comes from the field of forensic linguistics, which has focussed on similarity far more extensively, particularly in the area of plagiarism studies and detection.

The paper forms part of a larger project which uses the methodology of Modern Diachronic Corpus Assisted Discourse studies (MD-CADS, see the special issue of *Corpora*, 2010) to analyse the representation of GIRL and BOY in the UK press and investigates whether the definitions, characterisations and contexts of presentation have changed or remained stable over the last two decades. In doing so, it builds in particular on previous corpus studies of gender terms such as Holmes & Sigley (2000), Sigley & Holmes (2002), Pearce (2008), Caldas-Couthard & Moon (2010) and Baker (2010). More specifically, I analyse the occurrences of GIRL and BOY in the entire output of the Times, Guardian and Telegraph from 1993 and 2005, and then extend that analysis to a smaller corpus of the same newspapers from 2010. In this presentation, I will be limiting discussion to the analysis of the constants and similarities both in the time period, for instance the continued usage of GIRL to refer to women, and between the search items, for instance the way in which BOY is increasingly used with GIRL collocates.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. (2009) 'The BE06 Corpus of British English and recent language change.' *International Journal of Corpus Linguistics*. 14:3 312-337.

Baker, P. (2010) 'Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English.' *Gender and Language* 4.1: 125-129.

Caldas-Coulthard, R. & R. Moon. "'curvy, hunky, kinky": Using corpora as tools for critical analysis'. *Discourse & Society* 21(2): 99-133.

Gabrielatos, C. & P. Baker (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 pp. 5-38.

Holmes, J. and Sigley, R. (2001) 'What's a Word like *Girl* Doing in a Place Like This?', in A. Smith and P. Peters (eds.), *New Frontiers of Corpus Linguistics*, pp. 247-63. Amsterdam: Rodopi.

Pearce, M. (2008) 'Investigating the Collocational Behaviour of MAN and WOMAN in the BNC using Sketch Engine', *Corpora* 3(1): 1-29.

Sigley, R. and Holmes, J. (2002) 'Looking at *Girls* in Corpora of English', *Journal of English Linguistics* 30(2): 138-57.

#### **Yukio Tono (Tokyo University of Foreign Studies, JAPAN)**

Identifying new verb co-occurrence patterns as criterial features: Using ICCI and JEFLL

This paper aims to present an interim report on the analysis of younger learners' interlanguage development in English using the International Corpus of Crosslinguistic Interlanguage (ICCI). ICCI is a government-funded 5-year project of compiling corpora of English writings by primary and secondary school students in eight different regions (Japan, Spain, Israel, Austria, Poland, Hong Kong, Taiwan, and China), initiated by the author at Tokyo University of Foreign Studies.

The ICCI project focuses on data collection from younger learners of English. Most learner corpora available so far mainly cover intermediate to advanced learners at upper-secondary to university levels. In order to investigate the acquisition processes in its entirety, it is necessary to gather data at learning stages much earlier than that. To this end, we collected beginning-stage learners' data, by administering 20-minute in-class essay tasks without the use of a dictionary to primary and secondary school students. Another unique feature of ICCI is its comparability with the JEFLL Corpus (Tono 2007), a corpus of Japanese EFL learners' writings, covering 10,000 students from Year 7 to 12 (English is introduced in Year 7 in Japan). Together with JEFLL and ICCI, the data consists of more than 17,000 students ranging over 6-8 years of English learning at the beginning stage (the total size of ICCI and JEFLL is well over 1.5 million words).

In this paper, the overall design of the ICCI project will be described and an interim report on the research into the new verb co-occurrence patterns at early stages of learning will be presented as an example of the research using ICCI and JEFLL. The new verb co-occurrence patterns are said to serve as "critical features," linguistic features distinguishing proficiency levels of learners in terms of the Common European Framework of Reference (CEFR). Preliminary findings using Cambridge Learner Corpus (CLC) show that early stages of features, especially A1 and A2 are hard to identify, due to the lack of data in CLC. In this study, we first assigned CEFR levels to sampled compositions (200 samples from each country for each CEFR level, totaling approximately 1,000 samples for A1, A2 and B1 levels respectively). Then all the data were syntactically parsed, and verb co-occurrence patterns were extracted using a *tgrep*-like search engine. The results were statistically analysed in order to find significant differences in frequencies between different CEFR levels. The results showed how ICCI and JEFLL could fill the gap in identifying critical features for beginning-level learners.

Tono, Y. (ed.) (2007) *JEFLL Corpus: A Corpus of 10,000 Japanese EFL Learners*. Tokyo: Shogakukan.

**Aleksandar Trklja (University of Birmingham)**

Distinct(ive) features of translation equivalents

In this paper I will argue that only through a detailed description of lexico-grammatical similarities and differences between translation equivalences identified in a parallel corpus can we understand under what conditions a translation unit 'a' from a source language is translated as 'x', 'y' or 'z' into a target language. Contrastive language researches, which experienced a rebirth with the advent of large parallel corpora, have shown how meaning becomes visible through translation (Johansson 2007). Single words are usually ambiguous out of context and semantic relationships between a translation unit and its equivalents can be studied only if we focus on units of meaning (Teubert 2002, 2004).

The paper is divided into two parts which are parallel to two methodological phases. Part one describes briefly a procedure for extracting translation units and their equivalents in the parallel English-German corpus through the description of collocation profiles. It builds on the method introduced by Danielsson (2003). In the second part, I will show how the concept of distinctive features can be adopted for the study of similarities and differences between translation equivalences by means of reference corpora. The model will be illustrated on the example of the lexical item 'rise'. This item is used in several units of meaning and in my presentation I will focus only on the expression 'give rise to' and its corresponding German translation equivalents.

Danielsson, P. "Automatic Extraction of Meaningful Units from Corpora", in: *International Journal of Corpus Linguistics*. 2003, Volume 8, No 1 pp. 109-127.

Halliday, M.A.K. and Teubert, W. 2004. *Lexicology and corpus linguistics: an introduction*. London,

New York: Continuum.

Johansson, S. 2007. *Seeing through multilingual corpora*. Amsterdam, Philadelphia: John Benjamins.

Teubert, W. "Corpus Linguistics and Lexicography", in: *International Journal of Corpus Linguistics*. 2002, Volume 6, Special Issue, 125-154.

### Fanie Tsiamita (University of Liverpool)

#### Facing the data on *face*: Investigating the claims of Lexical Priming with respect to polysemy

This paper will focus on the claims of Lexical Priming (LP) with respect to polysemy, viz. that the collocations, semantic associations and colligations that a polysemous word is characteristically primed for will systematically differentiate its various senses, and that the different senses will avoid use of each other's primings (Hoey 2005: 81, 82).

The paper will explore these claims on the basis of the verb *face*. An examination of a corpus of fiction texts drawn from the BNC shows that *face* has visual and non-visual uses. The non-visual uses feature abstract *face*, e.g. *face challenges/problems/consequences*, and have a purely confrontational meaning. The visual uses encompass a purely directional sense of *face*, as in *coasts that face north; The two other machines face him*. The data on both the purely confrontational and the purely directional sense appear to support the claims of LP with respect to polysemy. But in the vast majority of cases the visual uses depart from the purely directional sense and involve an additional level of meaning, e.g.:

- (1) nearest and dearest." She had turned to **face** him. There was neither vehemence in her voice
- (2) This was yet another of his children to **face** him in defiance in this very kitchen.
- (3) not realise then that Rioja had stayed to **face** the killers. From the hut behind them

In these cases, we have more than a mere statement of the characters' physical position. In contexts of discourse, as in (1)-(2), there is a default expectation that the interlocutors will be looking at each other –or at least be face-to-face, so that establishing eye-contact would be a matter of split seconds–, as this is what daily experience of interaction with people in our western culture has taught us to expect. There is apparently something in the situation that makes an explicit mention of one character's positioning with respect to another particularly interesting or important. In cases like (3) there are strong lexical clues to point the reader towards a directional interpretation of *face* with a predominating confrontational element.

Whether such visual uses of *face* are to be interpreted as containing an element of confrontation or not is, however, not always a straightforward matter. The paper will therefore explore:

- in how far contextual clues are present that guide the reader in an interpretation of *face*;
- in how far such clues diverge from those already covered by LP;
- whether such clues indeed avoid the primings of the purely directional and purely confrontational uses;
- whether the reader should be accorded a more active role in the interpretation of what they read than they have been granted within the framework of LP.

Hoey, M. 2005. *Lexical Priming. A New Theory of Words and Language*. Abingdon/New York: Routledge.

**Agnès Tutin and Magda Florez (Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM, EA 609), Université Grenoble 3)**

*New approach, complex problem, promising results* : evaluative markers in French scientific writings in social sciences and humanities

In this study, we deal with evaluative markers relating to cross-disciplinary nouns of scientific writings like *approach, problem, results* ... Our goal is to explore the rhetorical and epistemological functions of these evaluative elements in three disciplines of social sciences and humanities (cognitive psychology, educational sciences, linguistics) through a semantic corpus study. Far from being neutral, scientific texts are now seen as argumentative texts with a pervasive authorial presence (Cf. Fløttum *et al.* 2006; Hyland 2005 ; Rinck 2006; Lores-Sanz *et al.* 2010), where persuasive strategies are widely used. Evaluation of scientific constructions deals both with authors' scientific constructions (e.g. *we propose a new method*) - and peers' scientific constructions (e.g. *X'approach is not fully appropriate* ...). The study of these devices helps to shed light on the criteria used to ensure scientific quality (e.g. novelty, salience, quality, inadequacy, for example). In line with previous studies in the field (Dahl 2008; Tutin 2010), we assume that these strategies vary significantly according to the discipline and according to the textual part (introduction, conclusion ...).

Evaluation, as outlined by Hunston and Thompson (2000), is a slippery notion. We adopt here their definition as "the broad cover term for the expression of the speaker or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about." (Hunston and Thompson 2005, 5), but we exclude affective lexicon. We only deal with explicit evaluative markers and adapt Kerbat-Orrechioni's (1980) definition of "subjective lexicon", dividing axiological lexicon (e.g. *appropriat*, "promising", in relation to a system of values) from non-axiological lexicon (e.g. *large, main* in relation to a norm). Our study specifically addresses adjectival and nominal evaluative markers relating to cross-disciplinary nouns (e.g. *the advantage of this approach, modest results*) and will study them in a large corpus of 2.5 M words including 50 French scientific articles and 5 French PhD dissertations in each of the 3 disciplines. The corpus is taken from the on-line Scientext corpus (<http://scientext.msh-alpes.fr>), a syntactically and structurally annotated corpus in TEI-XML format (see also Williams et Millon, 2009; Henderson et al. 2009). Our study will be twofold: we will first extract and study the evaluative markers from the whole corpus and study their semantic and rhetorical functions and their distribution in the text. Secondly, we will look more in detail at the discursive strategies that involve evaluative markers in a subset of 15 scientific articles.

Dahl Trine (2008). 'Contributing to the academic conversation: A study of new knowledge claims in economics and linguistics'. *Journal of Pragmatics* 40: 1184-1201.

Fløttum, Kjersti, Trine Dahl, & Torodd Kinn (2006). *Academic voices across languages and disciplines*. Amsterdam/Philadelphia: John Benjamins.

Henderson Alice, Tutin Agnès, Grossmann Francis, Barr Robert (2009). SCIENTEXT : A Corpus of French and English Scientific Texts., *British Association of Applied Linguistics Annual Conference*, 4 september 2009, Newcastle University.

Lores-Sanz Rosa, Mur-Duenas Pilar, Lafuente-Millan Enrique (2010). *Constructing Interpersonality: Multiple Perspectives on Written Academic Genres*. Cambridge: Cambridge Scholars Publishing.  
Hyland Ken (2005). *Metadiscourse*. London, New York: Continuum.

Kerbat-Orrechioni Catherine (1980). *L'énonciation : De la subjectivité dans le langage*. Paris:

Armand Colin.

Hunston Susan & Thompson Geoff (eds) (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Rinck, Fanny (2006). L'article de recherche en sciences du langage et en lettres. *Figure de l'auteur et approche disciplinaire du genre*. Thèse de doctorat en Sciences du Langage, sous la direction de F. Boch et F. Grossmann, Université de Grenoble 3-Stendhal.

Tutin Agnès (2010). 'Evaluative adjectives in academic writing in the humanities and social sciences'. In Rosa Lores-Sanz, Pilar Mur-Duenas, Enrique Lafuente-Millan *Constructing Interpersonality: Multiple Perspectives on Written Academic Genres*. Cambridge: Cambridge Scholars Publishing.

Williams Geoffrey & Millon Chrystel (2009). 'The General and the Specific : Collocational resonance of scientific language'. *Proceedings Corpus Linguistics 2009*. University of Liverpool. Available on : [http://ucrel.lancs.ac.uk/publications/CL2009/129\\_FullPaper.doc](http://ucrel.lancs.ac.uk/publications/CL2009/129_FullPaper.doc).

**Eleni Tziafa (Aristotle University of Thessaloniki)**

A register-diversified corpus in the Stock Market domain: The case of idiomatic and support verb expressions

One of the commonly accepted defining features of a text is representativeness, which is also a controversial issue, widely discussed in literature. The aim of this study is the adequate representation of a specialized language, by including different registers in a specialized corpus in the Stock Market domain, since analysis of the linguistic patterns across registers is of central importance for the linguistic description of a specialized language. The ultimate goal is to study and compare language varieties of formal and informal discourse from conversations in "trading slang" to academic prose. Collecting and creating structured bodies of specialized text, poses a great challenge to language technology, especially for languages with small numbers of native speakers such as Greek.

The texts under study are instances of different domain-specific registers. The text corpus consists of four sub-corpora of Greek texts in the Stock Market domain and it comprises over 18 million words. The time period covered by the corpus ranges from 1999 to 2010, a period marked by two major crises in Greece, a Stock Market crisis and a debt crisis.

Sub-corpus A consists of posted messages in public discussions in two internet forums, both dedicated to the stock market. This kind of forum appeared in Greece the last five years.

Sub-corpus B comes from journalistic texts, scanned from newspapers for the period 1999-2000 and consequently complemented with articles in an electronic format from 2000 to 2010, which were written at the same literary level.

Sub-corpus C comes from the website of the Athens Stock Market, which contains announcements, annual reports and articles dating from the year 2000. Sub-corpus C can potentially form a base for further study as parallel texts, since the texts included are accompanied with their English translations.

Sub-corpus D contains academic texts whose main focus are Money Markets and Stock Market Derivatives, which were provided from University modules. Furthermore, postgraduate and doctorate dissertations were used, available on-line.

All texts in the corpus were automatically annotated for part-of-speech and lemmatized through Unitex. This research actually forms part of the program for the development and gradual completion of the Greek version of Unitex - a text analysis system, already operating in many European languages. Using Unitex, verb/noun collocations were studied and more specifically idiomatic verbal phrases and support verb expressions. Hence, the aim of this case study is the identification, classification and analysis of these expressions in the Stock Market domain.

Benson, M. (1985) 'Collocations and Idioms', in R. Ilson (ed.) *Dictionaries, Lexicography and Language Learning* (ELT Documents 120), pp. 61–68. Oxford: Pergamon Press.

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

Biber, D., U. Connor & T.A. Upton (2007) *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

Biber, D. (2006) *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (1993) 'Using register diversified corpora for general language studies'. *Computational Linguistics*, 2: 219–41.

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

Kyriacopoulou, T. (2005) *L'analyse automatique des textes écrits : le cas du grec moderne*. University Studio Press, Thessaloniki.

McEnery, T. and A. Wilson (2001) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Sinclair J (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Williams G. (2002) "In search of representativity in specialised corpora: categorisation through collocation". *International Journal of Corpus Linguistics*, 7/1, 43-64.

Unitex, <http://www-igm.univ-mlv.fr/~unitex/index.php>

### **Peter Uhrig and Thomas Proisl (Universität Erlangen-Nürnberg)**

#### **A fast and user-friendly interface for large treebanks**

Although parsed corpora have been around for a long time, many are either comparatively small (e.g. ICE-GB) and/or come without an easy-to-use interface (e.g. Penn Treebank). The unintuitive nature of powerful corpus query tools such as TGrep2 has led researchers to create applications such as SearchTree (Nygaard/Bondi Johannessen 2004) or Treebank Search (Ghodke/Bird 2010). Due to their online interfaces, these tools provide considerable improvements in usability. However, they rely on Penn Treebank style phrase structure trees, which makes queries cumbersome for users who are used to more traditional labels such as "subject" or "direct object". We claim that a graphical interface (without a query language) with relatively simple labels of that kind is needed to lower the inhibitions of less technically minded linguists to use treebanks in their research. The most widespread and probably the most intuitive system of such labels is currently provided by Stanford Dependencies (SD, de Marneffe/Manning 2008). A big advantage of the SD framework is that it is built to read Penn Treebank style phrase structure trees. That means that an SD representation can

be obtained from dependency parsers trained on converted treebanks as well as from the output of many currently available phrase structure parsers (see Cer et al. 2010 for a comparison of speed and accuracy).

The dependencies do not necessarily form a tree structure, so they have to be represented more generally as a directed (possibly cyclic) graph. This is necessary to allow for instance for an object to depend on two coordinated verbs, and similar structures. Existing tree structure storage mechanisms thus cannot be used in our system. An overview of a high-performance system based on standard software and specifically designed for this purpose will be given in the paper.

In addition to finding concordance lines, the software offers the possibility to perform a collostructional analysis (Stefanowitsch/Gries 2003) of the collexeme variety (Stefanowitsch/Gries 2009:941). Thus it is possible to determine the association strength between (optionally partially lexically filled) syntactic structure in the form of a dependency graph and word forms or lemmata. For instance, we can look for the lemma “give” with a direct and indirect object node dependent on it where we only specify part-of-speech for the nodes, namely PRP (personal pronoun) for the indirect object and NNS (noun in plural) for the direct object. The direct object in turn has a dependent of the type determiner which is lexically filled by “the”. The collexeme searched for is the direct object plural noun. In the BNC, the most strongly associated form (though not the most frequent one) for this structure is “creeps” and while there are frequent literal examples (details, keys, ...) the list contains alternatives to “creeps” in a similar meaning, such as willies, shivers, horrors, jitters sorted by association measure.

The paper will include a live demonstration.

Cer, D. / M.-C. de Marneffe / D. Jurafsky / C. Manning (2010): “Parsing to Stanford Dependencies: Trade-offs between speed and accuracy”, in: Proceedings of the 7th Conference on International Language Resources and Evaluation, Valletta, Malta (LREC 2010).

de Marneffe, M. / C. Manning (2008): “The Stanford typed dependencies representation”, in: COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation.

Ghodke, S. / S. Bird (2010): “Fast Query for Large Treebanks”, in: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, USA, 267–275.

Nygaard, L. / J. Bondi Johannessen (2004): “SearchTree – A User-Friendly Treebank Search Interface”, in: Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004), 183–189.

Stefanowitsch, A. / S. Gries (2003): “Collostructions: Investigating the interaction of words and constructions”, in: *International Journal of Corpus Linguistics*, 8/2: 209–243.

Stefanowitsch, A. / S. Gries (2009): “Corpora and grammar”, in: A. Lüdeling / M. Kytö (eds.): *Corpus Linguistics: An International Handbook*. Berlin/New York: Walter de Gruyter, 933–952.

#### **Hiroko Usami (Lancaster University)**

How can corpora improve multiple choice grammar questions with possible answers?

While the main applications of general corpora to language teaching have been computer assisted language learning, data driven learning, and compiling teaching materials, it has only been recently that corpora have begun to be applied to language testing. Alderson 1996, Hughes 2003 and Barker

2004 have discussed the possibilities of such an approach, while Biber et al. 2004 and Hawkey and Barker 2004 have carried out practical studies in this area. Such research has focused on the creation of vocabulary tests and assessing learners' writing using corpora, while there has been little research focusing on multiple choice grammar questions using general corpora. On the other hand, among many proficiency tests, Japanese university entrance exams have been criticised for containing problematic items, due to the use of unnatural English and having more than one possible answer in multiple choice questions (e.g. Watkins et al. 1997; Kobayashi 2007). These problematic items have not been examined using corpora and corpus linguistic techniques.

Therefore in this talk, I present the extent to which corpus-based multiple choice grammar questions are effective, compared with existing non corpus-based multiple choice grammar questions. Firstly, I demonstrate that multiple choice grammar questions in Japanese university entrance exams testing a relative pronoun "who" used in an inserted clause can sometimes contain more than one possible answer. Evidence is provided via checks against the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), using corpus linguistic techniques such as collocations, concordances and clusters. Secondly, the test is improved by finding another possible answer based on using other frequent distracters presented in Japanese university entrance exams. To test the effectiveness of a set of new test questions created using this method, I give non corpus-based questions which contain multiple possible answers and corpus-based questions where a possible answer was replaced with another distracter to Japanese university students. The results are statistically analysed in terms of facility value, discrimination index and distracter analysis, and also the results of think-aloud protocol are examined in terms of qualitative analysis. The research shows that corpus-based questions where possible answers are replaced with other distracters are more effective, compared with original non corpus-based questions where multiple possible answers are contained in terms of statistical analysis. Thus it is recommended that general corpora containing authentic English should be used more often to check the English presented in exams in order to create more effective questions.

Alderson, J. C. (1996) "Do corpora have a role in language assessment?" In J. Thomas et al. (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman, pp. 248-259.

Barker, F. (2004) "Using corpora in language testing." *Modern English Teacher* 13, 2: 63-67.

Biber, D., S. M. Conrad, R. Reppen, P. Byrd, M. Helt, V. Clark, V. Cortes, E. Csomay and A. Urzua. (2004) "Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus." TOEFL Monograph Series No 25, Princeton, N. J.: Educational Testing Service.

Hawkey, R. and F. Baker. (2004) "Developing a common scale for the assessment of writing." *Assessing Writing* 9: 122-159.

Hughes, A. (2003) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Kobayashi, I. (2007) 'Daigaku Nyushi Mondai De Towareru oho No Keikou To Mondaiten'. [The trends and issues of grammar questions tested in university entrance exams]. *The English Teachers' Magazine* July 2007: 26-29.

Watkins, G., M. Kawakami and I. Kobayashi. (1997) *Korede Iinoka Daigaku Nyushi Eigo (Jyo)*. [University entrance English exams are OK? (volume 1)]. Tokyo: Taisyukan.

**Kateřina Veselovská (Charles University in Prague, Czech Republic)**

## A Corpus-based Study of the Constituent Negation in Czech

It has been widely accepted that there are two basic types of syntactic negation in Czech – the so-called sentential negation and constituent negation. The main criterion for distinguishing the two has been the fact that in the case of sentential negation the negative element *ne* is a prefix on the verb (1), whereas in the constituent negation, the particle *ne* immediately precedes the negated constituent (2a). However, according to the analysis of the Czech National Corpus SYN2005 [2], traditional approach to the negative sentences [1] usually neglects to mention that constituent negation involves another crucial distinctive feature – the contrastive character of the whole structure. In Czech, the contrastiveness is formally signalled by the adversative conjunction *ale* and the *ale*-clause is obligatory, as shown in (2b).

(1) *Petr včera ne-přišel.*

Peter NOM yesterday NEG came.

‘Yesterday, Peter didn’t come.’

(2)a. *Petr přišel ne včera, ale v neděli.*

Peter NOM came NEG yesterday but on Sunday.

‘Peter came not yesterday, but on Sunday.’

b. \**Petr přišel ne včera.*

Peter NOM came NEG yesterday.

In this talk I have two goals. (i) I will address a special type of constituent negation that has not been described in existing treatments of Czech negation, which I will call ‘contrastive negation’ (3). (ii) I will explore and formalize an analysis for common constituent negation and contrastive negation in Czech within a framework of dependency corpus, Prague Dependency Treebank 2.0 [4].

(3) *Petr ne-přišel včera, ale v neděli.*

Peter NOM NEG came yesterday but on Sunday.

‘Peter didn’t come yesterday, but [he came] on Sunday.’

Contrastive negation bears formal features of both sentential negation (prefix *ne-* on verb) and constituent negation (adversative paratactic structure). Semantically, however, the negative prefix does not negate the whole proposition (unlike in normal sentential negation – see [5]) and instead, it has scope only over one of the constituents. I will propose a dependency analysis and representation of this special type and relate it to the analysis of the standard negation structures. A crucial part of the analysis is an integration of information-structure constraints with the formal, morphosyntactic requirements of the patterns, using a tectogrammatical level of the Prague Dependency Treebank.

The analysis also leads to pointing out some commonalities and differences of constituent negation structures in Czech and English [3], e.g. the position of negative element *ne/not* in certain structures and the shape of the contrastive element (the adversative clause).

[1] Bernini, G.; Ramat, P. (1996) *Negative Sentences in the Languages of Europe. A Typological Approach*. Berlin, New York : Mouton de Gruyter.

[2] *Czech National Corpus - SYN2005*. Institute of the Czech National Corpus, Praha 2005. Accessible at WWW: <<http://www.korpus.cz>>.

[3] Haegeman, L. (1995) *The syntax of Negation*. Cambridge : Cambridge University Press.

[4] Hajič, J., E. Hajičová, J. Panevová, P. Sgall, J. Štěpánek, J. Havelka and M. Mikulová (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

[5] Hajičová, E. (1975). *Negace a presupozice ve významové stavbě věty*. Praha: Academia.

**Svetlana Vetchinnikova (University of Helsinki)**

**Semantic prosody as a communicative function of a unit of meaning**

In this paper I am going to argue that semantic prosody is an extremely valuable concept in the form proposed by Sinclair. Unfortunately, it has been popularized as a tendency of a word or lemma to co-occur with negative or positive collocates, which is undeniably an important observation but needs a different term, something like “attitudinal preference” (Hunston 2007) as it is, arguably, not the original phenomenon Sinclair had in mind. One of my principal concerns is that when semantic prosody is conceptualized as a kind of negative/positive “prosodic valence” of words, it does not prove to qualify as a psycholinguistically important phenomenon (Ellis et al. 2009).

Therefore, in this paper I am going to take Sinclair’s approach as my fundamental starting point and show that:

1. Semantic prosody as a concept is inseparable from the search for units of meaning which in turn is inseparable from the idiom principle idea. A unit of meaning is a sequence of lexical items which is produced not as a result of subsequent paradigmatic lexical choices and application of the rules of grammar but as a single choice of meaning and an activation of internal syntagmatic associations between its elements which glue them together in this sequence – an “occasion where one decision leads to more than one word in text” (Sinclair 1987: 321). This one decision is the choice of the semantic prosody, which is “on the pragmatic side of the semantics/pragmatics continuum” (Sinclair 1996/2004: 34).

2. Only an “independent” unit of meaning can be characterized by semantic prosody. The division of lexis into orthographic words or lemmas is not relevant in this respect.

3. As an obligatory component of a unit of meaning, semantic prosody should be clearly distinguished from semantic preference which is optional. For the same reason semantic prosody is not inherently evaluative, evaluation being just one type of semantic prosody.

4. Semantic prosody is not “hidden”, what is not readily available to intuition is a usage pattern of delexicalized words i.e. words which do not have an independent meaning of their own. Incidentally, this fact indirectly indicates that the mental lexicon is organized according to meanings rather than according to words.

In sum, semantic prosody as a kind of a pragmatic/functional meaning which keeps a unit of meaning together can serve as a bridge between corpus linguistics and psycholinguistics. It makes a unit of meaning fit into the pattern-finding and intention-reading conception of human language

acquisition (Tomasello 2003).

Ellis, Nick C., Eric Frey, & Isaac Jalkanen. 2009. 'The psycholinguistic reality of collocation and semantic prosody (1): Lexical access'. In Ute Römer & Rainer Schulze (eds.), *Exploring the lexis-grammar interface*, 89-114. Amsterdam: John Benjamins.

Hunston, Susan. 2007. 'Semantic prosody revisited'. *International Journal of Corpus Linguistics*, 12(2), 249-268.

Sinclair, John M. 2004. *Trust the text*. London: Routledge.

Sinclair, John M. 1987. 'Collocation: a progress report'. In Ross Steele & Terry Treadgold (eds.), *Language topics: Essays in honour of Michael Halliday*, 319-331. Amsterdam: John Benjamins.

Tomasello, Michael. 2003. *Constructing a language*. Cambridge, MA: Harvard University Press.

### **Benet Vincent (University of Birmingham)**

#### Modality and the V wh pattern

The advent of large computer-readable corpora has allowed researchers greater ease of access to language as it is used and has helped to highlight the close association between form and meaning pointed out by Sinclair (1991). Research into corpora has also provided evidence of the interdependence of paradigmatic and syntagmatic choices, which have traditionally been considered distinct (Hunston, 2003), thereby offering support to Sinclair's (1991; 2004) claim that the phrase is the 'essential building block of English' (Hunston, 2003: 58). An example of this interdependence is that the paradigmatic choice of *decide* rather than *decided* seems strongly related to the syntagmatic choice of a following *wh*-clause instead of a *that*-clause (Hunston, 2003). Moreover, the association between *decide* and the *wh*-clause can also be shown to be functionally motivated, in that, in most cases, the decision referred to has yet to be made, in contrast with *decided that*, where the decision has generally been made (ibid.).

The association between the base form *decide* and *wh*-clause coupled with the fact that the base form is also associated with modal or modal-like language led Hunston (2011) to hypothesise that modal-like language is attracted to the **V wh** pattern, that is, verbs that govern *wh*-clauses (Francis et al, 1996). However, it has yet to be demonstrated whether this observation about the lemma DECIDE reveals a systematic pattern applying to all verbs that frequently govern interrogative clauses.

This paper reports on a study that sets out to test Hunston's hypothesis by investigating verbs that frequently appear in the **V wh** pattern in the British National Corpus (BNC), using the CQP-edition of the online interface BNCweb (Hoffman & Evert, 2008). A quantitative approach is used to establish whether there is a significant attraction between *wh*-clauses and modal or modal-like language across verbs that govern the **V wh** pattern. This involves searching for instances of verbs complemented by *wh*-clauses and preceded either by a modal auxiliary or the infinitival operator *to*, a relatively simple task using the CQP-edition of BNCweb (Hoffman & Evert, 2008). A more qualitative methodology is used to identify the kinds of phraseologies that emerge from the data and to establish the extent to which queries based on the infinitival operator *to* followed by **V wh** can be said to result in hits containing 'modal-like expressions' (Hunston, 2011).

It is thought that this study will contribute to understanding of the relationship between paradigmatic and syntactic choice, as well as suggesting reasons why such a relationship might exist. There are also thought to be potential implications of this research in terms of producing more

accurate and therefore helpful grammatical descriptions which can inform English language pedagogy.

Francis, G., Hunston, S. and Manning, E. (1996) *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

Hoffman, S. & Evert, S. (2008) BNCweb (CQP-Edition). Online resource. Available at [<http://bncweb.lancs.ac.uk/>] (Accessed 5/10/2010)

Hunston, S. (2003) 'Lexis, Wordform and Complementation Pattern: a Corpus Study'. *Functions of Language* 10: 31-60.

Hunston, S. (2011) *Corpus approaches to evaluation: Phraseology and evaluative language*. New York/London: Routledge.

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

Sinclair, J. (2004) *Trust the text*. London: Routledge.

**Brian Walker (University of Huddersfield) and Dan McIntyre (University of Huddersfield)**

Modality and freedom of the press: A corpus based study of modality in Early Modern English journalistic writing and its relation to changing press freedom

According to historians of journalism, early newspapers were fairly anodyne publications. It was rare for writers to express an opinion about the news they reported, for fear of governmental reprisals. This paper reports on a project that investigates modality in Early Modern English news reportage, the results of which suggest a link between modality and changes in freedom of expression in print journalism.

In order to investigate the use of modality in journalistic writing we built and analysed a corpus of Early Modern English news, using texts from 1620 to 1720. Our corpus consists of around 500,000 words of press reportage which we obtained from the Burney Collection, which contains over 1270 newsbooks, newspapers, pamphlets and other news texts from 1600 - 1800. The corpus was constructed manually, transcribing the facsimiles of news publications available from the collection into machine readable files. The resulting corpus was marked-up for modality both formally and functionally. The formal mark-up was predominantly automated and captured modal auxiliary verbs, modal lexical verbs, modal adjectives and modal adverbs. The functional mark-up required manual analysis for which we used the model proposed by Simpson (1993). This model divides modality into three types: deontic, which expresses obligation and commitment; boulomaic, which expresses desire; and epistemic, which expresses knowledge and belief.

Using the mark-up, we ascertained the frequencies of different modal forms and functions, and their distribution across the corpus. We also carried out log-likelihood testing to determine the statistical significance of differences in modal frequencies across the time period represented by the corpus. We then attempted to correlate significant fluctuations in modality with wider socio-political events.

In our talk we will describe the construction of our corpus, the issues in annotating it, and the issues in applying a modern functional modality model to an older form of English. We will also present the results of our analysis. Our project attempts to relate changes in journalistic writing to changes in press freedom, and we will report a correlation between some of our figures and changes in legislation relating to printing and journalism. We will also show that the two-pronged form/function analysis enabled us to describe in more detail changes in modality in journalistic writing 1620-1720.

Simpson, Paul (1993) *Language, Ideology, and Point of View*. London: Routledge

**Fang Wang (University of Birmingham)**

A Corpus-based Discourse Analysis of Depression in Western and Chinese media

My research investigates different social constructions (Burr, 2003) of an increasingly relevant aspect of social life, namely mental disorders, more specifically, depression in Western and Chinese news media, aiming at delivering a contribution to people's understanding of the link between the discourse and the social reality of depression.

I understand discourse as all that has been said to the discourse object in question (Teubert, 2010). Therefore, to examine the meanings of a certain object, a researcher has to firstly, define the discourse that s/he is going to look at, and then discuss the meaning negotiations that have taken place within the defined discourse. The researchers' findings in this way, can only be considered as a construction on the meanings of the object in question, which should be further submitted to a wider discourse community.

In this research, to examine the meanings of mental depression in both UK and China, I define my discourse as represented by two large diachronic corpora consisting articles in which *depression* occurs in all national newspapers in both UK and China from 1980 to 2009. Furthermore, both the corpora are divided into five phases based on the frequency distribution patterns of the articles about depression. It is found that in the beginning phases of British subcorpus, 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 was more constructed as a biological disease and thus can be treated by antidepressants. The final phase of British subcorpus constructs depression as a rather complex problem that needs more scientific research, and more integrated forms of treatment are recommended. By contrast, in our Chinese subcorpus, it is found that depression has been always considered as being caused by external factors, such as problems of human relations 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.

My talk then, will firstly review the literature part of cultural roots and depression studies in both western and Chinese contexts, including the work of depression study in cross-cultural psychiatry, pioneered by Arthur Kleinman (1985). Then, I will show how I have used corpus research methodology to reveal such striking differences between British and Chinese subcorpora about depression. Therefore, this research not only provides a social constructionist research on how depression has been understood in both British and Chinese news media in the last 20 years, but also proves the role of corpus in revealing larger discourse patterns, those patterns that can frame our beliefs and understandings towards the world.

Burr, V. (2003) *Social Constructionism* (second edition). London: Routledge

Kleinman, A., & Good, B. (Eds.), (1985). *Culture and depression: Studies in the anthropology and cross-cultural psychiatry of affect and disorder*. Berkeley: University of California Press.

Teubert, W. (2010) *Meaning, Discourse and Society*. Cambridge: Cambridge University Press.

**Stephen Wattam (Lancaster University), Paul Rayson (Lancaster University), and Damon Berridge**

Folksonomies as Document Key Word Summary Engines

Folksonomies, socially-compiled taxonomies of data, are increasingly popular online, and have

recently attracted attention in computational linguistics and information retrieval as a source of easily processed linguistic data. They ostensibly offer what is generally a stumbling block to many NLP techniques --- a ready supply of human-reviewed text and summary information. The manually assigned tags in folksonomies can be seen as the counterpart to key words in corpus linguistics that are extracted automatically through tools such as WordSmith and Wmatrix.

This paper examines one folksonomy, the oft-studied del.icio.us social bookmarking service, in an attempt to assess the extent to which folksonomies (in this case tag-based) represent information already accessible using existing information retrieval metrics. The metrics compared have been chosen for their different philosophical justification, and include five different IR measures (Log-Likelihood, TFIDF, RFR, WordNet synonym connectedness and position of first occurrence), as well as two measures derived from tags themselves to act as a baseline. Log-likelihood is used in corpus linguistics as an alternative to the Chi-squared statistic to calculate key words.

Through the analysis of a large sample of web bookmarks tagged by users in del.icio.us, we establish that folksonomy tag usage does not fit commonly used patterns of summary length, and so breaks some assumptions regarding its use as a summary.

By taking the tags that are manually assigned and the key words produced from statistical measures when applied to the full text of the web pages that have been bookmarked, we are able to compare human and machine choices for key words. Through dimensionality reduction, we identify that tag measures do indeed appear to offer new information, and construct a series of classifiers designed to best predict tagging accuracy. These classifiers are based upon the reduced forms of the original IR metrics, and offer reasonable performance at the expense of complexity. One measure, log-likelihood, is identified as a particularly pure predictor for tag data.

We conclude that, for practical and parsimonious prediction, as well as a more comprehensive understanding of human tagging choices for searching documents, more complex language models must be used in order to identify the nature of, and extract, the tag-specific properties. These insights will also help to investigate the nature of key words in corpus linguistics.

**Xiangqing Wei (Bilingual Dictionary Research Center, Nanjing University), Dongbo Wang (Department of Information Management, Nanjing University), and Yundong Geng**

Some Changes in Discourse Features of Modern Chinese (1949-2009): Corpus evidence of English-Chinese language contact

The evolution of modern Chinese (1919- ) has long been marginalized if not neglected by many Chinese linguists. Two major factors may help explain the status quo. One is the common misconception in the academic circle that the history of modern Chinese is not long enough to be worth observing for any tangible signs of diachronic linguistic change. And the lack of available historical data from corpora is considered the other --- a practical limitation. However, a recent release of the diachronic data from the People's Daily Corpus (1949-2009) may hold out some hope for future studies. In this modern Chinese corpus, all Chinese words are segmented and each segmented word is tagged with its part-of-speech. The abundant data will obviously support various related linguistic studies, in which the study of English-Chinese language contact heavily counts. As a matter of fact, in the 90-odd years of modern Chinese history, the later 60-year part from 1949-2009 is definitely most eventful, in which the Chinese language itself has also undergone a lot of linguistic modifications. And within the possible parameters for the linguistic changes in modern Chinese, influences of the indirect language contact between English and Chinese languages did play a major role, especially during the latest 30 years when China implements firmly the reform and open policy put forward by the government. On the basis of the literature relevant to the study of modern Chinese language and its history, three major periods of English-Chinese language contact have been

distinguished. They are the period of loose connection between English and Chinese language (1949-1966), the period of little connection between English and Chinese language (1967-1978) and the period of close connection between English and Chinese language (1979-2009). This paper with hard evidence collected from this diachronic corpus depicts some changes in discourse features of modern Chinese, which hopefully will serve as a manifestation of English-Chinese indirect language contact. The investigation involves two major aspects concerning discourse analysis, namely the changes in discourse markers and the structural features. Several statistical methods of corpus-based research, including mutual information, likelihood ratio, and chi-square test, are used for the detailed concrete explorations of how the Chinese language has evolved from one historical period to another. The account of the explorations will be given comprehensively, with reference to discourse features of standard modern Chinese under the influence of indirect English-Chinese language contact in each different period. It is hoped that the present study will remedy the omission of corpus analysis of indirect English-Chinese language contact.

**Peter White (University of New South Wales)**

**Translatability and Attitudinal Meaning in the World's News Media**

The paper provides some fresh insights into cross-linguistic commensurability issues associated with the translation of explicitly attitudinal words and phrases (terms which convey either positive or negative assessments). The findings reported follow from a study employing semantically-annotated, parallel corpora of English-French and English-Chinese texts.

The paper is broadly concerned with the degree of attitudinal equivalence which is being achieved in the context of the translations into English of news media texts which are currently available to casual readers via the World Wide Web, and more specifically translations into English of French and Chinese news media texts. It considers this question of attitudinal equivalence both with respect to human translations (for example the English translation of *Le Monde Diplomatique* and the English translation of items from the *People's Daily*, the *Renmin Ribao*) and with respect to translations by the various automatic, machine-translation software packages which are currently available. It reports what are probably unsurprising findings that attitudinal equivalence is very much a hit-and-miss affair in the context of the machine translations. Perhaps more surprising are findings that the different translation software packages are not created equal in terms of their ability to handle explicitly attitudinal terms. The paper also reports findings that, according to the measure of "attitudinal equivalence" developed for the study, the human translations vary greatly from attitudinal term to attitudinal term with respect to of the degree of equivalence achieved. One key finding to be reported is that attitudinal equivalence seems to be related to how frequently the translated version of the term occurs in the Birmingham Bank of English corpus and in pages returned by a Google search of the Web. The more frequently occurring the translated term, the greater the likelihood that the translation will be deemed to be equivalent with respect to its attitudinal valeur. The less frequent, the more likely that the translation will be deemed incommensurate, and, interestingly, the more likely that the different machine translation packages will select different lexemes for their translation. These findings suggest more general conclusions with respect to a distinction between attitudinal terms which are more stable in terms of their attitudinal valeur and those which are less stable.

The methodology employed was as follows. A number of corpora were compiled from instances of original French and Chinese news media texts, linked with translations of those texts both by human translators and by machine translation software. Using Extensible Markup Language (XML), all instances of explicitly attitudinal terms were manually annotated by native speakers by reference to the taxonomy of attitude types and sub types as outlined by the appraisal framework (see for example, Martin & White 2005).

Martin, J.R., & White, P.R.R., 2005, *The Language of Evaluation – Appraisal in English*, Palgrave/Macmillan, Houndmills.

Garside, R., Leech, G., & McEnery, T (Eds), 1997, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Addison Wesley Longman, London.

### **Johannes Widmann (University of Tübingen)**

#### Lexical clusters in small pedagogic corpora

Pedagogic corpora as they have been conceived in the SACODEYL and BACKBONE ([www.uni-tuebingen.de/backbone](http://www.uni-tuebingen.de/backbone)) projects are an important step in bringing corpora closer to use in the communicative language classroom. Each corpus consists of interviews where the interviewees talk about pre-defined topics. With their focus on pedagogic purposes, the corpora provide teachers with a didactic infrastructure that harmonizes better with their teaching practice than it seems to be the case with typical large representative corpora (see Braun 2005 for an overview of the pedagogic concept behind the corpora).

In this talk I will look at the lexical potential of these corpora. Because of their carefully selected contents they allow for lexical tasks that go beyond classic data-driven learning. What is needed, however, are semi-automatic procedures for making the lexical potential of small corpora explicit to users and for integrating this lexical potential in a topic-based language learning approach. In order to find an answer to these two issues, I will present an analysis of the lexical potential of the English SACODEYL and BACKBONE corpora. The focus will be on various types of clusters, such as polywords, fixed expressions and collocations which are the key elements in a lexical approach as defined by Lewis (1993, 1997). I will extract such clusters using available corpus tools and methodologies. I will then look at the overall results across all interviews in order to analyze and describe the corpora with regard to their lexical cluster characteristics and I will discuss how much these automatic results are usable in language teaching with respect to a topic-based teaching approach.

The results will then be sorted per interview topic and I will compare the topics according to their lexical potential. Based on this lexical potential I will attempt to establish a pedagogical classification and sequencing of the topics. This classification will move from a description of the clusters to an interpretation of what the clusters are good for with respect to topic-based language learning. My interpretation will take into account a complementary inspection by hand which the automatic analysis has not revealed.

As a conclusion, I will argue that only once we achieve a systematic identification and pedagogical classification of topic-based multi-word units, we can contribute to making teachers aware of the advantages of using corpus techniques. It will be much easier for them to get a multi-faceted overview of the lexis of a text. Many researchers have claimed that such an awareness of corpus techniques is the key element for getting teachers more interested in using corpora (Frankenberg-Garcia 2006). My talk will thus demonstrate how even small corpora can fruitfully be exploited for a topic-based lexical approach.

Braun, S. (2005). "From pedagogically relevant corpora to authentic language learning contents", *ReCALL* 17:1, 47-64.

Frankenberg-Garcia, A. (2006). "Raising teachers' awareness to corpora", keynote speech at TaLC 7, *Bibliothèque Nationale de France*, 1-4 July 2006:

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M. (1997). *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

**Kate Wild (Lexical Computing Ltd.), Diana McCarthy (Lexical Computing Ltd.), Andrew Church (University of Brighton), and Jacquie Burgess (University of East Anglia)**

A Corpus Linguistics Analysis of Ecosystems Vocabulary in the Public Sphere

Over the past few decades there has been a growing body of research into the language used to discuss environmental issues. Research has included analysis of discourse features such as nominalization, passivization and modality (Schleppegrell 1997, Goatly 2001, Kuha 2009), as well as analysis of the semantics of environmental vocabulary such as 'nature', 'pollution' and 'carbon' (Mühlhäusler 2001, Nerlich and Koteyko 2009, Dillon 2010). Some studies are based on small samples of texts collected by the authors; others are based on larger corpora but focus on a small selection of linguistic features. Until now there has been no corpus linguistics study of a broad range of vocabulary related to the environment: our research aimed to fill that gap.

We carried out a study of over one hundred lexemes related to ecosystems and the environment. Our aim was to examine how these lexemes are used in public discourse, to discover their key collocates, and to consider how corpus evidence can indicate the evaluative uses of certain terms (cf. Hunston 2007). The study used purpose-built corpora of language relating to ecosystems – from academic websites, government websites, NGO websites, news media and blogs – and compared these with UKWaC. The interface used was Sketch Engine, which facilitated several aspects of the study:

- The Thesaurus feature helped us to generate a list of lexemes related to ecosystems and the environment. (These were then checked and supplemented by experts in the environmental field.)
- The Word Sketch feature allowed us to identify statistically significant collocates of these lexemes.
- The Sketch-Diff feature allowed us to compare the collocates of related lexemes such as urban/rural.

In addition, random samples of concordance lines were manually scrutinized in order to identify the subjectivity or objectivity of the texts, and the positive, negative or neutral connotations of particular lexemes as used in context.

We will present a selection of our findings, including:

- Evidence for the conceptualization of the environment: for example the framing of nature as a commodity in phrases such as 'ecosystem services' and 'heritage assets'; and the concept of nature as either including or separate from humans.
- Evidence for the subjective, emotionally charged or sceptical use of particular terms such as 'climate' and 'science'.
- Evaluative uses of terms such as 'green', 'wilderness' and 'expert'.
- The avoidance of stating agents that cause environmental problems, even in the language of environmentalists.

By combining statistical analysis with manual analysis, we offer both a quantitative corpus analysis and a qualitative discourse analysis of environmental language in the public sphere. We argue that conceptualization of the environment differs according to text type, but that there are several

recurring themes in the public data that we examined.

Denise Dillon (2010) 'People, environment, language and meaning: values in nature and the nature of "values"' *Language and Ecology* 3:2 [<http://www.ecoling.net/journal.html>]

Alwin Fill and Peter Mühlhäusler (eds) (2001) *The Ecolinguistics Reader: Language, Ecology and Environment* London: Continuum

Andrew Goatly (2001) 'Green Grammar and Grammatical Metaphor, or Language and Myth of Power, or Metaphors We Die By' in Alwin and Mühlhäusler, pp203-225

Susan Hunston (2007) 'Semantic prosody revisited' *International Journal of Corpus Linguistics* 12:2, pp249-268

Mai Kuha (2009) 'Uncertainty about causes and effects of global warming in U.S. news coverage before and after Bali' *Language and Ecology* 2:4 [<http://www.ecoling.net/journal.html>]

Peter Mühlhäusler (2001) 'Talking about Environmental Issues' in Alwin and Mühlhäusler, pp31-42

Brigitte Nerlich and Nelya Kotevko (2009) 'Compounds, creativity and complexity in climate change communication: The case of "carbon indulgences"' *Global Environmental Change* 19, pp345-353

Mary J. Schleppegrell (1997) 'Agency in Environmental Education' *Linguistics and Education* 9, pp49-67

#### **Annelore Willems and Gert De Sutter (University College Ghent, Ghent University)**

Grammatical complexity versus discourse status as determinants of PP placement in original and translated Dutch

The present corpus-based study discusses a well-known type of syntactic variation in Dutch, viz. prepositional phrase (PP) placement. Language users can put PPs either before V-final (A) or after V-final (B).

(A) *Niemand weet dat ik van mannen |hou|<sub>V-final</sub>.*

*No one knows that I with men am in love.*

(B) *Niemand weet dat ik |hou|<sub>V-final</sub> van mannen.*

*No one knows that I 'm in love with men.*

In this study we investigate which variables determine this variation, and how these variables relate to each other. Furthermore, we will investigate whether these factors have an identical effect in translated vs. non-translated texts, as previous research has shown that translated texts differ in a systematic manner from non-translated texts (Olohan & Baker 2000).

In order to do so, we extracted all PPs from the Dutch Parallel Corpus (DPC), which is a 10-million-word, parallel corpus of Dutch, English and French (Macken et al. 2011). It contains six different text types, but for the present study we opted to look only at PP variation in journalistic texts. After manually checking the relevance of the extracted corpus data, we obtained a dataset containing 1450 clauses with a PP either before V-final or after V-final. We coded the data for the variables grammatical complexity and discourse status, as previous research has shown that these have a significant impact on different kinds of constituent order phenomena (e.g., Arnold et al. 2000; Hawkins 1994; Wasow 2002; Van Bergen & De Swart 2010). Although these factors have already been described in earlier research into PP placement in Dutch (Jansen 1978), a simultaneous

comparison has not yet been conducted. The effect of both variables will be analyzed and interpreted by means of a binary logistic regression model.

Arnold, J. E., T. Wasow, A. Losongco & R. Ginstrom (2000). 'Heavyness versus newness: The effects of structure complexity and discourse status on constituent ordering'. *Language* 76 (1). 28-55

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Jansen, F. (1978). 'Hoe krijgt een spreker zijn woorden op een rijtje? Taalgebruikaspecten van de 'PP over V' constructie'. In: J. G. Kooij (ed.) *Aspekten van de woordvolgorde in het Nederlands*. Leiden: Vakgroep Nederlandse Taal- & Letterkunde. 70-104.

Macken, L., O. Declercq & H. Paulussen (2011) 'Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus'. *META* 56(2)

Olohan M. & M. Baker (2000). 'Reporting that in translated English: evidence for subconscious processes of explicitation?' *Across Languages and Cultures* 1(2): 141-158.

Van Bergen, G & P. De Swart (2010). 'Scrambling in spoken Dutch: Definiteness versus weight as determinants of word order variation'. *Corpus Linguistics and Linguistic Theory* 6-2. 267-295.

Wasow, T. (2002). *Postverbal behavior*. Stanford, CA: CSLI Publications.

**Geoffrey Williams and Chrystel Millon (Université de Bretagne-Sud)**

Palmer, Firth and Internet: Drawing together collocational threads

In corpus linguistics, the contribution of Palmer to collocation studies is often overlooked. However, Palmer's Second Report on English Collocations published in Japan in 1933 has been the inspiration for many major threads in phraseology, even if few have actually had access to the work itself. When Firth started looking into collocation, he almost certainly had never heard of Palmer.

Thus were the parallel worlds of overseas ELT and British academia at a time before Internet made data exchange and cross disciplinary exchange easy. The consequence is that traditional phraseology, and much pre-corpus lexicography, and corpus linguistics developed on parallel lines. Those lines were effectively drawn together in the COBUILD initiative, although the Palmer connection remained largely forgotten.

Part of the cause lies in the fact that these two approaches are based on radically different visions of language. Phraseologists and lexicographers seek to tame language so as to list and classify for inclusion in published works. This requires an essentially static vision of collocation where phraseological units are treated as if created ex nihilo and are simply found and classified on purely linguistic grounds as to what may and what may not be termed as a collocation.

The NeoFirthian approach developed by John Sinclair within the context of corpus linguistics is very different in that it places collocation at the heart of language as an essentially dynamic process in which meanings are created and exploited within textual contexts. This requires a much wider vision of collocation rather than simply reducing it to a series of part of speech groupings, with occasionally a smattering of pragmatic or linguistic considerations. The advantage of corpus linguistics is that it allows an analysis of dynamic collocation whilst providing the material for more reductive phraseological or computational exploitation of the data.

This paper intends to look at three issues; the development along parallel lines of phraseological and corpus linguist collocation, reinstating the place of Palmer whilst underlying the centrality of collocation as a language phenomenon, an overview of criteria for restricted collocation and finally how the threads can be drawn together in the extraction and analysis of collocational data from Internet.

**Nicholas Wood and Nicolai Struc (both Reitaku University, Japan)**

**Investigating Syntactic Complexity in L2 Narrative and Argumentative Writing**

Syntactic (or grammatical) complexity refers to the range and degree of sophistication of forms that surface in language production (Ortega, 2003), with the emergence of more elaborate syntactic patterning as learners develop and access more complex subsystems of language (Foster & Skehan, 1996).

This presentation will report on a corpus-based investigation of syntactic complexity in texts written in two genres - narrative and argumentative - produced by 25 L2 learners at the start of a university writing program and at the starts of their second and third years. The study can be considered a quantitative exploration of Rescher's (1998) three dimensions of ontological complexity: the compositional in terms of the relative frequency of syntactic constituents of text (words and clauses), the structural in terms of the occurrence of constructed units (clause types, sentence types and T-units), and function as evidenced by the use of specific constituents and structures in the production of genre.

Syntactic complexity genre differences in L1 writing have been noted by Crowhurst and Piche (1979), who found that T-unit length was significantly greater in argument than in narration, and by Crowhurst (1980), who argued that argumentative writing necessitates the logical structuring of propositions, and this is expressed syntactically in the more frequent use of subordination, and hence longer T-units, than narratives. Beers and Nagy (2009), in a more recent study of L1 texts, found that clause length positively correlated with quality for persuasive essays, while clauses per T-unit positively correlated with quality for narratives but negatively with quality for persuasive essays.

In a study concluded in 2010 (to be published) of narrative and argumentative texts produced by 170 L2 learners at the starts of their first and second years at university, we found statistically significant differences on nine measures of fluency and complexity between texts produced in the two genres. The present research tracks the development 25 of those 170 students into their third year.

The presentation will provide a brief background to methodological issues and then report on the findings of this longitudinal study. The investigation aimed not only to compare the syntactic complexity evidenced in the two genres, but also to develop and utilize an analytical approach that could examine complexity on the basis of a learner's own orthographic production, examine objective syntactic relations (e.g., between a simple sentence and following subordinated clause fragment) by the use of sentence reconstruction, and examine the range of clause types used by individual learners and by the population as a whole. The researchers also developed a statistically-based sentence variety index by which to evaluate the range of sentence types used in text, a tool that may have wider application.

Beers, S.F. & Nagy, W.E. (2009). 'Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?' *Reading and Writing*, 22, 185-200.

Crowhurst, M. & Piche, G.L. (1979). 'Audience and Mode of Discourse Effects on Syntactic Complexity in Writing at Two Grade Levels'. *Research in the Teaching of English*, 13, 101-109.

Crowhurst, M. (1980). 'Syntactic Complexity and Teachers' Quality Ratings of Narrations and Arguments'. *Research in the Teaching of English*, 14, 223-231.

Ortega, L. (2003). 'Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing'. *Applied Linguistics*, 24, 492-518.

Rescher, N. (1998). *Complexity: a philosophical overview*. Transaction Publishers: New Brunswick, New Jersey.

**Chi Cheung Ruby Yang (Lancaster University)**

Gender representation in Hong Kong English textbooks: An investigation of collocations of gendered terms in a small corpus

According to Hong Kong Federation of Women (2006), gender stereotyping and gender-based biases still exist in Hong Kong, and gender stereotyping is commonly found in teaching materials and textbooks. In the English lessons of many schools in Hong Kong, a large part of learner input is still provided by textbooks. Because of this reason, a small corpus of Hong Kong primary English textbooks was built and the collocations of gendered terms in the corpus were analysed in order to find out if gender stereotyping is still an issue in the recently published textbooks. Baker (2008) suggests that corpora are useful tools in language and gender research.

The purpose of this study is to investigate if gender bias and gender stereotyping still exist in the recently published textbooks in Hong Kong. To achieve the aim, a small corpus of textbooks (a series of twelve New Magic English textbooks for Primary One to Primary Six students) was compiled and the corpus software AntConc was used to analyse the collocations of gendered terms He/he, She/she, Man/man, Woman/woman/women, Boy/boy/Boys/boys, and Girl/Girls/girls in the corpus. From the results, it seems that the writers of this textbook series have become aware of the issue of gender equality and made attempts to avoid gender stereotyping. Therefore, females were no longer considered as delicate or weak but were described as even stronger than males. In domestic and occupational roles, females were no longer portrayed only as housewives who did all the household work, but they also worked in society as a doctor or a school principal who takes care of the whole school, and males did share the household work with females at home. Besides that, males and females engaged in different sports activities. Not only can males be good at sports, but also females, and a PE teacher can be a female.

However, the stereotyped images of males wearing shorts, Jeans or shirts and having big feet and females putting on skirts or dresses and having small hands still exist, and the masculine generic terms for occupation (fireman and postman) could still be found. In addition, while it was found that the male terms have more collocates and both males and females were attributed with positive characters, the female term (the node word She/she in this corpus) still has more positive collocates. It is impossible for the textbook writers to count the exact number of collocates of different semantic groups for the gendered terms in order to ensure that the textbooks are totally free of gender-bias. Nevertheless, it is suggested that the gender-neutral terms should be used to replace those masculine generic terms in the names of occupation, given the fact that textbook may have strong impact on children's development of values and attitudes.

Baker, P. (2008). 'Eligible' bachelors and 'frustrated' spinsters: Corpus linguistics, gender and language'. In K. Harrington, L. Litosseliti, H. Sauntson, & J. Sunderland (Eds.), *Gender and language research methodologies* (pp. 73-84). New York: Palgrave Macmillan.

Hong Kong Federation of Women (2006). *Submission of NGO report to Committee on the Elimination of Discrimination against Women on the implementation of CEDAW in Hong Kong*. Retrieved 4 July,

2010, from <http://www.hkfw.org/eng/commenteCEDAWreport.html>

# Pecha Kuchas

**Cristina Acunzo (São Paulo Catholic University, Brazil, PUC-SP)**

Teaching English as a Foreign Language to Professionals Using Corpora

Teaching English as a foreign language to professionals of specific fields has been a great challenge, as the materials available in the market fail to meet these students' necessity of communicating professionally. In order to fill this gap, this research aims at developing a methodology for the preparation of classes and materials using corpora for the teaching of English as a foreign language to professionals in Advertising. In this research, Corpus Linguistics (Sinclair, 1991; Berber Sardinha, 2004) provides the main theoretical framework and the Theory of Complexity (Morin, 1999) provides some principles based on which the tasks for the classes are designed. The basic methodology of this research is: (1) the construction of a corpus of 1 million words composed of written texts and transcriptions of videos which provide information on the world of Advertising; (2) the analysis of the corpus through its comparison with the British National Corpus, used as reference, in order to identify the distinctive lexicogramatical patterns of the area using tools of the Wordsmith Tools suite; (3) the selection of texts and the lexicogramatical patterns and the development of the classes and tasks, centered on the concordance and on the text (Berber Sardinha, 2009); (4) use of the activities in class to get students' reactions. Based on the pilot study, we believe the results of this research will contribute greatly to future development of classes and materials for the teaching of English as a foreign language with the use of corpora in specific contexts.

Berber Sardinha, T. (2004). *Linguística de Corpus*. São Paulo: Manole.

\_\_\_\_\_. (2009). *Pesquisa em Linguística de Corpus com Wordsmith Tools*. Campinas: Mercado de Letras.

Morin, E. (1999). *La tetê bien faite. Repenser la réforme. Réformer la pensée*. Paris: Seuil.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

**Susan Bell (University of Glasgow)**

A Corpus of Modern Scottish Gaelic

The Celtic and Gaelic subject area at the University of Glasgow is home to the Digital Archive of Scottish Gaelic, a British Academy recognised project that will digitally preserve a wide range of Scottish Gaelic texts.

This work-in-progress presentation will be based on my ongoing College of Arts funded PhD research at the University of Glasgow.

It will cover how I am organising these digital resources into a corpus in order to investigate the orthographic history of Scottish Gaelic. The talk will discuss aspects of the corpus design including sampling, representativeness, tagging and markup. The issues around building a corpus with a 'non-standardised' language will also be discussed.

**Hanno Biber and Evelyn Breiteneder (Institute for Corpus Linguistics and Text Technology)**

500 000 000 tokens

In the following presentation an entire corpus of considerable size will be presented. The issue of representing a very large corpus in a format that offers only very limited space is paradigmatic for the general task of representing a language by just a small collection of texts and by just a small sample of the language. The AAC - Austrian Academy Corpus operated by the Institute for Corpus

Linguistics and Text Technology consists of more than 500 million running words and several thousands of texts representing a wide range of different text types have been collected, digitized and annotated. Among the sources, which cover manifold domains and genres, there are literary journals, newspapers, novels, dramas, poems, advertisements, essays, travel accounts, cookbooks, pamphlets, political speeches as well as plenty of scientific, legal, and religious texts, to name just a few. The AAC was founded several years ago and is a corpus research initiative concerned with establishing and exploring large electronic text corpora and conducting scholarly research in the fields of digital text corpora. The texts that have been integrated into the collections of the AAC are German language texts of important historical and cultural significance. The historical period covered by the corpus is ranging from the 1848 revolution to the fall of the iron curtain in 1989. In this period significant historical changes with remarkable influences on the language and the language use in the German speaking areas can be observed and examined. The AAC corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the linguistic and textual properties of these texts.

#### **Yuhua Chen (Pearson)**

To Wordlist or Not to Wordlist? The Dilemma and Challenges for Second Language Learning and Testing

Wordlists have been an important tool for second language learning. The Academic Word List (Coxhead, 2000), which was compiled from academic texts with corpus approaches, for example, has been widely used for EAP teaching and learning. However, in a general second language learning context, whether or not to provide a common graded wordlist, how to compile such a list if needed and what to be included in the list, has been controversial. This is particularly disputed in the field of language testing, where the fear is that coaching schools would simply ask the students to memorise the whole wordlist in order to pass the tests.

In this presentation, it will be argued that wordlists which can highlight the lexical items learners are most likely to encounter in the real world do have an indispensable role in the discourse of language learning and testing. Yet compilation of a wordlist should be collaborative work between language teachers, learning material publishers and exam boards instead of establishing a wordlist purely on the basis of word frequency in authentic texts. In addition, a good vocabulary syllabus should include the information more than just a list of individual words. The contextual information such as frequent word combinations, parts of speech, core meanings of ambiguous lexical items should also be covered. The challenges and the process in a collaborative project with an ESL publisher, language schools, and a language test developer will be discussed. The implications for providing such a vocabulary syllabus will also be addressed.

Coxhead, A. (2000). 'A new academic word list'. *TESOL quarterly*, 34 (2): 213-238.

#### **John Flowerdew (City University of Hong Kong)**

Small corpora, larger corpora and discourse analysis in the study of lexical cohesion

Acknowledging that there may be degrees of gradience, Bednarek (2009) argues for a three-pronged approach to corpus study: small-scale corpus analysis, large-scale corpus analysis, and manual analysis of individual texts. This paper exemplifies this approach by showing some ongoing work on lexical cohesion. Small corpora are typically about 100,000 words in size. Their advantage is that they allow the research to become very familiar with the corpus content and thus facilitate discourse analysis and that hand tagging of features that would not be susceptible to automated tagging can be done. Their limitation is that the frequency data they provide may not be representative. The study reported in this paper employs a large small corpus (some 600,000 words). The corpus has been tagged using a semi-automated process. Once tagging has been done, quantitative and qualitative data can be instantly retrieved. The large number of instances of given patterns revealed by the concordancer allows for analysis of individual examples in context (discourse analysis). Where

examples of minor categories are few in number, then reference corpora such as BNC and COCA can be used to verify if such patterns are significant or not. The procedures referred to above will be exemplified during the presentation, using the corpus in question, along with a concordancer.

Bednarek, M. (2009). 'Corpora and discourse: a three-pronged approach to analyzing linguistic data', HCSNet Summerfest '08, Cascadilla Proceedings Project, Somerville, MA, USA, 19-24.

**Stephen Jeaco (Xi'an Jiaotong-Liverpool University; University of Liverpool)**

Developing an intuitive screen design for a learner-centred concordancer

While learners and teachers may be using more materials based on patterns from corpora, the impact of corpus applications on self-study and in the classroom has not been as great as the shift in the academic or publishing fields. Indeed, it would seem that of the vast numbers of language teachers working around the world, only a relatively small number attempt to motivate learners to use concordancers, often finding that learning to navigate the user interfaces requires a deep understanding of linguistic jargon and that learners only experience a limited amount of success in being able to process snippets from authentic sentences which have been decontextualised.

Factors which may be holding teachers back from learning to use and teach corpus tools are:

- a) traditional KWIC concordance output is almost completely cut away from its context (Hunston, 2002).
- b) the amount of detail which concordances can provide to a learner can be confusing (Kennedy, 1998).
- c) interpretation of grammatical patterns is not easy (Gaskell & Cobb, 2004).
- d) exploration using carefully selected concordance lines seems to take too long (Thurstun, 1996)
- e) software is not usually designed specifically with learners in mind (Anthony, 2004)

This presentation of work in progress will outline some research into how manipulating visual elements of the screen can improve learner understanding of concordance output and improve performance of scanning and matching tasks. Elements such as context window size, window shape, the formation of windows, and font colour and size will be tested.

Anthony, L. (2004). *AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit*. Retrieved from [www.lextutor.ca](http://www.lextutor.ca)

Gaskell, D., & Cobb, T. (2004). 'Can learners use concordance feedback for writing errors?', *System*, 32(3), 301-319.

Hunston, S. (2002). *Corpora in applied linguistics*, Cambridge : Cambridge University Press, 2002.

Kennedy, G. D. (1998). *An introduction to corpus linguistics*, London : Longman, 1998.

Thurstun, J. (1996). 'Teaching the Vocabulary of Academic English via Concordances'.

**Rafidah Kamarudin (University of Birmingham)**

Influence of L1 on the use of phrasal verbs by Malaysian learners of English

Phrasal verb is one of the linguistics features which presents great difficulty to many language learners (Dagut and Laufer 1985; Hulstijn and Marchena 1989; Laufer and Eliasson 1993; Granger

1998; Liao and Fukuya 2004, Anna and Schmitt 2007). Various reasons have been highlighted to the problems faced by learners in understanding and using this language element including the nature of phrasal verb itself as well as cross linguistic factors. In this presentation, I will highlight some preliminary findings from my ongoing research into the use of phrasal verbs amongst Malaysian learners of English. Specifically, I will present an analysis of phrasal verbs with particle *up* and *down*, comparing usage in a native speaker corpus (the Bank of English) and EMAS, a corpus of English produced by Malaysian learners. My analysis so far has revealed that cross linguistic aspect, in particular learners' L1 (Malay) is one of the main reasons which influences learners' understanding and use of these phrasal verbs. Finally, I will conclude the presentation by discussing some pedagogical implications with respect to phrasal verbs particularly in Malaysian classrooms.

Anna, S., & Schmitt, N. (2007). "Native and nonnative use of multi-word vs. one-word verbs". *IRAL, International Review of Applied Linguistics in Language Teaching*, 45 (2): 119-140.

Dagut, Menachem & Batia Laufer. 1985. "Avoidance of phrasal verbs – a case for contrastive analysis." *Studies in second Language Acquisition* 7: 73-80.

Granger, Sylviane. 1998a. "Prefabricated patterns in advanced EFL writing: collocations and formulae." In: Cowie, Anthony P. (ed.). *Phraseology: theory, analysis and applications*. Oxford: Clarendon Press. 145-160.

Hulstijn, J. H., & Marchena, E. (1989). "Avoidance: Grammatical or semantic?" *Studies in Second Language Acquisition* 11: 241-252.

Laufer, Batia & Stig, Eliasson. 1993. "What causes avoidance in L2 learning? L1-L2 difference, L1-L2 similarity, or L2 complexity?" *Studies in Second Language Acquisition* 15: 35-48.

Liao, Yan & Yoshinori J. Fukuya. 2004. "Avoidance of phrasal verbs: the case of Chinese learners of English." *Language Learning* 54 (2): 193-226.

#### **Hui-Chuan Lu and Cheng-Yu Chang (National Cheng Kung University, Taiwan)**

##### Corpus-based Tool of Error Detection and Improvement Suggestion for Learners of Spanish

In this paper, we will present a developed tool of error detection and possible candidate suggestion based on two heterogeneous corpora in order to make up the lack of similar resources in learning Spanish and benefit foreign language learning in the future.

Our error detector is mainly based on an error-annotated learners' corpus (Corpus of Taiwanese Learners of Spanish, CATE (Lu, 2010)) and it can detect basic errors on several linguistic levels. The main idea is, given an error segment in an original text written by Taiwanese learners, there must exist a mapping in the revised one by native speakers of Spanish. Based on this mapping relation, we not only construct an error detector but also label the disorder permutation as an error for any incoming new text.

On the other hand, the improvement suggestions result from two complementary directions for learners. In the implementation, for each beginning of the error, we cluster the related phrases from revised texts in CATE. These clusters will be one of our two types of suggestion candidates. Besides, we also provide suggestion candidates from the corpus of online resources (e.g. wikipedia, Spanish news and academic articles). Once the segment of an incoming new article has been reported as an error, the top-n candidates from dual directions will appear simultaneously for the correction.

Finally, we will show experimental results of efficiency by undertaking an evaluation task from the

point of view of users including learners and educators.

Lu, Hui-Chuan. 2010. "An Annotated Taiwanese Learners' Corpus of Spanish, CATE." *Corpus Linguistics and Linguistic Theory*. 6(2), 297-300.

**Katarzyna Marszalek-Kowalewska (Adam Mickiewicz University, Poznan, Poland)**

Study of 'Sender's Phraseology': *Phrasemes* in the Modern Persian Language

'Sender's Phraseology' is the term developed by Polish linguist Wojciech Chlebda. According to this Researcher, all conventional, repeated multiword units belong to one group: *phrasemes*. Thus, the term phraseme is hyperonym for idioms, proverbs, citations etc. What is new and original here, is the fact that it embraces also the whole list of conventional, repeated multiword units that so far have not been counted as elements of phraseology.

The starting point of my research was the dissatisfaction with the application of existing theory (or rather theories) of phraseology in Persian lexicography. The lack of consistency leads to the situation that very often the same unit is described as idiom in one dictionary and as proverb in another. Moreover, dictionaries are full of old phraseological elements while modern, widely-used, conventional multiword units are simply bypassed.

Therefore, the aim of my research is to prepare an index of Persian *phrasemes* of such a kind - *phrasemes* that do really belong to discourse. To achieve it, I am compiling corpus of modern Persian texts. There are three kinds of texts that would go under scrutiny: newspaper articles, blogs and film subtitles. In order to extract *phrasemes*, special structural, semantic, metatextual, formal and frequency indicators will be used.

Dmitrij Dobrovol'skij, Elisabeth Piirainen. 2005. *Figurative Language Cross-cultural and Cross-linguistic Perspectives*. Amsterdam: Elsevier.

Ghaffari, Seif. 2005. *A dictionary of idioms Persian-English*. Teheran: Farhang Albarz.

Ghanbari, Abdollah. 2007. *Rahnama English-persian Dictionary of Proverbs*. Tehran: Rahnama Press.

Chlebda, Wojciech. 2003. *Elementy frazematyki: wprowadzenie do frazeologii nadawcy* [Basics of phrasemes study: introduction to sender's phraseology] . Łask: LEKSEM

Habibian, Simi. 2002. *1001 Persian-English Proverbs*. Maryland: IBEX Publishers.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: a corpus-based approach*. Oxford: Clarendon Press.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

**Coral Calvo Maturana (University of Granada, Spain)**

Jackie Kay's poetic discourse: A corpus stylistics research

This presentation deals with my on-going thesis research, a corpus stylistics study of Jackie Kay's poetic collection 'The Adoption Papers' (1991). This Scottish contemporary writer has always been studied from a feminist and post-colonial point of view; however, very little research, if any, has been done as regards her language. My thesis aims at describing 'The Adoption Papers' from a linguistic point of view, achieving interpretations which might agree or disagree with previous studies as well as adding new insights into the reading of her poetry. 'The Adoption Papers' is

framed in three poetic voices – adoptive mother, birth mother, and daughter – who tell us about the adoption process from their own perspective, according to their role. The voices of the three female characters are interwoven in the poem, being a slight typographical difference the only trace that distinguishes their individual discourse.

In this research, I make use of Corpus Linguistics research tools so as to carry out the study of the three different points of view expressed in the poem and their conceptualization of ‘maternity’ and ‘adoption’. With this aim in mind, I create three different corpora which include the words of each character. Through Wordsmith Tools 5.0. (Scott, 2008), I use three wordlists which specify words and their frequency in each poetic voice, which I later explore through concordances. In this way, I focus on the differences as regards modality and those words that in Kay’s poem are considered key in the process of adoption – such as ‘know’, ‘want’, ‘mother’, ‘blood’ and ‘colour’.

Adolphs, S. 2006. *Introducing Electronic Text Analysis. A practical guide for language and literary studies*. Oxon: Routledge.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Carretero, M. e Hidalgo, E. (2005) *For every man hath business and desire* or what modality can do for hamlet; In Martínez-Dueñas, JL., Pérez, C., McLaren, N. y Quereda, L. (2005) *Towards an understanding of the English Language: Past, Present and Future*. Granada: Universidad de Granada. (pp. 41-47)

Clear, J. 1993. “From Firth principles: Computational Tools for the Study of Collocation”. In Baker, M., Francis, G. y E. Tognini-Bonelli (eds.) 1993. *Text and technology in honour of John Sinclair*. Amsterdam y Philadelphia: Benjamins.

Corbett, J. and Anderson, W. 2009. *Exploring English with Online Corpora. An introduction*. Hampshire: Palgrave Macmillan.

Fischer-Starcke, B. 2009. ‘Keywords and frequent phrases of Jane Austen’s *Pride and Prejudice*. A corpus-stylistic analysis’. In *International Journal of Corpus Linguistics* 14:4, pp. 492-523.

Halliday, M.A.K. (1985), *An introduction to functional grammar*. Londres: Edward Arnold.

Kay, J. 1991. *The Adoption Papers*. Northumberland: Bloodaxe Books.

McIntyre, D. and Busse, B. 2010. *Language and Style*. Hampshire: Palgrave Macmillan.

Moreno Jaén, M. 2009. Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional. Thesis. Granada: University of Granada.

Palmer, F.R. (2001). *Mood and Modality*. Cambridge: Cambridge University Press.

Partington, A. 1998. *Patterns and Meanings. Using corpora for English Language Research and Teaching*. Series: Studies in Corpus Linguistics. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.

Quereda Rodríguez-Navarro, L. (1983). *A morphosyntactic study of the English Verb Phrase*. Granada: Universidad de Granada.

Rayson, P. 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>

Scott, M. y Tribble, C. 2006. *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Scott, M. 2008. *Wordsmith Tools 5.0*. Oxford: Oxford University Press

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

Stubbs, M. 2001. *Words and Phrases*. Corpus Studies of Lexical Semantics. Oxford: Blackwell.

**Danielle Noble (University of Newcastle, Australia), Brian Budgell (Canadian Memorial Chiropractic College, Canada), and Tracy Levett-Jones (University of Newcastle, Australia)**

A corpus linguistics study of the undergraduate nursing curriculum

Health-care has its own language, and each discipline has a distinctive lexicon which can be extensive and complex. However, to date, no studies have documented the language learning burden of nursing students, or how they acquire the language of their discipline. This presentation describes an ongoing project to characterize the language used in the Bachelor of Nursing program at the University of Newcastle, Australia, and to document the process of language acquisition.

A corpus of approximately four million words has been created from learning materials used in the undergraduate degree. Using the program WordSmith tools, keywords have been identified by comparison to a corpus of general English. The keywords are incorporated into questions on language assessments which will be administered over three years as students progress through the course, permitting mapping of vocabulary acquisition.

The study has identified recurring phrases and approximately 5000 keywords, which characterizes the nursing language at the University of Newcastle. A lexicon has been generated for each year of study, which each quartile of the 3 respective cohorts is expected to have mastered. Preliminary assessments have identified some inconsistencies between the nursing language used throughout the nursing degree and the communicative competence of these students.

This information can be used by others involved in course design to target the materials to the level of communicative competence of the students. Identification of the vocabulary will also better equip the nursing students with transitioning into the workforce.

Boyчук Duchscher, J.E., & Cowin, L.S. (2004). 'The experience of marginalization in new nursing graduates'. *Nursing Outlook* 52 (6), 289-296.

Budgell, B., Miyazaki, M., O'Brien, M., Perkins, R., & Tanaka, Y. (2007). 'Developing a corpus of the nursing literature: A pilot study'. *Japan Journal of Nursing Science*, 4, 21-25.

Corlett, J. (2000). 'The perceptions of nurse teachers, student nurses and preceptors of the theory-practice gap in nurse education'. *Nurse Education Today*, 20, 499-505.

Donnelly, T.T., McKiel, E., & Hwang, J. (2009). 'Factors influencing the performance of English as an

additional language nursing students: instructors' perspective'. *Nursing Inquiry*, 16 (3), 201- 211.

Krautshaid, L.C. (2008). 'Improving communication among healthcare providers: preparing student nurses for practice', *International Journal of Nursing Education Scholarship*, 5 (1), Article 40.

Parker, B., & Myrick, F. (2010). 'Transformative learning as a context for Human Patient Simulation', *Journal of Nursing Education*, 49 (6), 326-332.

The Australian Institute of Health and Welfare and the Australian Commission on safety and Quality in Health Care. (2007). Sentinel events in Australian public hospitals 2004-05. Canberra: AIHW. Retrieved from <http://www.aihw.gov.au>

The Centre for Biomedical and Health Linguistics (2009) The Test of English for bioMedical Purposes. Retrieved from <http://www.bmhlinguistics.org/joomla2/tebpm>

**Patricia Rodríguez-Inés (Universitat Autònoma de Barcelona)**

**SMS Language: Revelations from a Personal Corpus**

SMS language is known for its brevity, a characteristic due in part to restricted space for full words and lengthy texts, but also attributable to such language's playful, creative nature, an aspect of which is the use of expressions from foreign tongues (when available to the senders and receivers of messages). The study to be presented is based on an approximately 15,000-word personal corpus compiled by the author since 2000. While mainly a monolingual Spanish resource, the corpus also contains instances of English and Catalan. The corpus is not only analysed as a unit but is also divided into several subcorpora according to each message sender's relationship with the receiver (i.e. the author of the study). Although the size and representativeness of the corpus might be deemed problematic, the resource's strong points include its homogeneity, in that all the message senders are in one way or another related to a single receiver, and the fact that all the sources of its content are known. Analysis is expected to shed light on both the use of certain features of SMS language in a "multilingual" community and whether proximity between communicating parties affects the characteristics of the actual SMS language being used.

**Ute Römer and Matthew Brook O'Donnell (University of Michigan)**

**Looking at paragraphs in academic writing: Corpus and pedagogical perspectives**

Corpus linguistic methodology is beginning to move beyond a central focus upon words/phrases and their local contexts (through KWIC and collocational analysis) to consider lexical patterns at the level of larger discourse units. Hoey (2005) expands the notion of colligation to include the possibility that words and phrases have associations with specific text locations. Römer (2010) combines this insight with corpus-driven phraseology analysis to reveal the uneven distribution of n-grams and phrase-frames across academic book reviews.

Using a new corpus of student academic writing, the Michigan Corpus of Upper-level Student Papers (MICUSP), we created a phraseological database that records the position of each word in its sentence, paragraph and text and allows us to extract frequency and text-positional information for words, n-grams and phrase-frames. In this paper we focus on phraseological items where the majority of occurrences are in paragraph-initial (e.g. *in addition to*) or paragraph-final positions (e.g. *in the future*).

To gauge the educational validity of the results from our corpus analysis we constructed an online questionnaire presented to experienced EAP instructors. We asked them to rate items identified as paragraph-initial and paragraph-final as to whether they consider them worth teaching and might call attention to their functions as discourse structuring devices. Based on the instructors' feedback we then created revised lists of academic phrases which relate items to text structure. We believe

that these lists will be of relevance in the context of teaching and learning academic writing.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

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.

**Nawel Toumi (University of Reading)**

A comparative corpus based analysis of reflexive metadiscourse in L1 and FL research articles

Today, the overwhelming majority of impacted journals publish in English. While publishing in these journals is the aim of every scholar, who looks for promotion and other academic rewards, writing in English is difficult for Non Native English (NNE) academics, who have different writing conventions in their national disciplinary context. In this situation, it is useful to conduct specific comparative studies of the local and the native English socio-cultural contexts. Such analyses are insightful to NNE scholars, they inform about the necessary adjustments to be made in order to increase publication chances in international impacted journals. To serve this purpose, this study comparatively analyses reflexive metadiscourse use in research articles (RAs) written in English by native English and Tunisian researchers. The analysis investigates reflexive metadiscourse in a corpus of 100 RAs from 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 and Earth and Planetary Sciences RAs as samples of the hard sciences.

Basing my analysis on corpus linguistics (Hunston, 2002; McEnery, Xiao & Tono, 2006), this presentation explains the methodology used. The corpus for this study was individually compiled and processed manually and electronically. The procedure ranges from compiling the corpus, through coding the texts to extracting and comparing the metadiscursive instances.

This analysis will help identify divergence and similarities in the two cultural groups. The results will enable us to give more accurate guidance to Tunisian writers (or other FL writers) who wish to report their research in English for an international audience.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London/New York: Routledge.

**Atro Voutilainen and Krister Linden (Department of Modern Languages, University of Helsinki)**

Specifying a linguistic representation with a grammar definition corpus

Linguistic representations (LRs) constitute the specification (linguistic descriptors and application guidelines) for practical tasks such as corpus annotation. Usually, LRs are specified by applying an initial set of descriptors to limited text samples and refining the descriptor palette and associated documentation on the basis of issues that emerge during the initial analysis.

Descriptive grammars easily consist of thousands of pages, which suggests that natural language has a very large number of rules, some of a more general kind, the majority covering specific, low-frequency phenomena ("long tail").

LR design based on corpus samples runs the risk of specifying the LR on a limited, possibly random subset of natural language phenomena. When applying the resulting LR to larger corpora, linguistic phenomena not covered in the specification phase are likely to emerge, with harmful consequences such as problematic interpretation of undocumented analyses and/or unintuitive further adjustments to the LR.

We argue for using large descriptive grammars as a more systematic and wide-coverage basis for LR specification. The sample sentences in a descriptive grammar illustrate the syntactic structures of the language; they can also be used as text basis of a grammar definition corpus (GDC).

We describe completed work with a 16,000 sentence GDC, based on a large descriptive grammar of Finnish (Hakulinen et al, 2004), with focus on specifying a dependency syntactic representation and its application guidelines. We also outline benefits for treebanking, for design of formal language models for parsing purposes, and for design and evaluation of traditional descriptive grammars.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho 2004: *Iso suomen kielioppi [A large grammar of Finnish]*. Helsinki: Suomalaisen Kirjallisuuden Seura.

**Eros Zanchetta (University of Bologna), Marco Baroni (University of Trento), and Silvia Bernardini (University of Bologna)**

Corpora for the masses: the BootCaT front-end

This presentation introduces the BootCaT front-end, a graphical interface for the BootCaT toolkit (Baroni & Bernardini 2004).

The application implements an iterative procedure to bootstrap specialized corpora and terms from the web requiring only a list of seeds as input (a seed is a term that is expected to be typical of the domain of interest).

The front-end is a "wizard" that guides users through the BootCaT procedure allowing them to create a web corpus in a few minutes (theoretically, it should be possible to create a corpus in 6 minutes and 40 seconds).

Unlike other similar existing GUIs (such as JBootCaT, WeBoCa and WebBootCaT), BootCaT front-end is under active development (new features are added on a fairly regular basis) and is available for free. The application is cross-platform and runs on Windows, Mac and Linux.

New features currently being considered include: Unicode support, inclusion of non-HTML files (i.e. pdf, doc) in the corpus, exclusion of specific sites/domains, exclusion of documents that do not conform to specific licenses (i.e. Creative Commons).

#### Reference

M. Baroni and S. Bernardini (2004), "BootCaT: Bootstrapping corpora and terms from the web". Proceedings of LREC 2004.

# Colloquia

## Colloquium 1

**Michael Handford (Tokyo University), Almut Koester (University of Birmingham), Martin Warren (The Hong Kong Polytechnic University), and Svenja Adolphs and Kevin Harvey (University of Nottingham)**

Professional discourse and corpora

Format: The colloquium will include an introduction, four papers and time for audience discussion:

Introduction: 10 minutes

4 papers: 20 minutes each

Questions and discussion: 30 minutes (2-3 minutes after each paper and 15 minutes for general discussion)

One of the most exciting developments in corpus linguistics over the past decade has been the growth of smaller, specialized corpora of professional communication, whether spoken, written or multimodal. Such corpora allow the fine-grained analysis of linguistic items and features as they occur in their particular professional environments. This colloquium consists of four papers, two of which will focus on spoken professional discourse and two on written or electronically mediated discourse. The 'spoken' papers will show how corpora can be used in the analysis of two areas of communication more typically associated with discourse analysis than corpus linguistics, namely narratives and intercultural communication. Similarly, the 'written' papers discuss profession-specific language in financial services, engineering and health communication. The four papers will raise several issues of import in professional communication, such as face work, obligation, problem-solving, genre, culture and identity construction, and will show how corpora can provide a fresh perspective on discourse phenomena and contribute to a more robust description of professional discourse.

Paper 1: Signalling culture in professional spoken discourse

In intercultural communication, 'culture' is often used as a synonym for nationality. However, in professional intercultural communication, culture can also be framed in terms of national, professional, and institutional memberships, with language constituting particular face concerns, goals, practices and identities, and with nationality arguably being the least relevant variable. This paper analyses texts from a one million word corpus of authentic business meetings (CANBEC, Handford 2010), and other recordings\*, showing how selected statistically significant interpersonal features, such as pronouns and modals, may signal cultural differences in the unfolding discourse. The paper will also show how certain items, such as place deictics, may be 'key' in one professional culture, such as the construction industry, but not in others. The paper therefore attempts to show how charges of circularity in describing discourse as 'intercultural' can be addressed through the combination of corpora and contextual information.

\*The research referred to here was substantially supported by a grant from the Japanese Society for the Promotion of Science (Project no. 22520390).

Michael Handford, PhD  
English Language

Department of Civil Engineering  
Tokyo University

Paper 2: Hypothetical reported speech in business meetings

When spoken business discourse is compared to everyday communication, one of the top 15 keywords is *if*. While *if*, like many other keywords in business discourse, is multifunctional, for instance serving politeness functions as in the very frequent 4-word cluster *if you look at* (usually used to direct participants to some information or screen), a recurring pattern across many meetings is *if + pronoun + verb relating to speech*, for instance *if you say* or *if you were to say*. Closer examination of this pattern in longer extracts reveals that it often introduces direct reported speech, which usually begins with a marker, for instance *if you say*, “*Well...*” (Handford, 2010). However, this form of reported speech does not refer to actual events, but instead allows the participants to enter into hypothetical discussions, for example negotiating about price. According to Myers (1999), such hypothetical reported speech has been found in a variety of contexts, such as newspaper or radio reporting, and discussion groups. He also states that it may be particularly common in institutional settings, as it simultaneously manages to dramatise events (or imagined events), thereby making them more interesting, and also manages to create an impression of objectivity because of the implied ‘detachment’ of reported speech.

This talk will look at some typical examples of hypothetical reported speech in CANBEC (Cambridge and Nottingham Business English Corpus), and will also briefly introduce some teaching materials based on the research.

Almut Koester  
Department of English  
University of Birmingham  
a.j.koester@bham.ac.uk

Paper 3: Profession-specific phraseologies

This paper uses two profession-specific corpora compiled with the help of professionals in the financial services and the engineering sectors in Hong Kong. The corpora are the 7.3m Hong Kong Financial Services Corpus and the 9.2m-word Hong Kong Engineering Corpus (both are available at <http://www.engl.polyu.edu.hk/RCPCE/>).

The focus of this paper is on identifying phraseologies which are profession-specific in the sense of either being unique to, or more frequently found in, financial services or engineering texts. While there are indeed profession-specific phraseologies across the corpora as a whole which can be classified as register-specific, the study shows that there are profession-specific phraseologies in the two corpora which are genre-specific. The paper, therefore, also has implications for the ways in which studies of phraseology can be used to uncover the aboutness of register-specific corpora, genres-specific corpora and individual texts to help us better understand the language use of different professional communities.

Acknowledgements

The research described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5474/09H).

## Colloquia

Martin Warren  
Research Centre for Professional Communication in English  
English Department  
The Hong Kong Polytechnic University

### Paper 4: Advice in email health communication

This paper examines how corpus linguistic methods can be harnessed in analysing health communication. Focusing on electronic health discourse, we examine patterns and commonalities in the linguistic routines of young people when formulating health concerns to professionals online. The past decade has seen a proliferation of opportunities to use the internet for health related advice and information, and many new sites provide opportunities for users to construct identities, formulate problems and seek solutions concerning health related issues.

The paper reports on a study of a one million word corpus of emails sent to a UK-based website - the Teenage Health Freak website - a health forum tailored to the needs of young people to seek advice anonymously from health professionals. Our interrogation of the email data aims to demonstrate how corpus analysis is able to reveal micro patterns of communication (such as personal attitudes towards and beliefs about health and illness), as well as macro discursive patterns that potentially reflect broader cultural and economic trends in contemporary society: in particular the phenomenon of psychiatrization, the process whereby emotional complaints, once considered to fall within the scope of everyday experience, are perceived as pathological and hence susceptible to medical intervention.

Svenja Adolphs and Kevin Harvey  
School of English Studies  
University of Nottingham

Handford, M. (2010). *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Holmes, J. (2005). 'Story-telling at work: a complex discursive resource for integrating personal, professional and social identities'. *Discourse Studies* 7 (6): 671-700.

Koester, A. 2006. *Investigating Workplace Discourse*. London, Routledge.

Myers, G. (1999). Functions of reported speech in group discussions. *Applied Linguistics*, 20, 3: 376-401.

Orr, J. (1990). 'Sharing knowledge: celebrating identity: War stories and community memory among service technicians'. In D.S. Middleton and D. Edwards (Eds), *Collective Remembering*, 169-189

Vasquez, C. (2009). 'Examining the role of face work in a workplace complaint narrative'. *Narrative Inquiry* 19 (2): 259-279.

### Colloquium 2

**Tony Berber Sardinha (Sao Paulo Catholic University, Brazil), Cristina Mayer Acunzo (Sao Paulo Catholic University, Brazil), Marcia Veirano Pinto (Sao Paulo Catholic University, Brazil), Patricia Bértoli-Dutra (UniToledo University, Brazil), Renata Condi de Souza (São Paulo Catholic University, Brazil)**

Recent Perspectives on Multi-Dimensional Analysis

Long Colloquium (4 hours)

Title: Recent Perspectives on Multi-Dimensional Analysis

Rationale:

Multi-Dimensional Analysis (MDA) is a methodology introduced by Biber (1988, *inter alia*) that allows for the identification of underlying parameters of variation in corpus data, typically across different registers. Dimensions of variation, in turn, are patterns of cooccurrence of linguistic features underlying the registers of a language. An example of dimension of variation (for English) is 'Interaction versus Information' (Biber, 1988), which maps, along a scale, how different registers are more or less interactive or more or less informational. As the name implies, MDA typically reveals a multitude of dimensions, each representing a scale of variation. MDA makes extensive use of statistical techniques, notably Factor Analysis (FA), for the extraction of factors that are then interpreted both linguistically and situationally to indicate dimensions of variation. Previous research includes analysis of both whole languages and individual registers. Examples of the former are the descriptions of English (Biber, 1988; Crossley & Louwse, 2007; de Mönink, et al., 2003; Lee, 1999), Korean (Kim & Biber, 1994), Somali (Biber & Hared, 1994), Nukulaelae (Besnier, 1988), Gaelic (Lamb, 2008) and Spanish (Biber, et al., 2006; Parodi, 2007); examples of the latter are analyses of conversation (Biber, 2004), sitcoms (Quaglio, 2009) and research articles (Biber, et al., 1994).

In this colloquium, all papers address this basic research question: what are the underlying dimensions of variation in the corpus? The basic methodology is the following: (1) The corpus is tagged for selected features, using manual, automatic or semi-automatic procedures, the output is checked for accuracy, and corrections are made if necessary; (2) Counts are taken for each feature, which are then normalized, and standardized; (3) An initial FA is run, and the number of factors in the data is established; (4) A subsequent rotated FA is conducted for the specified number of factors; (5) Factors scores are computed for each text on each factor; (6) Factors are interpreted in terms of underlying dimensions of variation.

We will present recent studies that answer this question while both complementing and pushing the boundaries of MDA, both in a synchronic and a diachronic perspective. The first one is a full register variation analysis of Brazilian Portuguese, a language not before documented in the MDA literature. Other studies each present a detailed look at individual registers found in the media, an area where a noticeable gap exists in previous research, including movies, songs and magazine articles. Another one investigates variation in student writing. A final study looks at variation in metaphor use, exploring a major feature of human language that has not received attention in the MDA literature. In addition to part-of-speech, studies in the colloquium incorporate semantic analysis, an annotation level not explored in the MDA literature, as well as tagging for linguistic metaphor.

Audience discussion: 40 minutes

Dimensions of variation in Brazilian Portuguese

Tony Berber Sardinha, Cristina Mayer Acunzo; Carlos Kauffmann (Folha de S.Paulo News Organization, Brazil)

This is a synchronic study of register variation in Brazilian Portuguese. Portuguese is an important European language, the second largest Romance language, and the Brazilian variety accounts for 90% of its native speakers. To date, no MDA study has been carried out on Portuguese. For this investigation, a 9.5 million word sample of the 1-billion-word Brazilian Corpus was chosen,

comprising major spoken and written registers. It was tagged for POS using the Palavras tagger. The FA suggested a number of relevant factors, which will be presented and discussed in the presentation.

#### Dimensions of variation in Hollywood: the language of comedy and drama

Marcia Veirano Pinto

This is a diachronic study of representative American movies based on a 350K-word corpus of 16 comedies and 16 dramas from 1940 to 2009. Selected variables were computed to match Biber's Dimension 1, to see to what extent the language of Hollywood pictures approached spontaneous dialog at the interactive end of the scale. Results show that neither comedy nor drama resembled authentic dialog, but both are close to personal letters, spontaneous speeches, and interviews. No incremental trend in dimension scores was found over time. However, a cluster analysis identified groupings of movies, which are interpreted in terms of what they might suggest for an understanding of naturalness in film language.

#### Dimensions of variation in British and American pop music

Patricia Bértoli-Dutra

This is a diachronic study of UK and US pop music. The corpus is a 1.2-million word collection of over 6K lyrics recorded by 32 different artists from 1940 to 2009, including a variety of styles, from rock to punk to country. It was tagged for both part-of-speech, semantics (through a locally developed tagger), and lexical bundles (through specially designed software that mined the 1-trillion-word Google Corpus). FA indicated seven different dimensions. Close inspection of these dimensions showed how singers and bands and musical styles vary across time. Results show that the dimensions suggest a different view of styles, thus pointing to the market-driven nature of current style labels, which seem to cater for the interests of the music industry rather than reflect a concern for composition style.

#### Dimensions of variation in Time Magazine

Renata Condi de Souza

This is a diachronic study of Time magazine. A 1.3 million-word corpus of texts was collected, comprising more than 300 texts, spread over a period of 70 years, from the 1930's to the present, and tagged for part of speech. The FA indicated two factors, which were interpreted in terms of the communicative properties of the texts over time. The results indicated that the dimensions were unrelated to time of publication, suggesting that the style of Time did not change gradually over time, at least with respect to the dimensions.

#### Dimensions of variation in learner language

Denise Delegá and Tony Berber Sardinha (both São Paulo Catholic University, Brazil)

The corpus used for this study is the whole ICLE, the International Corpus of Learner Language, comprising 17 different nationalities. The texts were tagged for part of speech; frequencies for each feature were counted and normalized, and then mapped onto Biber's 1988 dimensions. The mean scores for each dimension were then compared across subcorpora, each representing a different

mother tongue background. The results suggested a wide range of variation in student writing across ICLE subcorpora, especially with respect to the dimension for argumentation, suggesting that writing an argumentative essay in EFL is a task that learners around the world accomplish in different ways.

#### Dimensions of variation in metaphor use

Tony Berber Sardinha (São Paulo Catholic University, Brazil)

Although metaphor is a key characteristic of ordinary language, so far it has been ignored in MDA research. The aim of this paper is to report on a synchronic study of variation in metaphor use across several registers of Brazilian Portuguese. A 50K word corpus was tagged for part of speech and manually annotated at the word level for metaphor, using an identification system based on features of both MIP (Metaphor Identification Procedure) and MIV (Metaphor Identification through Vehicles). The manual markup was complemented by automatic analysis carried out by the Metaphor Candidate Identifier, a software program designed to retrieve metaphors from corpora. Each metaphorically used word was tagged a variable based on source domain, target domain, vehicle morphology, metaphor type, vehicle word class, vehicle probability level, among other features. Other variables were added, such as for morphosyntax and metaphor density. The FA indicated three factors, which were interpreted in terms of their communicative purpose, showing differences across registers with respect to metaphor use. These will be reported in the presentation.

Besnier, N. (1988). 'The linguistic relationships of spoken and written Nukulaelae registers'. *Language*, 64, 707-736.

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

Biber, D. (2004) 'Conversation text types: A multi-dimensional analysis'. In Gérald Purnelle, Cédric Fairon, and Anne Dister (eds.), *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, 15-34. Louvain: Presses universitaires de Louvain.

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). 'Spoken and written register variation in Spanish: A multi-dimensional analysis'. *Corpora*, 1(1), 1-37.

Biber, D., Finegan, E., Oostdijk, N., & de Haan, P. (1994). 'Intra-textual variation within medical research articles', *Corpus-based research into language* (pp. 201-222). Amsterdam: Rodopi.

Biber, D., & Hared, M. (1994). 'Linguistic correlates of the transition to literacy in somali: Language adaptation in six press registers'. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 182-216). Oxford: Oxford University Press.

Crossley, S., & Louwse, M. M. (2007). 'Multi-dimensional register classification using bi-grams'. *International Journal of Corpus Linguistics*, 12(4), 453-478.

de Mönnink, I. M., Brom, N., & Oostdijk, N. H. J. (2003). 'Using the MF/MD method for automatic text classification'. In S. Granger & S. Petch Tyson (Eds.), *Extending the scope of corpus based research : New applications new challenges* (pp. 15-25). Amsterdam: Rodopi.

Kim, Y.-J., & Biber, D. (1994). 'A corpus-based analysis of register variation in Korean'. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 157-181). Oxford: Oxford University Press.

Lamb, W. (2008). *Scottish gaelic speech and writing : Register variation in an endangered language*. Belfast: Cló Ollscoil na Banríona.

Lee, D. Y. W. (1999). *Modelling variation in spoken and written language: The multi-dimensional approach revisited*. Tese de doutoramento, Department of Linguistics and Modern English Language, Lancaster University, UK.

Parodi, G. (2007). 'Variation across registers in Spanish: Exploring the El-Grial PUCV corpus'. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 11-53). London: Continuum.

Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. Natural conversation*. Amsterdam: John Benjamins.

# Workshops

## Workshop 1

**Alistair Baron (Lancaster University), Paul Rayson (Lancaster University), and Dawn Archer (University of Central Lancashire)**

Dealing with spelling variation in historical corpora: Using VARD to standardise spelling variants from the EmodE period

In papers presented at the previous corpus linguistic conferences (Archer et al. 2003; Rayson et al, 2005, 2007; Baron and Rayson, 2009), and elsewhere (Baron et al, 2009) a strong case has been made for the need to provide standardised versions of EmodE corpora alongside the original published editions. Without standardisation of spelling, commonly-applied corpus linguistics methods such as frequency and key words analysis and part-of-speech and semantic tagging are much less accurate. For example, it has been estimated that around 60% of word types and 35% of word tokens in a variety of corpus samples dating from 1500-1600 are historical variants. Although standardisation could be viewed as a spell checking or translation problem, it has been shown that existing tools such as Microsoft Word are unable to cope with the variety of variants in historical corpora. Our solution is the VARD (Variant Detector) software which allows corpus compilers and users to standardise spelling in corpora before the text is used for corpus analysis.

This session will be a two-hour hands-on workshop using the VARD software with a mixture of presentations and significant audience participation. We will begin with two short presentations each of 20 minutes. The first presentation will be by Anu Lehto from the University of Helsinki team who have recently published the corpus of Early Modern English Medical Texts (EMEMT, see Taavitsainen and Pahta, 2010); she will describe the team's experience of training and applying the VARD tool to develop a standardised version of the EMEMT corpus. The second presentation will be given by the workshop organisers and present an overview of the methods used in the VARD software and the accuracy of the tool.

The remainder of the workshop time (1 hour 20 minutes) will be devoted to hands-on exercises with the workshop participants using the software directly to manually standardise corpus samples. We will provide samples from Early English Books Online but workshop participants will be invited to bring their own corpora to test. As VARD can also be used to deal with other forms of spelling variation, such as in SMS (Tagg et al, 2010), participants would be welcome to bring texts from other sources. The hands-on component will include four main parts. First, familiarisation with the VARD user interface. Second, step-by-step training on the standardisation procedure itself. Third, the participants will independently standardise a selection of corpus samples. Finally, guidance will be given on how to tailor the linguistic resources and rules within VARD to improve the accuracy of the standardisation procedure.

By the end of the workshop, participants will understand how to use the VARD software to standardise spelling variants in EmodE corpora, how to export both original and standardised versions for use in other corpus linguistic software and how much training is required for their own corpora. Participants will be provided with copies of our previous studies on standardising historical corpora, a copy of the VARD software for academic use and a user manual. Since no computer labs are available at the conference venue, participants should bring their own laptops.

Archer, D., McEnery, T. Rayson, P. and Hardie, A. (2003). 'Developing an automated semantic analysis system for Early Modern English'. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper

number 16. UCREL, Lancaster University, pp. 22-31.

Baron, A. and Rayson, P. (2009). 'Automatic standardization of texts containing spelling variation, how much training data do you need?' In M. Mahlberg, V. González-Díaz and C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference, CL2009*, University of Liverpool, UK, 20-23 July 2009.

Baron, A., Rayson, P. and Archer, D. (2009). 'Word frequency and key word statistics in historical corpus linguistics'. *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.

Rayson, P., Archer, D. and Smith, N. (2005). 'WARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora'. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. *Proceedings from the Corpus Linguistics Conference Series on-line e-journal* 1 (1). ISSN 1747-9398.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). 'Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora'. In *Proceedings of Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Taavitsainen, I. and Pahta, P. (eds.) (2010). *Early Modern English Medical Texts: Corpus description and studies*, Benjamins, Amsterdam.

Tagg, C., Baron, A. and Rayson, P. (2010). "I didn't spel that wrong did i. Oops": Analysis and standardisation of SMS spelling variation. In *ICAME 31 Abstracts*, 108-109, Gießen, Germany.

## **Workshop 2**

**Andrew Dickinson (unaffiliated), Gill Francis (unaffiliated), and Jane Bradbury (King's Norton Boys' School, Birmingham)**

**Skylight and CALA: Classroom resources for teachers and language learners**

The workshop will have two main parts: the Skylight part and the CALA part. Each will consist of an introduction followed by hands-on audience participation.

Part 1: Skylight: an online classroom corpus resource.

Andrew Dickinson will first demonstrate the use of Skylight, a program that provides online corpus access for teachers and students in the classroom. Users select a particular corpus or corpora from the Skylight home page; for example, a teacher could direct students to a corpus of Shakespeare's plays, a corpus of Wikipedia text, or a large general web corpus.

The interface has been designed to be very clear and easy to interact with. Users can make all the familiar types of query, entering either a single word or a phrase, and viewing returned concordance lines or collocate lists. However, the range of operations is more limited, for the sake of simplicity; for example alphabetical sorting is limited to 1, 2, or 3 words to the right or left, and collocates are shown in order of significance, but without statistics.

Participants will then be free to explore Skylight themselves for a short time; we will have suggestions for words and phrases that anticipate the activities in Part 2. After this we move on to some specific activities, all of which are examples from our growing store of ideas to be developed both in CALA and in an as-yet-untitled collection of resources for ESOL and EAP classrooms. This leads into Part 2.

## Part 2: Corpus awareness activities

Gill Francis and Jane Bradbury will talk about CALA, the working title of a corpus-based workbook for teachers and students, which we first introduced in a paper at the BAAL/CUP seminar in Birmingham, June 2010. The activities are based on sets of concordance lines and/or collocation lists which focus on the use of language in different text-types. CALA is aimed at the upper secondary levels of UK schools, though we are also working on ideas for an ESOL resource along similar lines.

Gill will then present and invite participation in a range of activities, covering as many different types as time allows. Most of these are aimed at the ESOL classroom, and include items focusing on the uses of very common 'grammar' words like 'it' and 'that'. One activity, for example, shows how each of the more frequent uses of 'that' is identifiable by its lexical collocates. She also presents an activity addressing the issue of non-standard pronoun use, intended for a GCSE English class. This is entitled "Is it 'Sam and I' or 'me and Sam'?". The search results illustrate the flexibility of pronoun use in different types of text.

Finally, Jane will give examples of how Skylight can be used to aid teaching and learning in line with the new AQA GCSE English specifications. This will lead into hands-on activities in which participants can try a range of activities for themselves.

To conclude: through these resources, we aim to help get teachers and students into the swing of working with corpus evidence, and prepare them for developing their own ideas through the use of the Skylight program online on a day-to-day basis. Ideally we'd like to see Skylight as somewhere to turn to in class when any sort of query or observation about language comes up.

Gill Francis & Andrew Dickinson (2009) 'SKYLIGHT: A classroom corpus resource for language learners'. Article 414, *Proceedings of the Corpus Linguistics Conference CL2009*, online, University of Liverpool, 20-23 July 2009 (eds. Michaela Mahlberg, Victorina González-Díaz, Catherine Smith).

**Workshop 3****Alannah Fitzgerald (Durham University) and Shaoqun Wu (Waikato University)**

## Open Education Principles for Designing and Developing Digital Language Learning and Teaching Collections

This 1-2 hr conference workshop will be of interest to any practitioner of applied corpus linguistics who is interested in the development and design of language teaching and learning collections based on linguistic content derived from digital prose libraries and search engines such as Google and Yahoo.

A web-based demonstration of English language learning and teaching collections from the Flexible Language Acquisition Project (FLAX) from Waikato University's Greenstone Open Source Software Lab in New Zealand will be presented to familiarise the audience with technical aspects aimed at providing a simple and free concordancing user interface for non-specialist end users, namely language teachers and learners.

Learning support materials will also be demonstrated with easy to follow activities for exploiting the FLAX and British National Corpus collections based on a new Hefce-funded sibling project, TOE TOE (Technology for Open Education - Training with Open E-resources) for developing open educational resources (OER) for language teachers and learners in conjunction with two British universities, the Open University at Milton Keynes and Durham University.

The OER to be presented in this workshop build upon existing OER knowledge and practices that are

## Workshops

relevant to the English for Academic and Specific Purposes (EAP & ESP) communities within the Higher Education sector. This workshop will also be of relevance to general foreign language learning and teaching communities interested in corpora development with demonstrations and activities designed to provide dedicated learning support in the following areas:

1. Location and search of EAP-relevant OER located in repositories and networked communities on the web, using powerful aggregation tools to share, recommend and collaborate.
2. Materials and curriculum development guidelines for the EAP teaching community working within an OER framework, including: discovery, use, reuse and repurposing of existing digital materials. This will also include guidelines that promote an understanding of the different licensing standards for the use, promotion and sharing of OER.
3. Learning support for incorporating free web-based concordancers (e.g. FLAX and the British National Corpus) for data-driven language learning.

### **Workshop 4**

**Eva Hajicova (Charles University in Prague)**

Annotation of Discourse Relations in Large Corpora

AIM:

Annotation of discourse is commonly considered to be one of the next steps in corpus annotation and as such it has been the objective of several current research projects of corpus linguistics. However, with some exceptions such as the Penn Discourse Treebank (PDTB) or the discourse annotation scheme of the Prague Dependency Treebank (PDT), the state-of-the-art proposals deal with individual aspects or phenomena rather than with a systematic representation of discourse in annotation schemes. Discourse annotation of corpora is, of course, a rather complex task and there are many issues open for discussion, such as whether and how to develop a layer of discourse annotation "above" a syntactic layer, what features of the syntactic sentence annotation are useful or even indispensable for the annotation of discourse, what is the relation of the information structure of a sentence to discourse relations, what is the place of coreference annotation in the discourse annotation scheme etc. The purpose of the workshop is thus to create a forum for discussing these and other related issues.

ORGANIZATION:

Expected participants: The participation will be open to all interested scholars. The proposer has also addressed some prominent scholars from teams that already have a respectful experience in the development and application of discourse annotation scheme. For instance, the participation has been accepted by a prominent member of the PDTB team (Prof. Bonnie Webber, Univ. of Edinburgh) and Prof. Manfred Stede (Leipzig). Also the discourse and coreference analysis and annotation in the Prague Dependency Treebank will be introduced (Lucie Mladova, Charles University in Prague).

Duration of the workshop: 2 hours

### **Workshop 5**

**David Wible (National Central University, Taiwan) and Nai-Lung Tsao**

The StringNet Lexico-Grammatical Knowledgebase and its LexChecker Applications for Lexicography and Language Teaching

This workshop introduces a lexico-grammatical knowledgebase of English called StringNet and a suite of its applications called LexChecker.

A StringNet Navigator web interface supports keyword searches and navigation of the knowledgebase. StringNet (Wible and Tsao 2010) itself consists of hybrid n-grams, which, unlike traditional n-grams, can include part-of-speech (POS) grams. Thus, not only the string consider yourself lucky but also the patterns consider [prn rflx] lucky, consider yourself [adj], [verb] yourself lucky, inter alia. With StringNet Navigator, a click on any POS slot provides a pop-up showing the exact words attested in BNC in that slot in that pattern and their frequency in that slot. Each hybrid n-gram links to all BNC examples of it.

StringNet exploits the POS slots to capture subordinate and super-ordinate relations among and between hybrid n-grams. These are navigated by following `!parent!` or `!child!` links beside each hybrid n-gram listed in search results. For example, the two distinct n-grams consider yourself lucky and count yourself lucky are related by a common `!parent!` [verb] yourself lucky. Conversely, the hybrid n-gram consider yourself [adj] is the common parent of the `!children!` consider yourself lucky and consider yourself fortunate. This structure represents a dense relational dimension of new, navigable lexical knowledge.

LexChecker applications of StringNet include error detection and correction (Tsao and Wible 2009), also implemented as a `!query doctor!` a proxy for the common, risky practice of using Google searches for English error checking. LexChecker installed as a web-browser toolbar also can determine, for any string of text that a user mouse-selects in a webpage, whether that string is a frozen expression or an instance of a more general pattern and what that pattern is. LexChecker can also actively detect and highlight lexico-grammatical patterns in a webpage, unprompted. All tools and functions to be demonstrated are freely accessible.

Nai-Lung Tsao and David Wible. 'A Method for Unsupervised Lexical Error Detection and Correction', North American Association of Computational Linguistics (NAACL) Conference, Workshop on Innovative Use of NLP for Building Educational Applications, Boulder, Colorado, May 31-June 5, 2009.

David Wible and Nai-Lung Tsao 'StringNet as a Resource for the Discovery and Investigation of Linguistic Constructions', North American Association of Computational Linguistics (NAACL) Conference, Workshop on Extracting and Using Constructions in Computational Linguistics, Los Angeles, June 6, 2010.

# Posters

**Kirsten Ackermann (Pearson), Adam Kilgarriff (Lexical Computing Ltd.), David Tugwell (Lexical Computing Ltd.), and John H.A.L. de Jong (Pearson)**

The Pearson International Corpus of Academic English (PICAE)

Academic English is the register that people seeking to study at an institution where English is the language of instruction aim to master. While there are high-quality corpora of some variety of Academic English, most, like MICASE or BAWE, cover only a specific text type.

As part of the development programme for Pearson Test of English Academic, it was thus decided to compile an academic corpus that would comprise spoken and written data from five major English-speaking countries in order to support the objective to ground PTE Academic on an accurate rendition of the English that students will need to understand and produce in academic settings. PICAE includes curricular English as found in lectures, seminars, textbooks and journal papers. It also samples extracurricular English that students will encounter - from university administration to transcripts of broadcasts. The corpus has been cleaned, lemmatised and POS-tagged and is available for research.

**Nilanjana Banerji (Educational and Children's Division, Oxford University Press), Vineeta Gupta (Educational and Children's Division, Oxford University Press), Adam Kilgarriff (Lexical Computing Ltd.), David Tugwell (Lexical Computing Ltd.), and Kate Wild (Lexical Computing Ltd.)**

Oxford Children's Corpus: a corpus of writing for children

Language written for children is a text type of great interest. It needs to be lively and engaging, with simple vocabulary and structures, and is central to learning to read. We have created the Oxford Children's Corpus, a corpus of 28 million words, one third as published by OUP and two thirds from websites for children such as BBC Schools, History on the Net, Classic Readers, Nickelodeon, Neopets, Mrs. Mad, children's parts of the National Health Service, and British Museum sites. The material is fiction, textbooks, and informational and magazine material, as targeted at 7- to 14-year-olds. It is largely contemporary, though we have also included classics such as *The Wind in the Willows* as they are widely-read and influential. We have thoroughly cleaned, lemmatised and POS-tagged the input and loaded it into the Sketch Engine, a leading corpus tool, where it is available for research.

**Hanno Biber and Evelyn Breiteneder (Institute for Corpus Linguistics and Text Technology)**

Corpus Structures. The AAC Container as an example of organizing texts in a corpus

In the following paper corpora are regarded as complex containers of text. In order to be able to read, access and investigate the texts of a corpus in a digital environment, these containers must be constructed in a functional way. The corpus for which such a container is going to be developed is the AAC - Austrian Academy Corpus, a corpus of culturally and historically significant German language texts from the period between 1848 and 1989, which has been built at the Austrian Academy of Sciences in Vienna in the past years. More than 500 million running words of text have been scanned, converted into machine-readable text and annotated by means of XML-related standards. The AAC has collected thousands of literary objects and sources written by thousands of authors, representing an astonishing range of different text types. The texts that are systematically integrated into this large digital corpus have to be organized in a well structured way and provided with extensive metadata which is one important source of information about the texts. Specific metadata models and standards have been developed and become highly sophisticated methodologies for the description of corpora and their texts. The question of the structural organization of a large digital text corpus will be addressed from a new perspective of a large text corpus that has been built on the basis of specific thematic selection principles. In order to be able to investigate the texts and the language of the texts in such a large digital corpus of retrodigitized

texts the corpus needs to be structured in a specific way so that the results given by means of applying analytical corpus linguistics methods and tools are useful for textual scholars in several ways. In making digital model editions of texts which are part of the overall corpus, some questions could be answered whereby the tools have become powerful and critical reading instruments equipped with fully searchable databases of the texts, with various indexes, search tools for lexicographic and linguistic research as well as navigation aids in a functional graphic design interface providing the reader with a complex research environment to read, study and access the texts from various entry points. Corpus research methods have to take into account the various textual representations of historical periods and their developments so that corpus based analysis of the texts are of value for linguistic, literary, and cultural studies as well as related fields.

**Susie Caruso (University of Calabria, Italy)**

**A Corpus-based Metaphor Analysis of News Reports on the Middle East 'Road Map' Peace Process**

Metaphor is acknowledged as playing a central role in our understanding of how language, thought and discourse are structured and media discourse as having a persuasive and powerful role in reproducing discourse. The research presented in this paper aims to illustrate that the four main parties in the 'Road Map' peace process (US, UK, Israel, Palestine) viewed the events from different perspectives, thus producing different discourses. A corpus of news reports taken from four English language newspapers within the four societies involved are analysed using conceptual metaphor theory and critical metaphor analysis. This paper presents the findings of a corpus-based metaphor analysis looking at an ongoing peace process from different points of view. The results show that metaphor is a predominant feature of peace discourse, and the types of metaphors used to conceptualise peace are varied. Moreover, though the newspapers are reporting the same events, there are often different discourses that emerge.

**Gao Chao (University of International Business and Economics, Beijing, China)**

**Nativized features of English verbs in China's English Newspapers**

This poster is an attempt to explore nativized features of English verbs in China's English Newspapers and investigate whether nativized features are intelligible and acceptable.

The main method includes a corpus-based study and a brief questionnaire survey. The corpora adopted were China's English Newspaper Corpus (CCEN) and the newspaper part of the British National corpus (NBNC). The results show that: 1) the distribution of senses varied in CCEN and NBNC. Semantic specification/ broadening /variation can be found in CCEN. CCEN tends to use more recurrent collocations than NBNC and also tends to use the Verb + Noun + Preposition group. 2) Nativized English in China's context can be understood and accepted by both native and non-native speakers of English. However, native and non-native English speakers' interpretations of the verb collocations varied. English proficiency is the most important factor that affects intelligibility while intelligibility is the uppermost factor that affects acceptability.

**Pearl Shih-ping Cheng (Queen's University Belfast)**

**The Effectiveness of Using Corpora to Improve the Quality of Chinese-English Translation by Students**

This research aims to find out how to assist Taiwanese university students to improve the quality of Chinese-English Translation by using corpora as reference tools. The researcher is interested in finding out how students feel about the corpus-based translation course and how has it helped the quality of the students' Chinese-English translation? In which aspects of their translation skills have been improved through the training of using corpora? What happened during the process of implementing the corpus-based translation curriculum? How does the instructor feel about the implementation of the curriculum? An action research will be conducted by designing a corpus-based translation curriculum and implementing it with a group of student participants who are Taiwanese university students undertaking English-majors. These students will be taught by the researcher as the course instructor. The objective of the research is to find out how effective it is to

teach the students to use corpora as a tool of collocation references in improving the quality of their Chinese-English translation. The effectiveness of the curriculum will be evaluated by observing the students during the learning process, interviewing students, evaluating online student feedback on Moodle, pre-testing and post-testing on the students' collocation competency, and limited error analysis on student assignments for quality of translation.

Aston, G. (1999) 'Corpus use and learning to translate', *Textus*, 12(2), pp. 289-314.  
<http://www.sslmit.unibo.it/~guy/textus.htm>, last accessed 14/04/2011.

Baker, M. (1995) 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research', *Target*, 7(2), pp. 223-243.

Bowker, L. (1998) 'Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study', *Meta*, 43(4), pp. 631-651.

Carr, W. and Kemmis, S. (1986) *Becoming Critical: Education, Knowledge and Action Research*. London: Routledge Falmer.

Cohen, L., Manion, L. and Morrison, K. (2007, 6th Edition) *Research Methods in Education*. Routledge: London and New York.

Ferraresi, Adriano (2009) 'Google and beyond: web-as-corpus methodologies for translators', *Revista Tradumàtica* 7,  
<http://webs2002.uab.es/tradumatica/revista/num7/articles/04/04.pdf>, last accessed 23/04/2011.

Holmes, J. (1972 [1988]) 'The name and nature of translation studies', in Holmes, J. (Ed.) *Translated!: Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi, pp. 66-80.

Laviosa, S. (1998) 'The Corpus-Based Approach: A New Paradigm in Translation Studies', *Meta*, 43(4), pp. 474-479.

Laviosa, S. (2003) 'Corpora and Translation Studies', in Granger, S., Lerot, J., Petch-Tyson, S. (Eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. New York: Rodopi.

Liao, P. (2009) 'The Implications and Implementation of Communicative Translation Teaching', *Compilation and Translation Review*, 2 (2), pp. 65-91.

McCarthy, M. and R. Carter (2002 [2004]) 'This that and the other: Multi-word clusters in spoken English as visible patterns of interaction', Reprinted in McCarthy, M. (Ed) (2006) *Explorations in Corpus Linguistics*. New York: Cambridge University Press.

Munday, J. (2001) *Introducing Translation Studies: Theories and Applications*. London and New York: Routledge.

Possamai, V. (2009) 'Catalogue of Free-Access Translation-Related Corpora', *Revista Tradumàtica* 7,  
<http://webs2002.uab.es/tradumatica/revista/num7/articles/09/09.pdf>, last accessed 23/04/2011.

Reppen, R. and Simpson, R. (2002) 'Corpus Linguistics', in Schmitt, N. (Ed) *An Introduction to Applied Linguistics*. London: Arnold.

Robson, C. (2002, 2nd Edition) *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*. Oxford: Blackwell.

Rodríguez-Inés, P. (2009) 'Evaluating the process and not just the product when using corpora in translator education', in Beeby, A., Inés, P., Sánchez-Gijón, P. (Eds) *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam: John Benjamins.

Rodríguez-Inés, P. (2010) 'Electronic Corpora and Other Information and Communication Technology Tools: An Integrated Approach to Translation Teaching', *The Interpreter and Translator Trainer (ITT)*, 4(2), pp. 251-282.

Xiao, R. (2010) 'Can 'translation universals' survive in Mandarin? Idioms, word clusters and reformulation markers in translational Chinese'. *Proceedings of the Using Corpora in Contrastive and Translation Studies (UCCTS) 2010 conference*, Edge Hill University, 27-29 July 2010.

Xiao, R. and Yue, M. (2009) 'Using Corpora in Translation Studies: The State of the Art', in Baker, P. (Ed) *Contemporary Corpus Linguistics*. London and New York: Continuum.

Zanettin, F. (1998) 'Bilingual Comparable Corpora and the training of translators', *Meta*, 43(4), pp. 616-30.

Zanettin, F. (2002) 'DIY corpora: the WWW and the translator', in Maia, B., Haller, J. and Ulrych, M. (Eds) *Training the Language Services Provider for the New Millennium*. Oporto: AstraFlup, pp. 239-48.

#### **Telma de Lurdes São Bento Ferreira (Lexikos Cursos e Traduções)**

##### Corpus Linguistics and Authenticity in Textbooks: the case of Portuguese as a Foreign Language

This poster aims to show the results of an investigation of indicative aspects of authenticity in a textbook for the teaching of Portuguese as a Foreign Language, starting from the premise that even non-authentic texts may show characteristic elements of authenticity, and that these elements can be detected using the methods from Corpus Linguistics. As such, this analysis is based on Corpus Linguistics and the concepts of authenticity (Berber Sardinha, 2007; Nunan, 1989), idiomaticity (Sinclair, 1991), and lexical bundles (Biber et al., 1999).

We developed and applied a methodology for identification of authenticity in corpora that, in summary, is based on the lexico-grammatical analysis of the texts involved in a search for patterns that might provide evidence of authenticity (or otherwise) of teaching material, given that the frequency and quantity of the patterns found are expected to reflect the actual usage of language.

#### **Sheena Gardner (Coventry University)**

##### Perspectives on the disciplinary discourses of academic argument

Studies of student writing indicate disciplinary variation in the frequency of metadiscourse (Hyland 2005:57). For example, English student essays use 'less metadiscourse to explain the shape of the essay and less overt direction of the reader towards the arguments' than Sociology essays (Bruce 2010:162). As a central purpose of Essays in our classification is to develop an argument (Gardner and Nesi 2008; Nesi and Gardner forthcoming), searches for propositions introduced by framing metadiscourse can tell us more about the nature of academic argument in Sociology than in English.

This study aims to redress this imbalance and uncover disciplinary differences in propositional discourse of academic argument across Classics, English, Law, Philosophy and Sociology in 440 university student essays in the BAWE corpus. The findings from three analyses using SketchEngine

are compared: Clusters, key words with collocations and causal conjunctions. Although for English informative examples are those where key words collocate, this technique does not work equally across disciplines.

Bruce, I. (2010) 'Textual and discoursal resources used in the essay genre in sociology and English' *Journal of English for Academic Purposes*, 9, 153-166.

Gardner, S. and H. Nesi. (2008) 'A new categorisation of university student writing tasks' Paper presented at Language Issues in English-medium Universities: A Global Concern. University of Hong Kong, 18-20 June 2008. available at [www.coventry.ac.uk/BAWE](http://www.coventry.ac.uk/BAWE) (manuscript forthcoming in *Applied Linguistics*)

Nesi, H. and S. Gardner (forthcoming) *Genres Across Disciplines: Student Writing in Higher Education*. Cambridge Applied Linguistics Series. Eds. C. Chapelle and S. Hunston Cambridge University Press.

Hyland, K. (2005) *Metadiscourse*. London: Continuum.

#### **Mônica Holtz (Technische Universität Darmstadt)**

Lexico-grammatical properties of abstracts and research articles. A corpus-based study of scientific discourse from multiple disciplines

The main goal of this PhD-thesis is to gain insight into linguistic characteristics of abstracts in direct comparison with their respective RAs, finding differences and similarities between them. Abstracts themselves have been a "rather neglected social artefact of disciplinary life" (Hyland 2000: 83) and such a direct analysis comparing abstracts to their RAs has been largely disregarded by present linguistic research (Swales 1990: 181). For this reason, this study aims to systematically explore observable linguistic features at both lexical and grammatical levels, and evaluate them qualitatively and quantitatively. The investigation of linguistic variation between abstracts and their RAs across disciplines is another important goal of this study, since different communities may deploy linguistic features in discourse differently (Halliday & Martin 1993; Wignell et al. 1993; Wignell 1998). Finally, based on statistical evaluation of obtained data, this thesis aims to position abstracts and RAs in a broader linguistic context addressing the issue of the linguistic relationship between abstracts and RAs.

In order to investigate authentic usage of language, this study is performed over a corpus of abstracts and their respective RAs from scientific journal of several disciplines. The disciplines under study are computer science, linguistics, biology, and mechanical engineering. The design, processing, annotation and query of the corpus under study follows the current standards recommended by corpus linguistics methods (e.g., Biber 1993; McEnery & Wilson 2001; Sinclair 1991).

The criteria for the selection of linguistic features for the systematic quantitative analysis of the corpus follows not only preeminent work on corpus-based quantitative linguistic analysis (e.g., Biber 1988, 1995, 2006) but also is recursively based on primary data directly obtained from the corpus under study. Lastly, the evaluation of the results is substantiated by current and traditional statistical methods and practices (e.g., Baayen 2008; Gries 2006, 2007, 2008a,b, 2009).

This work, however, is not free from theoretical underpinnings. As Oesterreicher (2001: 1564) points out, theoretical assumptions are always present in any linguistic analysis. Desirable for this study is a linguistic theory that considers the functional variation of language and the context of situation in which this variation takes place, thereby delivering a systematic analytical framework for lexical and grammatical qualitative and quantitative analysis of linguistics features of this variation. Systemic Functional Linguistics (SFL; Halliday & Hasan 1989; Halliday 2004) fulfils these needs, since the

interest in functional variation of language is inherent in SFL (Halliday 2004: 33ff). Hence, SFL and corpus linguistics will be the theoretical and methodological underpinnings of this research.

This paper will show the results, discuss the methodology used, and present the outcomes of this thesis.

Baayen, R. H. (2008). *Analysing Linguistic Data. A Practical Introduction to Statistics using R*. London [u.a.]: Cambridge University Press.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). R'epresentativeness in corpus design'. *Literary and Linguistic Computing*, 8(4), 243-257.

Biber, D. (1995). *Dimensions of Register Variation. A cross linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2006). *Analytical procedures for the linguistic analyses*, vol. 23 of *Studies in Corpus Linguistics*, chap. Appendix A, (pp. 241-250). Amsterdam/Philadelphia: John Benjamins Publishing.

Gries, S. T. (2006). 'Exploring variability within and between corpora: some methodological considerations'. *Corpora*, 1(2), 109-151.

Gries, S. T. (2007). 'Cognitive linguistics and functional linguistics: Common assumptions and methods'. In S. T. Gries, & A. Stefanowitsch (Eds.) *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*, chap. Introduction, (pp. 1-17). Berlin, New York: Mouton de Gruyter.

Gries, S. T. (2008a). 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics*, 13(4), 403-437.

Gries, S. T. (2008b). *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, S. T. (2009). *Quantitative Corpus Linguistics with R. A practical introduction*. New York and London: Routledge.

Halliday, M. A. K. (2004). *An Introduction to Functional Grammar*. Revised by Matthiessen, C.M.I.M. London: Arnold, 3 ed.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. London, Washington D.C.: The Falmer Press.

Hyland, K. (2000). *Disciplinary Discourses. Social Interactions in Academic Writing*. Harlow: Pearson Education Limited, Longman. Michigan Classics Edition. The University of Michigan Press. 2004.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh Univ Pr: Edinburgh University Press, 2nd ed.

Oesterreicher, W. (2001). 'Historizität - Sprachvariation, Sprachverschiedenheit, Sprachwandel'. In M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.) *Language Typology and Language Universals / La typologie des langues et les universaux linguistiques / Sprachtypologie und sprachliche Universalien*, vol. 20.2, (pp. 1554-1595). Berlin, New York: Walter de Gruyter.

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

Swales, J. M. (1990). *Genre Analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Wignell, P. (1998). 'Technicality and abstraction in social science'. In J. R. Martin, & R. Veel (Eds.) *Reading science. Critical and functional perspectives on discourses of science*, (pp. 297-326). London [u.a.]: Routledge.

Wignell, P., Martin, J., & Eggins, S. (1993). 'The discourse of geography: Ordering and explaining the experiential world'. In M. A. K. Halliday, & J. Martin (Eds.) *Writing Science: Literacy and Discursive Power*, (pp. 136-165). London: University of Pittsburgh Press.

**Irina Iakovleva (Ulyanovsk State University)**

The corpus-based study of the Russian pseudosynonymous verb-preposition constructions: the CxG approach

This research focuses on three groups of Russian verb-preposition constructions: 1) constructions with "verbs of speech and thought": govorit' o YLoc/govorit' pro YAcc meaning 'to speak about Y'; 2) "verbs of sorrow" constructions: skuchat' o YLoc/skuchat' po YDat meaning 'to miss Y'; 3) constructions with "verbs of directed contact": bit' v YAcc/bit' po YDat meaning 'to bang against Y'. The constructions in every group used to be considered synonymous. Indeed, the interchangeability of the two constructions in every group is possible in overwhelming majority of examples, but according to Ruscorpora data, there are contexts that make this interchangeability impossible. The analysis of Ruscorpora data from the cognitive point of view shows that the semantic differences between the constructions in each group are caused by the restrictions which the construction as a whole imposes on the semantics of its components. In our case such restrictions are brought about by the semantic roles of the prepositions. As for "verbs of speech and thought" constructions, the prepositions o and pro are linked to different semantic roles. The preposition pro is linked to a complex role of theme and content and requires the agent argument in the position of the subject, while the preposition o is connected with the role of theme and doesn't impose such strict restrictions on the type of the subject: it may be agent, instrument or stimulus. As for constructions with "verbs of sorrow", the preposition o is connected with the role of theme and requires the agent argument in the position of the subject, while the preposition po is linked to the role of stimulus and requires the experiencer in the subject position. Thus, the o-construction implies a more controlled action than the po-construction. As regards the third group of the constructions, the preposition po here is connected with the role of patient and YDat refers to a kind of surface, while the preposition v is linked to the role of goal and YAcc refers to a kind of plane covering the cavity that is the place of destination. Thus, the constructions in the first group restrict the type of subject, those in the second group impose some restrictions on the type of the main verb, while the constructions in the third group differ in the type of object. The study of the acquisition of the two constructions in each group emphasises the importance of the semantic differences between them.

1. Fillmore, Charles J., Paul Kay & Catherine O'Connor. 'Regularity and Idiomaticity in Grammatical Constructions: The case of Let Alone'. *Language* 64. 1988.

2. Fried, Mirjam, Ostman, Jan-Ola. *Construction grammar: a thumbnail sketch. / Construction grammar in a cross-language perspective*. Amsterdam/Philadelphia, PA: Benjamins. 2004.

3. [http:// www.ruscorpora.ru](http://www.ruscorpora.ru)

**Irina Iakovleva (Ulyanovsk State University)**

The constructional approach to the historical corpus: the Casket Letters attributed to Mary, Queen of Scots

This study focuses on the constructional approach to the authenticity of the Casket Letters, the discovery of which helped Mary Stuart's forced abdication. The originals disappeared, so the Letters' authenticity can be judged only by the surviving copies which were repeatedly treated from different points of view: historical, psychological, etc. Analyzing the letters from the position of linguistics using the traditional statistical methods isn't supposed to be justified because of their comparatively small size, but the method of grammatical analysis can be successfully applied to texts of rather small size. Two corpuses were taken into consideration in this study of the Casket Letters' authenticity. The first one included four out of eight Casket Letters attributed to 24-year-old Mary Stuart. These four letters survived in the original French version. The texts were taken from [MacRobert 2002]. The second corpus consisted of 30 authentic Mary Stuart's letters which she wrote at the age of 18-26. The texts were taken from [Labanoff 1844]. The letters were analyzed on the basis of five parameters and resulted in revealing the following structures in the Casket Letters which weren't found in Mary Stuart's texts:

I. subject-predicate agreement: 1) the subject and the predicate don't agree in number and person; 2) the predicate agrees in number not with the subject but with the preceding complement; 3) the subject is used with the preposition *de*; 4) the absence of a finite verb predicate;

II. coinstantiation in the participial complement 'control' structures: 1) the implicit/explicit subject of the participial phrase doesn't coincide with that of the main clause; 2) the absence of the main clause on which the participial complement depends;

III. theta-criterion: Letters 3, 5, 6 have structures in which the predicate has two complements with one and the same semantic role;

IV. negative constructions: 1) the absence of the first *ni* in the construction *ni...ni...ne+V* or its substitution with *ou*; 2) the negative meaning of the construction *ne faire que*; 3) *V1...ni+V2*, where it is *V1* that has the negative meaning; 4) *...ne+ V1... V2+ou+ V3*, where verbs *V2* and *V3* have the negative meaning;

V. Constructions with conjunctions *que/qui*: 1) *...tell...qui +S+Pred*, where *qui* is used as a complement and refers to an inanimate object; 2) *que* is repeatedly used after the participial phrase.

Thus, though we can't faultlessly judge the Casket Letters' authenticity/forgery in the absence of the originals, there is no doubt that the Casket Letters possess some constructs passing on from one letter to another and having rather high frequency for texts of such a small size, but not occurring in Mary Stuart's authentic texts which outnumber the Casket Letters greatly.

1. Labanoff, A., ed. (1844). *Lettres, instructions et memoires de Marie Stuart*. Paris.

2. MacRobert, A. E. (2002). *Mary, Queen of Scots and the Casket Letters*. London/New York: Tauris.

**Steve Issitt (University of Birmingham)**

How an L2 learner corpus can identify areas of quantifiable improvement in students' written discourse

Empirical measures of development in writing can be difficult to establish, especially as the act of measurement itself involves so many complex and interconnected operations. In the wider context it may also be problematic to declare a programme of instruction as likely to result in improvement, if clear sets of criteria cannot be identified. In other words, it may behove researchers, teachers and course designers to be able to state with at least some confidence, what a course in English can produce in terms of learner outcomes. With this in mind, I have compiled a logitudinal learner corpus of approximately 80,000 words collected over a three-year period from preessional students at the University of Birmingham. This corpus consists of a series of essays which were completed at both the beginning and end of an intensive (20/15/10 and 6 week) EAP programme over three succesive years. The students were given the same title and the same length of time to answer. Papers were matched and the contents analysed according to a range of lexicogrammatical features that previous research (eg Biber et al 1999) has shown to be typical of written academic English. Features of the pre and post course scripts are then compared to a variety of reference corpora including BAWE and the written component of the BNC. A picture emerges of fairly consistent improvement which occurs over a relatively short time period. This finding tends to challenge previous research, which has argued that improvements cannot be detected for periods of instruction less than one year. I conclude that after periods of instruction ranging from as little as 6 weeks to 5 months on an intensive English for academic purposes programme, the students' writing shows greater structural flexibility and that this can be evidenced by reference to corpora.

Biber, D., Johansson, S., Leech, G., and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.

**Yuichiro Kobayashi (University of Osaka / Japan Society for the Promotion of Science)**

A Corpus-Based Approach to the Unnaturalness of Non-Native Metadiscourse

Why do native speakers often tell English writings written by non-native speakers from those written by native speakers? What is the "unnaturalness" found in the former? The purpose of this study is to extract characteristics of English argumentative essays written by Japanese EFL learners. This study draws on the Japanese component of the International Corpus of Learner English (ICLE) and the American component of the Louvain Corpus of Native English Essays (LOCNESS). The method is based on discriminant analysis with stepwise method, and the explanatory variables are the frequencies of ten semantic categories in Hyland's list of metadiscourse markers. With the accuracy of 90% over the entire set of corpus texts, the result of discriminant analysis shows that four categories (self-mentions, hedges, boosters, and frame markers) are very powerful discriminant index. This study illustrates how the four discorsal categories are used "unnaturally" in the non-native speakers' writings.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer-Verlag.

Crismore, A., Markkanen, R., & Steffensen, M. (1993). 'Metadiscourse in persuasive writing: A study of texts written by American and Finnish students'. *Written Communication*, 10, 37-71.

Granger, S. (Ed.) (1998). *Learner English on computer*. London: Longman.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. New York: Continuum.

Hyland, K., & Tse, P. (2004). 'Metadiscourse in academic writing: A reappraisal'. *Applied Linguistics*, 25, 156-177.

Kobayashi, Y. (2009). 'Profiling metadiscourse markers in native and non-native English'. *Lexicon*, 39, 1-17.

Vande Kopple, W. (1985). 'Some explanatory discourse on metadiscourse'. *College Composition and Communication*, 36, 82-93.

**Masumi Kojima (Gifu City Women's College)**

**An argument-based approach to validate S: A newly developed measure of lexical richness**

The purpose of this study is to validate S (Kojima, 2010), a newly developed approach to assess the lexical richness of written text produced by second language learners of English. Several measures have been proposed to examine the learners' productions regarding the frequency information of vocabularies used in the text, for example, Lexical Frequency Profile (Laufer & Nation, 1995) and P\_Lex (Meara & Bell, 2001). None of these measures, however, seem satisfactory (Daller & Xue, 2007; Meara, 2005). A new measure, which is called S, is calculated by randomly sampling words from a transcript to produce a curve of the cumulative ratio of lexes against the frequency levels for the empirical data. By adjusting the value of the parameter S, the computer programme finds the best fit between this empirical curve and the theoretical curves, which is calculated using a mathematical model. To develop a mathematical model of the curve, I use a logarithmic assumption of the frequency distribution of vocabulary (Zipf, 1935). The theoretical curves indeed fitted well with the empirical curves of my data, which indicates that the measure is a promising one. Kojima (2010) shows that the advantages of S over Lexical Frequency Profile and P\_Lex are that S can be used to evaluate shorter texts and can sensitively evaluate the texts of various levels of learners.

To further validate S, the present study employed an argument-based approach that was proposed by Kane (2006) and examined six hypotheses. For this purpose, written texts from 60 Japanese learners of English, half of them are in the intermediate levels and the others are in the high intermediate levels, are collected and analyzed on the basis of S. Each learner produced two pieces of written work. Corresponding data of English native speakers were also evaluated on the basis of S. The results indicated the following: (1) The fitness of the theoretical curves to the empirical curves of the real data is satisfactory, 2) inner consistency of S is high, 3) S is independent of the text length, 4) S values calculated from each subject's two essays correlate significantly, 5) S is moderately but significantly correlated with the scores of the Productive Vocabulary Levels Test (Laufer & Nation, 1999) and the scores of the Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001), 6) S can significantly discriminate between the two levels of learner groups and native English speakers. Those results strongly support S's validity as a new measure of lexical richness. S is hopefully applied to various study purposes such as exploring the relationship between learners' mental lexicons and their vocabulary use in their written or spoken productions.

Daller, H., & Xue, H. (2007). 'Lexical richness and the oral proficiency of Chinese EFL students'. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge University Press.

Kane, M.T. (2006). 'Validation'. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp.17-64). Westport, CT: American Council on Education and Praeger.

Kojima, M. (2010). 'Assessing S: A new measure of lexical richness'. Paper presented at the 2010 conference of the American Association for Applied Linguistics, Atlanta, GA.

Laufer, B., & Nation, P. (1995). 'Vocabulary Size and Use: Lexical Richness in L2 Written Production'. *Applied Linguistics*, 16 (3), 307-322.

Laufer, B., & Nation, P. (1999). 'A vocabulary-size test of controlled productive ability'. *Language Testing*, 16, 33-51.

Meara, P. (2005). 'Lexical Frequency Profiles: A Monte Carlo analysis'. *Applied Linguistics*, 26 (1), 32-47.

Meara, P., & Bell, H. (2001). 'P\_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts'. *Prospect*, 16 (3), 5-19.

**Cedric Krummes (Bangor University), Sylvia Jaworska (Queen Mary, University of London), and Astrid Ensslin (Bangor University)**

The Use of Discourse-Structuring Sequences by Advanced Learners of German: Corpus-Driven Investigations

This poster investigates discourse-structuring formulaic sequences produced by advanced learners of German and native speakers of German (cf. Wray 2002). The analysis is based on data collected for the 3-year AHRC/DFG project 'What's Hard in German?' (WHiG), which has been carried out by Bangor University and Humboldt-Universität zu Berlin since July 2009. Both data sets are part of the German error-annotated learner (parent) corpus FALKO (see Lüdeling 2008, Lüdeling et al. 2008, and Zeldes et al. 2008).

Following Stubbs (2002), our approach uses the *data* (and not a theoretical framework) as a starting point ('corpus-driven') in order to investigate the frequency and the use of formulaic sequences (also referred to as lexical bundles, or clusters). We aim to examine (1) the degree to which learners of German produce formulaic sequences compared to their native-speaker counterparts, (2) the types of sequences they over- and underuse compared to native speakers, especially with regard to discourse, reference and stance (cf. Chen and Baker 2010), and (3) the extent to which learners modify canonical structures, for instance, but not exclusively by L1-interference.

An in-depth understanding of which sequences are attested in the L1 corpus, but absent or modified (non-canonical) in the L2 corpus, will provide an impactful contribution to better teaching and learning materials for advanced learners of German.

Chen, Y.H. and Baker, P. (2010) 'Lexical Bundles in L1 and L2 Academic Writing', *Language Learning & Technology* 14 (2): 30–49.

Lüdeling, Anke. (2008). "Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora". In: Walter, Maik/Grommes, Patrick (eds.), *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung*. Tübingen: Niemeyer, 119-140.

Lüdeling, Anke/Doolittle, Seanna/Hirschmann, Hagen/Schmidt, Karin/Walter, Maik. (2008). "Das Lernerkorpus Falko". *Deutsch als Fremdsprache* (2), 67-73.

Stubbs, Michael. (2002). "Two quantitative methods of studying phraseology in English". In: *International Journal of Corpus Linguistics*, 7 (2), 215-244.

Zeldes, Amir, Lüdeling, Anke and Hagen Hirschmann. (2008). "What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data". 3rd Workshop on "Quantitative Investigations in Theoretical Linguistics 3" (QITL-3). Helsinki, Finland, 2-4 June 2008. Available from [http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes\\_et\\_al.pdf](http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes_et_al.pdf) (accessed 26 November 2010).

Wray, Alison. (2002). *Formulaic Language and the Lexicon*. Cambridge: CUP.

**Hui-Chuan Lu, Chien-Ting Ye, and Ya Jie Cheng (all National Cheng Kung University, Taiwan)****A Corpus-based Study of Phi-Feature in Literary Works**

In this study, we will integrate two disciplines on text-oriented study by applying corpus-based approach to the analysis of twentieth-century literary works, Gilman's *The Yellow Wall-Paper* and Herland by using WordSmith Tools, a lexical analyzer. With our primary study results showing that these two works features the morphological contrasts in number and gender of Nouns, Verbs, Pronouns and Adjectives, we reinforce the already existing results of literary studies in terms of narrative theory as well as those on the idea of new motherhood featuring a feminist tendency in Gilman's works. To find the pattern, we highlight the correlation function to bring out the contrast in both works, thus interpret the inconsistency and further find literary theories to explain the discrepancy. By adopting corpus-based approach on literary study, we intend not only to facilitate literary study with a more efficient method, but also to discover a diverse aspect of linguistic study.

Gilman, Charlotte Perkins. 1996. *The Yellow Wallpaper*. London: Little, Brown and Company.

Gilman, Charlotte Perkins. 1998. *Herland*. Mineola: Dover Publications.

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

**Katarzyna Marszalek-Kowalewska (Adam Mickiewicz University, Poznan, Poland)****English technical loanwords in Farsi - corpus-based studies**

Corpus linguistics is one of the most modern approaches in the study of language. It is constantly developing and expanding the scope of its study. Although, the most well-known corpus-based studies are of the British tradition, the field of corpus linguistics has been noticed and applied in various language communities.

The aim of this paper is to present the application of corpus research into the fields of lexicology, discourse and sociolinguistics. It focuses on assessing Iranian language policy (which is characterized by heavy linguistic purism) towards English lexical borrowings in Farsi. There are two studies that will be described. The first one was conducted for my MA thesis in 2009 "English borrowings in Farsi: a lexicographic and corpus-driven study of technical vocabulary". One of the purposes of this research was to study English loanwords in Farsi. Although there is quite rich monograph concerning Farsi borrowings in English (see for example Ahmad Mirfazalin's *Farsi words in English* from 2006) the literature on English borrowings in Farsi is hardly encountered. It seems that this area has not been investigated thoroughly so far. What is more, there are detailed studies concerning Arabic, Turkic, French or even Russian borrowings in Farsi, yet, English borrowings seem to be rather omitted in that discussion.

The tool used in that research was PLDB (Persian Linguistic Database) first, and actually the only constantly improved corpus of the Persian language. I would present a comparative corpus-driven study of certain English borrowings and their Farsi counterparts proposed by Iranian linguistic purists. These lexical borrowings belong to one semantic group - technology. The first study attempted to verify the differences in usage between certain English loanwords and their Farsi counterparts. This usage relates to collocations, register and frequency. By means of Persian Linguistic Database the question about the successfulness of the Iranian language policy towards this particular semantic group was addressed and it turned out that Iranian Language policy is not very successful. And that particular statement was the starting point for my second research.

Therefore, the hypothesis was: Iranian language policy is not particularly successful when it comes to English technical loanwords. In order to check if it is right or wrong I compiled my own corpus of Persian texts from 2010. The data comes from blogs, newspaper articles, film subtitles and song

lyrics. The results of the second study will be presented and then compared with the outcome of the first study. Certain similarities and differences will be also addressed. Finally, the question of the successfulness of the Iranian language policy towards English technical loanwords is going to be answered.

Assi Mostafa, Mohammad Haji Abdolhosseini. 2000. "Grammatical tagging of a Persian Corpus", *International Journal of Corpus Linguistics* 5(1) pp. 69-82.

Aryanpour Dictionary. 2009. (<http://www.aryanpour.com/>).

Bashiri, Iraj. 1994. "Russian loanwords in Persian and Tajiki", in Mehdi Marashi (ed.), 109-140.

Haugen, Einar. 1950. "The analysis of linguistic borrowing", *Language* 26: 210-231.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.  
Marashi, Mehdi. 1994. *Persian Studies in North America: Studies in honor of Mohammad Ali Jazayeri*. Bethesda: Iranbooks.

Mirfazalin, Ahmad. 2006. *Varzegane Farsi dar Englisi* [Farsi words in English]. Tehran: Farhangmoaser.

Moshiri, Mahshid. 1993. *Dictionnaire des Mots Européens en Persian* [The dictionary of European words in Persian]. Tehran: Alborz Publications.

Schuetze, Hinrich. (1995). "Distributional Part-of-Speech Tagging", *From Texts to Tags: Issues in Multilingual Language Analysis*. Online Proceedings of the ACL SIDGAT Workshop.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

The Academy of Persian Language and Literature. 2005a. *A collection of terms approved*. [No indication of publisher.].

Weinreich, Uriel. 1970. *Languages in contact: Findings and problems*. The Hague: Mouton.

Winford, Donald. 2003. *An introduction to contact linguistics*. Malden: Blackwell Publishers Inc.

**Peter Menke and Florian Hahn (both Universität Bielefeld, Germany)**

Corpus-driven development of a Gesture Typology based on the Bielefeld Speech and Gesture Alignment Corpus

We introduce the multi-modal Bielefeld Speech and Gesture Alignment (SaGA) corpus, and approaches to data-driven development and validation of a gesture typology.

The SaGA corpus contains 25 route-description dialogues. 280 minutes of video/audio material were transcribed (39.435 words). Additionally, boundaries of ~6000 gestures were marked, and ~5000 of these were annotated in detail and rated (Lücking et al. 2010).

We set up a typology of iconic gestures on the basis of one dialogue (~400 gestures). These types are grouped into classes according to their spatial complexity (Hahn and Rieser 2010, Rieser 2010).

Afterwards, we attempt to verify them by performing classification tasks. These results might aid in automated classifications of the remaining gestures. With features as input, gestures can be (1)

sorted into six classes, and (2) assigned the correct spatial dimensions.

Hahn, F. and Rieser, H. (2010): 'Explaining Speech- Gesture Alignment in MM Dialogue Using Gesture Typology'. In: P. Lupowski and M. Purver (eds.), *Aspects of Semantics and Pragmatics of Dialogue*. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue. Polish Society for Cognitive Science. Poznan 2010, 99-111.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). 'The Bielefeld Speech and Gesture Alignment Corpus (SaGA)'. In M. Kipp, J.-C. Martin, P. Paggio & D. Heylen (Eds.), *LREC 2010 Workshop: Multimodal Corpora—Advances in Capturing, Coding and Analyzing Multimodality*.

Rieser, H. (2010). 'On Factoring out a Gesture Typology from the Bielefeld Speech-And-Gesture-Alignment Corpus (SAGA)'. In: Kopp and Wachsmuth (Eds.), *Proceedings of GW 2009*. Springer, pp. 47-61.

#### **Katerina Mojziso** (Charles University in Prague)

##### Discourse functions of the cleft construction in English and Norwegian

The aim of the paper is to compare the use of the cleft construction in English and Norwegian from the point of view their discourse functions. Previous comparative studies of this kind analysed either direct translations between the two languages (Gundel 2002) or combined direct translations with the analysis of original texts (English-Swedish clefts, Johansson 2001). However, as these methods have some drawbacks (interference in the case of translation from a close language, or difficult comparability of two original texts), it seems fruitful to use the intermediary of another language which is typologically and genetically more distant. This study is hence based on a corpus of English and Norwegian texts translated from the Czech original. Czech, unlike English and Norwegian, is an inflectional language with a relatively free word order and it employs primarily other focusing devices than the cleft construction (focusing adverbs, word order etc.)

The study is confined to IT-clefts and focuses on their discourse functions. The classification of discourse functions analysed is based on Hasselgård (2004) and it is complemented by an account of additional irretrievable information based on Firbas (1995). The typology of discourse functions presented in this paper includes: contrast, selection, identification, summarizing function, topic launching and topic linking. All of the analysed discourse functions are closely connected with the functional sentence perspective (FSP). There are two main types of the cleft construction from the FSP point of view: the first type carries only one focus (in the main clause), the second type carries two foci (one in the main clause and one in the TH-clause). All discourse functions include clefts of both FSP types, but some functions are more likely to be represented by one of these types (e.g. the topic linking function is more likely to be attached to a two-foci cleft). The results show that all the identified discourse functions operate in both English and Norwegian and that the most frequent function in both languages is the contrastive function. However, there are also differences: the summarizing function and the identification (i.e. functions which are directed backwards in the text) are more common in Norwegian. The Norwegian cleft sentences of this type often correspond to English reversed WH-clefts. The topic launching and topic linking functions (directed forwards in the text), on the other hand, are more common among the English examples.

Firbas, J. (1995) "Retrievability span in functional sentence perspective" *Brno Studies in English*. Brno : Masarykova univerzita, pp. 17-45.

Gundel, J. K. (2002) "Information structure and the use of cleft sentences in English and Norwegian" In Hilde Hasselgård et al (eds) *Information Structure in Cross-Linguistic Perspective*, 113-128. Amsterdam: Rodopi.

Hasselgård, H. (2004) "Adverbials in it-cleft constructions" In *Language and Computers* 49, 195-212.

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

**Danielle Noble (University of Newcastle, Australia), Brian Budgell (Canadian Memorial Chiropractic College, Canada), and Tracy Levett-Jones (University of Newcastle, Australia)**

A corpus linguistics study of the undergraduate nursing curriculum

Health-care has its own language, and each discipline has a distinctive lexicon which can be extensive and complex. However, to date, no studies have documented the language learning burden of nursing students, or how they acquire the language of their discipline. This presentation describes an ongoing project to characterize the language used in the Bachelor of Nursing program at the University of Newcastle, Australia, and to document the process of language acquisition.

A corpus of approximately four million words has been created from learning materials used in the undergraduate degree. Using the program WordSmith tools, keywords have been identified by comparison to a corpus of general English. The keywords are incorporated into questions on language assessments which will be administered over three years as students progress through the course, permitting mapping of vocabulary acquisition.

The study has identified recurring phrases and approximately 5000 keywords, which characterizes the nursing language at the University of Newcastle. A lexicon has been generated for each year of study, which each quartile of the 3 respective cohorts is expected to have mastered. Preliminary assessments have identified some inconsistencies between the nursing language used throughout the nursing degree and the communicative competence of these students.

This information can be used by others involved in course design to target the materials to the level of communicative competence of the students. Identification of the vocabulary will also better equip the nursing students with transitioning into the workforce.

Boyчук Duchscher, J.E., & Cowin, L.S. (2004). 'The experience of marginalization in new nursing graduates'. *Nursing Outlook* 52 (6), 289-296.

Budgell, B., Miyazaki, M., O'Brien, M., Perkins, R., & Tanaka, Y. (2007). 'Developing a corpus of the nursing literature: A pilot study'. *Japan Journal of Nursing Science*, 4, 21-25.

Corlett, J. (2000). 'The perceptions of nurse teachers, student nurses and preceptors of the theory-practice gap in nurse education'. *Nurse Education Today*, 20, 499-505.

Donnelly, T.T., McKiel, E., & Hwang, J. (2009). 'Factors influencing the performance of English as an additional language nursing students: instructors' perspective'. *Nursing Inquiry*, 16(3), 201- 211.

Krautshaid, L.C. (2008). 'Improving communication among healthcare providers: preparing student nurses for practice', *International Journal of Nursing Education Scholarship*, 5 (1), Article 40.

Parker, B., & Myrick, F. (2010). 'Transformative learning as a context for Human Patient Simulation', *Journal of Nursing Education*, 49 (6), 326-332.

The Australian Institute of Health and Welfare and the Australian Commission on safety and Quality in Health Care. (2007). Sentinel events in Australian public hospitals 2004-05. Canberra: AIHW.

Retrieved from <http://www.aihw.gov.au>

The Centre for Biomedical and Health Linguistics (2009) The Test of English for bioMedical Purposes. Retrieved from <http://www.bmblinguistics.org/joomla2/tebpm>

**Matthew Brook O'Donnell (University of Michigan)**

The Adjusted Frequency List: A New Method to Extract Cluster-sensitive Frequency Lists from Corpora

The importance of multi-word chunks is well established in psycholinguistic (Erman & Warren 2000; Ellis 2003) and corpus linguistic (Sinclair 1991; O'Keeffe et al 2006) circles. However, many of our computational tools and methods still focus on individual words as the foundational units of analysis.

The adjusted frequency list is 'cluster sensitive' method for collecting n-grams, boosting the rank of larger word sequences ('on the other hand') and reduces the counts of their component parts ('on the', 'the other hand', etc.). Using sections of the BNCBaby corpus, a number of comparisons of unadjusted (standard) and adjusted frequency 1- to 5-gram lists are compared. For example, the top 10 items in a standard combined 1- to 5-gram frequency lists from BNCBaby-Demographic are: i, you, the, it, and, a, to, that, yeah, oh. The adjusted frequency method produces a cluster-sensitive list: i don't know, and, the, do you want, one two three, i don't think, of, in, two three four, a.

The simple method presented here, along with other more complex techniques that have been recently proposed (Gries & Mukherjee, 2010), demonstrates how corpus analysis continues to validate the importance of chunking in the investigation and description of language.

Ellis, N. C. (2003). 'Constructions, chunking, and connectionism: The emergence of second language structure'. In C. Doughty & Long, M.H. (Eds.), *Handbook of second language acquisition*. Oxford: Blackwell: 33-68.

Erman, B. & Warren, B. (2000). 'The idiom principle and the open choice principle'. *Text* 20 (1): 29-62.

Gries, S. & Mukherjee, J. (2010). 'Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes'. *International Journal of Corpus Linguistics* 15(4): 520-548.

O'Keeffe, A, McCarthy, M. & Carter, R. (2006). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

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

**Gill Philip (University of Macerata)**

Contexts of situation in text, in experience, and in the mind

Although Firth's (1957) "context of situation" is often referenced in corpus linguistics studies, no concerted effort has been made to map it onto corpus data. This poster, summarizing a major thread of the author's recent monograph (Philip 2011), does precisely that, demonstrating how Sinclair's (1996) unit of meaning can be regarded as the linguistic realization of the context of situation. Since the unit of meaning can admit variable elements (especially within semantic preference), the poster also makes recourse to the context of situation and its mental representation: in image schemata (Lakoff 1987) and metaphoremes (Cameron and Deignan 2006). Corpus evidence of these cognitive elaborations of the context of situation, in the form of creative variants of idiomatic phrases, is offered to support the main argument of the poster, which is that the context of situation is the place where meanings in the mind and meanings in language converge.

Cameron, L. and A. Deignan. 2006. 'The emergence of metaphor in discourse'. *Applied Linguistics* 27, 671-690.

Firth, J.R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Lakoff, G. 1987. 'Image metaphors'. *Metaphor and Symbolic Activity* 1 (3), 215-225.

Philip, G. 2011. *Colouring Meaning. Collocation and connotation in figurative language*. Amsterdam/Philadelphia: John Benjamins.

Sinclair, J.M. 1996. 'The search for units of meaning'. *Textus* 9, 75-106.

**Yufang Qian and Mengdi Ye (both Zhejiang University of Media and Communications)**

Discursive constructions of low-carbon economy in the UK, US and Chinese press

The term low-carbon economy first came into being in the UK Energy White Paper "Our energy future: creating a low carbon economy" in 2003. The background of the coinage of low-carbon economy was the severe challenge that the global warming had brought to the human existence and development. From the "Kyoto Protocol" to the "Bali Road Map", all countries in the world are doing their efforts to find a solution to climate change.

Scholars from various disciplines have conducted a plenty of valuable researches, but few of them are from linguistics perspective. This study focuses on the discursive constructions of low-carbon economy in the UK, USA and Chinese press by merging corpus methods and critical discourse analysis. The purpose of this study is to explore how media construct the discourses around low carbon economy in developing and developed countries and how discourse is socially shaped while socially shaping.

The newspapers in the UK and USA started to report low-carbon economy in 2004. One year later the term appeared in Chinese newspapers. How discourses of low-carbon economy are filtered via a wider range of phenomena including national interest and other political and social factors? How do these newspapers construct discourses around low-carbon economy? What are the value and policy orientation in each country? In order to answer these questions, this study focuses on the texts relating to low-carbon economy from major newspapers in China, UK and US.

By doing so, we will merge corpus techniques and CDA methods. Corpus techniques and CDA methods are able to complement each other. Corpus techniques such as keyness, clusters, collocation and concordance, can be applied to identify frequent and significant language patterns, which are indications of linguistic traces of particular discourses. While CDA methods examine the relevant political, economic, social and cultural process. This contextual analysis is invaluable, adding quantitative balance to the more quantitative analyses.

**Ida Ruffolo (University of Calabria, Italy)**

Identifying the perception of nature in travel promotion texts through a corpus-based discourse analysis

The increase of public concern about environmental issues has led to a surge of environmental appeals in advertisements (Hansen, 2002), in which advertisers make essential (mis-)use of the terms "nature" and "natural" (Harré et al., 1999). This research aims at revealing how nature and what is regarded as natural are described and employed by advertisers in travel promotion texts in order to attract ecotourists. In particular, by tracing the discourses of nature through a corpus-based discourse analysis, the study investigates the meaning of the terms "nature" and "natural" in order to understand whether or not their usage in tourism advertising is deceptive. To this end,

collocations of the terms “nature” and “natural” along with their concordances will be analysed to identify linguistic patterns in order to carry out a qualitative analysis (Baker, 2006).

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Hansen, A. (2002). Discourses of Nature in Advertising. *Communications: European Journal of Communication Research*, 27(4), 499-511.

Harré, R., Brockmeier, J., Mühlhäusler, P. (1999). *Greenspeak. A Study of Environmental Discourse*. Thousand Oaks, California: Sage Publications.

**Majdi Sawalha and Eric Atwell (both University of Leeds)**

Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus

The Arabic Internet Corpus is one of several large corpora collected for Translation Studies research at <http://corpus.leeds.ac.uk/internet.html> alongside Internet Corpora of English, Chinese, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian and Spanish (Sharoff, 2006). The Arabic Internet Corpus consists of about 176 million words. Initially it consisted of raw text, with no further processing such as lemmatization or part-of-speech tagging. In this paper we show how we added the lemma and root for each word.

Arabic is a morphologically rich and highly inflectional language. Hundreds of words can be derived from the same root; and a lemma can appear in the text in many different forms due to the glutination of clitics at the beginning and at the end of the word. Therefore, lemmatization and root extraction is necessary for search applications, to enable inflected forms of a word to be grouped together. We used the lemmatizing part of an Arabic morphological analyzer (Sawalha and Atwell, 2009, Sawalha and Atwell, 2010) to annotate the Arabic Internet Corpus words at two levels; the lemma and the root, illustrated in Figure 1. The morphological analyzer is relatively slow. In initial tests it processed 7 words per second, because the analyzer has to deal with orthographic issues, spell checking of the word's letters, short vowels and diacritics and the large dictionaries provided to the analyzer. An estimate execution time for lemmatizing the full Arabic Internet Corpus was 300 days using ordinary uni-processor machine.

To reduce the processing time of the whole task, we used the power of HPC (High Performance Computing). NGS (National Grid Services) aims to enable coherent electronic access for UK researchers to all computational and data based resources and facilities required to carry out their research, independent of resource or researcher location. We used the huge computational power of NGS to lemmatize the Arabic internet corpus and we gained massive reduction in execution time. We divided the Arabic Web Corpus into half-million-word files. Then we wrote a program that generates scripts to run the lemmatizer for each file in parallel. The output files are combined in one lemmatized Arabic Internet Corpus, comprising 176 million word-tokens, 2,412,983 word-types, 322,464 lemma-types, and 87,068 root-types.

By using the NGS we massively reduced the execution time of processing the 176M-word corpus to only 5 days. It might have been a few hours, had we been able to allocate enough CPUs to process all files strictly in parallel; NGS provides virtual parallel processing on a reduced set of CPUs. After the output files were combined into one lemmatized Arabic Web Corpus, 10 random samples, of 100 words each, were selected to evaluate the accuracy of the lemmatizer. For each sample, we computed the accuracy of the root and lemma analysis. We found that the average root and lemma accuracy was consistent across samples. The average root accuracy was about 81.20% and the average lemma accuracy was 80.80%; see Figure 2.

لعله	عل	علل		طويلا	طويل	طول	
أن	أن	أن	S OP_WORD	.	.	.	
يكون	كان	كون	STOP_WORD	.	.	.	
كابوسا	كابوس	كيس		طويلا	طويل	طول	
ويستفيق	يستفيق	فوق		،	،	،	
منه	منه	منه	STOP_WORD	وجلست	جلس	جلس	
على	على	على	STOP_WORD	البيوت	بيت	بيت	N_BP
الأشياء	أشياء	شيأ		ساكنة	ساكن	سكن	
الأليفة	أليف	ألف		،	،	،	
والطيبة	طيب	طيب		مطرقة	مطرق	طرق	
والحبيبة	حبيب	حب		،	،	،	
.	.	.		والمصايح	مصايح	صبح	
وامتد	امتد	مدد		الصفراء	صفراء	صفر	
الشارع	شارع	شرع		المقرورة	مقرور	قرر	
الضيق	ضيق	ضيق		تنزف	نزف	زفف	
				ضوءا	ضوء	ضوأ	

Figure 1: Sample of lemmatized sentence from the Arabic Internet Corpus

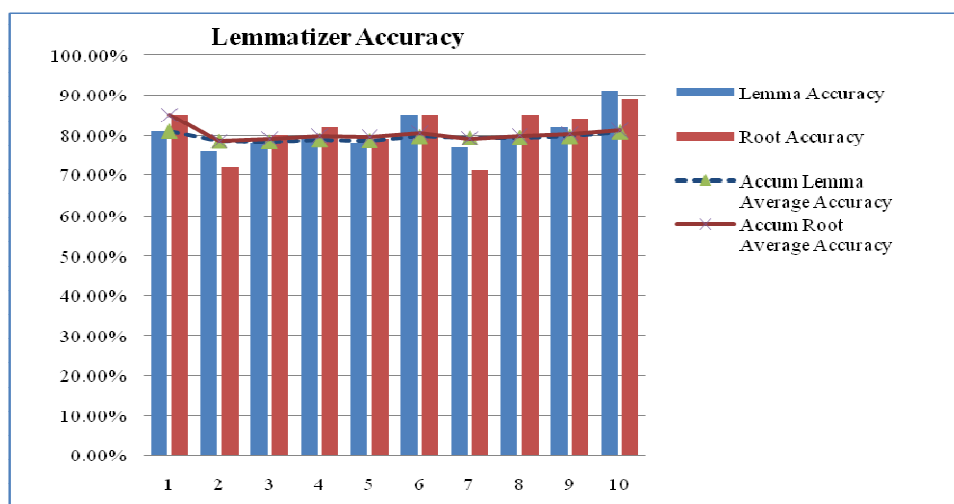


Figure 2: Lemma and root accuracy of the lemmatized Arabic internet corpus

Sawalha, Majdi; Atwell, Eric (2009). *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic*. in: Proceedings of the 5th International Corpus Linguistics Conference CL2009, 20-23 July 2009, Liverpool, UK.

Sawalha, Majdi; Atwell, Eric (2010). *Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text*. in: Proceedings of the Language Resource and Evaluation Conference LREC 2010, 17-23 May 2010, Valletta, Malta.

Sharoff, Serge (2006). Creating General-Purpose Corpus Using Automated Search Engine Queries. In M. Baroni and S. Bernardini (eds.). *WaCky! Working papers on the Web as Corpus*, pp. 63-98.

Bologna: GEDIT.
<b>Ulrike Schneider (University of Freiburg)</b>
Does, uh, Frequency Play a Role? - On the Placement of Pauses and Discourse Markers
The proposed poster is concerned with the influence of frequency-based and probabilistic factors on the placement of filled as well as unfilled pauses and discourse markers.
Recent studies have established links between frequently used sequences of elements and the places of occurrence of hesitations (Bybee 2007; Erman 2007; Fox, Maschler and Uhmman 2010; Kapatsinski 2005; Shriberg and Stolke 1996; Tily et al. 2009). My current project investigates on a theoretical level in how far probabilistic factors compete with other attractors of pauses and discourse markers, such as syntactic boundaries, and on a methodological level whether more complex probabilistic scores such as the Mutual Information Score (MI) and the upcoming Lexical Gravity G (Daudaravičius and Marcinkevičienė 2004; Gries and Mukherjee forthcoming), which takes type as well as token frequencies into account, are more accurate predictors than simpler measures such as transitional probability and bigram frequency. The proposed poster shows first results from a pilot study based on Switchboard NXT.
Bybee, Joan 2007. <i>Frequency of Use and the Organization of Language</i> . Oxford: OUP.
Daudaravičius, Vidas and Rūta Marcinkevičienė 2004. 'Gravity Counts for the boundaries of collocations'. <i>International Journal of Corpus Linguistics</i> 9 (2): 321-48.
Erman, Britt 2007. 'Cognitive processes as evidence of the idiom principle'. <i>International Journal of Corpus Linguistics</i> 12 (1): 25-53.
Fox, Barbara A., Yael Maschler and Susanne Uhmman 2010. 'A cross-linguistic study of self-repair: Evidence from English, German and Hebrew'. <i>Journal of Pragmatics</i> 42: 2487-505.
Gries, Stefan Th. and Joybrato Mukherjee forthcoming. 'Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes'. <i>International Journal of Corpus Linguistics</i> .
Kapatsinski, Vsevolod M. 2005. 'Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair'. <i>Berkeley Linguistics Society</i> 30: 481-92.
Shriberg, Elizabeth and Andreas Stolke 1996. 'Word predictability after hesitations: A corpus-based study'. <i>Proceedings of the International Conference on Spoken Language Processing</i> . (Vol. 3). Philadelphia, PA: 1868-71.
Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari and Joan Bresnan 2009. 'Syntactic probabilities affect pronunciation variation in spontaneous speech'. <i>Language and Cognition</i> 1 (2).
<b>Lucia Specia and Wilker Aziz (Research Group in Computational Linguistics, University of Wolverhampton, UK)</b>
Using a parallel corpus to learn semantic correspondences between two languages
Parallel corpora have been serving as the basis to extract different types of information for language processing applications. Common uses include extracting probabilistic word (Brown et al., 1990) or phrase dictionaries (Koehn et al., 2003) for statistical machine translation and multilingual information retrieval, and learning syntactic transfer rules for machine translation (Hoang and Koehn, 2010). In this paper we propose a method to exploit parallel corpora in order to learn legitimate shallow semantic correspondences between the two languages. An English-Spanish

parallel corpus is first processed to produce shallow semantic representations. We concentrate on semantic role labels, since these can be produced automatically with satisfactory accuracy. Semantic labels, along with base-phrase information, are used to generate shallow semantic 'trees'. These are then used to learn a model of the expected correspondences in term role labels between English sentences and translations into Spanish.

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: 'A statistical approach to machine translation'. *Computational Linguistics* 16(2), 79-85 (1990)

Hoang, H. and Koehn, P.: 'Improved translation with source syntax labels'. Workshop on Statistical Machine Translation and MetricsMATR, p. 409–417 (2010)

Koehn, P., Och, F.J., Marcu, D.: 'Statistical phrase-based translation'. In: *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. pp. 48-54 (2003)

**Anna Stermieri (University of Modena and Reggio Emilia)**

**Play and Performance: metadiscursive strategies in academic theatre reviews**

In the context of a study approaching the academic theatre review (ATR) as a genre, this analysis aims at identifying metadiscursive strategies associated with the realization of the communicative purposes of the ATR as a carrier genre (Tse and Hyland 2010), contributing to the delivery and promotion of knowledge. The study looks at how argument is built in ATRs, and what resources are considered appropriate by the community of discourse owning the genre.

The methodological framework includes concepts derived from discourse and genre analysis, and corpus linguistics, keeping in mind Hyland's works on metadiscourse (Hyland 2005). The corpus under analysis includes texts collected from 1991 and 2001 issues of 5 international academic journals.

The poster describes the metadiscursive patterns found, including both explicit appeals to a sense of community and specific forms of construction of identity. The patterns appear representative of the disciplinary and professional orientation of the genre.

Hyland, K. (2005) *Metadiscourse*. London/ New York: Continuum.

Tse, P., Hyland, K. (2010) 'Claiming a territory: Relative clauses in journal descriptions'. *Journal of Pragmatics*, 42, 1880-1889.

**Laurel Stvan (University of Texas at Arlington)**

**Sugar Makes You Sweet: Polysemy and Cultural Beliefs about Causation**

Lay terms in health discussions can show multiple senses not recognized as polysemes. Since polysemous terms have related senses—often in the same domain—contexts give fewer clues for disambiguation. Identifying conflated senses reveals misunderstandings reinforcing cultural beliefs about the causes of health and illness.

Methods: Track words expressing causation (causes, makes, becomes, turns into) collocated with 8 ambiguous terms (fat, stress, cholesterol, oil, hard, cold, hot, sugar) in the Corpus of American Discourses on Health.

1) "I am persuaded we are on a wrong scent in supposing moist, or cold air, the causes of that disorder we call a cold. (Ben Franklin, 1773, Green 2008)

2) "When you eat ice cream, the fat in the ice cream becomes fat in your body. So if you eat a lot of ice cream, you might become fat. If you don't, you're gonna stay skinny." (Little Miss Sunshine, 2006)

**Sylwia Twardo (University of Warsaw)**

Should we POS tag learner corpora?

Tagging errors in learner corpora is time-consuming and expensive. In order to make the task easier it would be useful to be able first to tag the parts of speech automatically. However, this may be feasible for learner corpora written by students who make few spelling mistakes. Removing texts with spelling mistakes from the analysis would blur the results. Hence it is worth-while to find a way of POS tagging corpora with spelling errors.

The present author POS tagged a learner corpus consisting of texts written by students at B1, B2 and C1 using CLAWS and analysed the errors in POS tagging. It was found that there were two kinds of them: those caused by the spelling mistakes in the texts and the errors made by the tagger and that both kinds of errors in tagging were systematic. The erroneous POS tags were tagged with the help of Spejd, a tool for rule based disambiguation.

Buczyński A., Wawer A., (2008). 'Shallow parsing in sentiment analysis of product reviews'. In: Proceedings of the Partial Parsing workshop at LREC 2008, pp. 14-18.

Buczyński A., Przepiórkowski A., (2008). 'Demo: An Open Source Tool for Partial Parsing and Morphosyntactic Disambiguation'. In: Proceedings of LREC 2008.

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>

**Tomio Uchida (Meisei University)**

Investigating Formulaic Use of *Of*-phrases by Non-advanced Learners: Findings from a Japanese EFL Learner Corpus

Learners of English with Japanese as a mother tongue tend to underuse prepositions regardless of their length of learning experience, and it is particularly true of the preposition *of* as shown by analyses of the JEFLL Corpus, a collection of writings by 10,000 Japanese high school students, and the ICLE. This study has focused upon expressions incorporating *of*-phrases and investigated the use of those by Japanese non-advanced learners in the JEFLL Corpus in detail. The items and frequencies of such sequences as prepositional phrases or phrasal verbs in the corpus and those in bilingual learner dictionaries were compared. Results showed that the variation both in lexical and collocation items was much less in learner English. It implies that underuse of the lexical item *of* by Japanese non-advanced learners results from a lack of competence in constructing semantic and/or structural units with *of* rather than lexical knowledge of individual words.

Tono, Y. (2007). *Nihonjin chuukousei ichimannin no eigo koopasu JEFLL Corpus*. Tokyo: Shogakukan.

**Zigrīda Vincēla (University of Latvia)**

Linguistic Variation in EFL Students' Virtual Texts of Different Registers

Linguistic variation received growing attention since the multidimensional analysis (MDA) has been proposed by Biber (1988) and applied (Xiao, McEnery, 2005) to investigate the texts of different registers. Biber advocates that it is accurate to view register differences as continuous dimension of variation distinguishing texts. EFL students of the University of Latvia experience difficulties in register-determined choice of linguistic features in different virtual writing situations. The purpose of the poster is to present linguistic variation found in a part-of-speech annotated corpus of EFL students' texts grouped according to communicative purpose. The texts were written during students' participation in project *IDEELS* (Intercultural Dynamics in European Education through

Online Simulation) and e-course *English Academic Writing* virtual writing activities. The results, obtained by MD framework (three linguistic dimensions), reveal variation among the texts the communicative purpose of which students had explored in detail due to their participation in the virtual writing activities.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University press

Xiao, Z., McEnery, A. (2005). 'Two approaches to genre analysis: three genres in modern American English'. *Journal of English Linguistics*, 33 (1), March, pp. 62-82.

**Natalia Vinogradova and Richard Moot (INRIA, LaBRI, University of Bordeaux)**

Temporal and Discourse analysis in Corpus of the French Narratives of 19th century

This work is realized in the context of the ITIPY project which aims to provide a solution for automatic extraction of the itineraries. As a database for this project a corpus of the French narratives of 19th century about the voyages in the Pyrenees is used. In (Loustau, 2008) we find already some results on the way of extracting itineraries. The problem is that not much attention is paid to the temporal analysis. In this context we draw our attention to the temporal analysis of the text (Boguraev and Ando, 2005; Muller and Denis, 2010). Moreover, we suppose that the temporal analysis itself is not always sufficient to reconstruct the correct order of events and we need also to look at the discourse structure of the text (Lascares and Asher, 1993). Let us look at the simple example:

Je partis en conséquence pour Bagnères-de-Luchon une seconde fois et <...> j'arrivai à Luchon le 17 juillet.

/I went consequently to Bagnères-de-Luchon for a second time and <...> I arrived to Luchon on the 17th of July./

In this paragraph one event (partis /went/) follows the other (arrivai /arrived/), forming a simple narration. We have in this case two verbs in Passé Simple (which corresponds more or less to English Past Indefinite). This time is not ambiguous in French and presents a finished action. Nevertheless, different discourse relations can exist between two phrases in Passé Simple (narration, elaboration, etc.) In this case we deal with a Narration (one event logically follows the other).

Now we shall take a more complex example:

Nous nous dirigeons donc vers Fost qui est un peu éloigné <...> Nous traversâmes la Garonne au dessus d'Estenos pour se diriger à droite sur Luchon, à gauche sur Saint Béat. La plaine que nous traversons est celle que nous avons vu la veille.

/So we are going to Fost that is a little bit far <...> We crossed Garonne over Estenos so that to go to Luchon to the right, to Saint Béat to the left. The plain we are crossing is the same we had seen the day before./

We have a sequence of Présent – Passé Simple – Présent in this paragraph (Présent corresponds to English Present Indefinite). The French Présent is ambiguous between a progressive, a narrative and a habitual reading (in our corpus, the habitual reading does not occur frequent enough for us to take it into consideration in our analysis). The choice between these readings influences the discourse relation between phrases. A sequence Progressive Présent - Passé Simple, as in the current example, excludes both the Background and Narration relation between the two events and leaves place to Elaboration. The second Présent (in the relative clause) is also progressive. Following precedent

discourse that gives us one more time an Elaboration. In the full paper, we are going to analyse, using the *Présent* and *Passé simple* as examples, the different possible discourse relations presupposed by one or the other temporal sequence.

Boguraev, Branimir and Rie Ando. 2005. 'TimeML -- compliant text analysis for temporal reasoning'. In Kaelbling, Leslie Pack and Fausto Giunchiglia, editors, Proceedings of IJCAI05.

Lascarides, Alex and Nicolas Asher. 1993. 'Temporal Interpretation, Discourse Relations, and Commonsense Entailment'. In : *Linguistics and Philosophy*, Springer, Vol. 16, p. 437-493.

Loustau, Pierre. 2008. *Interprétation automatique d'itinéraires dans des récits de voyages. D'une information géographique du syntagme à une information géographique du discours*. PhD Thesis, Université de Pau et des pays de l'Adour.

/Loustau, Pierre. 2008. *Automatic interpretation of itineraries in the corpus of french voyage narratives. From geographical information of clause to geographical information of discourse*. PhD Thesis, University of Pau./

Muller, Philippe and Pascal Denis. 2010. 'Comparison of different algebras for inducing the temporal structure of texts'. In Proceedings of the 24th International Conference on Computational Linguistics (Coling 2010).

**Salmah Yaakob (University of Birmingham)**

Disciplinary differences in social science and physical science BASE lecture introductions

Listening to lectures can be problematic due not only to the demands of real-time processing of information, but also the non-homogenous nature of lectures compared to other academic genres (Thompson 1994). Lecturers can mitigate the potential problem faced by listeners through the use of rhetorical functions to structure their lectures. Building on Thompson(1994)'s model of generic moves of lecture introductions, this study identifies the disciplinary differences in functions and sub-functions and their linguistic realisations in 20 social science lectures and 15 physical science lectures from the BASE corpus\*. This study uses a combination of genre analysis and corpus analysis to investigate the differences. The possible applications of findings include helping international students understand university lectures and encouraging lecturers to become more listener friendly.

Thompson, S. (1994). "Frameworks and contexts: A genre-based approach to analysing lecture introductions." *English for Specific Purposes* 13(2): 171-186.

\*The recordings and transcriptions used in this study come from the British Academic Spoken English (BASE) corpus. The 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.

**Eros Zanchetta (University of Bologna)**

Corpora for the masses: introducing CARCASS / CORPSE, an open-source client-server architecture for easy corpus access

This poster introduces a new corpus manager designed specifically to target language professionals and students with limited computer skills.

The Corpus Archive Search System (CARCASS) is a client that allows users to query corpora (including very large annotated ones) located on remote computers running the Corpus Server (CORPSE).

The client is a cross platform, user-friendly GUI that aims to replicate the full potential of the Corpus Query Language: queries can either be built with the help of a wizard or entered directly using the

## Posters

CQP syntax. The system supports positional (e.g. part-of-speech) as well as structural (e.g. text\_id) attributes. Unlike similar applications, the interface does not use a web browser.

The server component – built on top of OpenCWB (<http://cwb.sourceforge.net>) – manages user access and implements a sophisticated permission mechanism capable of restricting access to specific corpora and allocating computing time to users.

The system is free and open-source software.