Corpus Linguistics 2015
Lancaster University

Pre-conference workshop: "Academic corpora: Development, exploration and application"
Convenors: Paul Thompson (University of Birmingham) & Vander Viana (University of Stirling)
Date: 20 July 2015 (Monday)
Format: Full-day (9:30-17:30)
Venue: Charles Carter A17 (Please note that registration will take place in the George Fox foyer)

- 09:00-09:30: Registration
- 09:30-10:00: "PSLW-Corpus: A learner corpus in concept, content, and execution" (Heejung Kwon, Robert Scott Partridge & Shelley Staples)
- 10:00-10:30: "A corpus-based analysis of student talk in the university setting" (Eniko Csomay)
- 10:30-11:00: "Shell-nounhood in student writing: A multifaceted analysis of *ways* and *problems* in third-year undergraduate writing across four disciplines" (Miguel-Angel Benitez-Castro & Paul Thompson)
- 11:00-11:30: Coffee break
- 11:30-12:00: "A corpus-based analysis of reporting verbs in nonnative graduate student papers in Applied Linguistics" (Yasemin Bayyurt & Selahattin Yılmaz)
- 12:00-12:30: "Exploring disciplinary variation in English Language and Literature: A multidimensional analysis of PhD theses" (Vander Viana)
- 12:30-13:00: "Investigating speech act verbs to describe register variation" (Melanie Andresen)
- 13:00-14:00: Lunch
- 14:00-14:30: "Information density in scientific writing: Exploring the SciTex corpus" (Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis & Elke Teich)
- 14:30-15:00: "Use of topic models as a means to explore a corpus of academic English" (Dominik Vajn & Akira Murakami)
- 15:00-15:30: "Scholarly publication and writing development: Exploring interaction in research articles written by the same academic authors over time" (Suthee Ploisawaschai)
- 15:30-16:00: Coffee break
- 16:00-16:30: "Comparing the usefulness of academic word list and academic vocabulary list: A corpus-based critical evaluation" (David D. Qian)
- 16:30-17:00: "Towards a Norwegian academic vocabulary list" (Ruth Vatvedt Fjeld, Janne Bondi Johannessen, Kristin Hagen & Arash Saidi)
- 17:00-17:30: "Vocabulary test development with EAP specialized corpora" (Magdolna Lehmann)
- 19:00: Dinner

PSLW-Corpus: A learner corpus in concept, content, and execution

Heejung Kwon (Purdue University)
Robert Scott Partridge (Purdue University)
Shelley Staples (Purdue University)

The Purdue Second Language Writing Corpus (PSLW-Corpus) is our new learner corpus designed as a collaborative, research-community building project. Our corpus consists of multiple finished papers and drafts created by students in our introductory writing course for international students. Writings are grouped around 5 specific projects: narrative, proposal, literature review, interview report, and argumentative essay. PSLW-Corpus includes a robust dataset of de-identified information about the student writers' language ability, years in school, major/discipline, and country of origin among others. The breadth of data on the writers and the size of the corpus – nearly 2 million tokens and growing – make this corpus a powerful new scholarly tool; however, the importance extends beyond corpus linguistics. With many robust learner corpus building projects, it would have been easier to heed advice to use currently available resources rather than undertaking such a labor-intensive project (Reppen 2010). Nevertheless, despite the growing richness of English language learner corpus and documented construction processes (Bloch 2009, Hana et al 2014, Nessi & Gardner 2012, Römer & O'Donnell 2011), we have recently undertaken to build a learner corpus of our own.

We believe our learner corpus fills a gap in the currently available corpus in a variety of ways: first, it consists of multiple writing samples from novice language learners' writing process from a first year writing course covering a number of familiar academic writing genres. The population represents the current situation at many large U.S. academic institutions and our samples are tied to a host of background data for each writer. It is specifically focused description, not error analysis (Granger 2003, Hana et al 2014). We believe such a corpus allows for a wide range of studies involving language acquisition of a population that is representative of current international student contexts in large US universities. PSLW-Corpus also helps us develop more research by collecting and processing an enormous amount of student-generated writings that previously were underutilized in helping writing instructors' pedagogical improvement and enactment of the teacher-researcher persona (Mills 2002).

While PSLW-C was ultimately created and presented as a research tool, it was initially conceived of as a community-building project that highlights our desire to create more interactive, collaborative research amongst our fellow PhD students in disciplines concerned with issues of second language writing. Creating an infrastructural project as a centerpiece of our Community of Practice (Wenger, 1998) makes good sense in terms of research development and securing long-term sustainability for our growing monitor corpus, and yet we are aware that this is not the standard motivation behind corpus creation (Reppen 2010).

This presentation documents the process of construction highlighting the recursive steps involved: conceptualization; pitch to department chair, faculty lead, fellow instructors, IRB, and Registrar; reiterative attempts to secure storage space; standardizing processing of texts; granting access to fellow collaborators; undertaking initial research projects; and promoting and sustaining the corpus beyond our time in the program.

A corpus-based analysis of student talk in the university setting

Eniko Csomay (San Diego State University)

During the past three decades, a growing number of corpus-based studies have investigated language use in the academic setting. Researchers have compared the lexico-grammatical patterns used in the university context to other registers outside of the university such as face-to-face conversation and newspaper language (Biber et al. 2002) or provided comprehensive descriptions of spoken and written registers within the university (Biber and Conrad 2009). Other studies looked at only one register within the university setting, university classrooms, and found variation in language use associated with levels of instruction (Csomay 2002), disciplines (Csomay 2005), and participants (Csomay 2007) in that context.

This study investigates patterns of language use in student talk in two academic contexts. More specifically, similarities and differences are explored in the way students use particular lexico-grammatical patterns when they present in the classroom and when they present at a professional, academic symposium on campus. First, a 400,000 word corpus was compiled using two datasets: 1) 76 segments of student talk extracted from a large corpus of classroom discourse; and 2) 76 segments of student presentations recorded at a student research symposium. Both recordings were transcribed. Second, several computer programs were developed to process the transcribed texts to automatically identify grammatical patterns, and to count a number of lexico-grammatical features in the texts. Third, using a multi-dimensional analytical framework, the co-occurrence patterns of a large number of linguistic features were identified. Finally, student language use in the two contexts is compared based on the dimensions of linguistic variation in the university setting (Biber and Conrad 2009).

Preliminary results show differences in language use in the two situations identified, supporting earlier findings of variation in language use associated with aspects of the contextual differences. Pedagogical implications are discussed.

Shell-nounhood in student writing: A multifaceted analysis of *ways* and *problems* in third-year undergraduate writing across four disciplines

Miguel-Angel Benitez-Castro (University of Granada)
Paul Thompson (University of Birmingham)

Over the past forty years, nouns such as *fact*, *idea* or *warning* have received considerable attention from numerous approaches. These nouns share what Schmid (2000: 13) refers to as '[t]he property of shell-nounhood', which is associated with three assumptions, namely that shell nouns are abstract and semantically unspecific, that they refer to long discourse segments, and that they are preceded by deictically specific determiners (e.g. *the/this assumption*) and followed by complement clauses (e.g. *the fact that*).

The study of shell-noun phrases has chiefly focused on their use in academic writing. Despite the widespread interest in this genre, research to date has been primarily concerned with specific text types, particularly with published research articles (e.g. Gray & Cortes 2011; Cortes 2013; Flowerdew & Forest 2015). The widespread reliance on published academic prose contrasts with the more limited attention that shell-noun use has received in the study of discipline-specific student writing. This is apparent, for example, in the considerable attention that is devoted to only a small group of shell patterns (e.g. Aktas & Cortes 2008; Sing 2013). Research in this area is also primarily concerned with quantitative analyses. Qualitative insights, by contrast, are often restricted to small subsets of data and uses (e.g. Caldwell 2009). Finally, in cases where shell-noun use is explored in multidisciplinary corpora, the analysis tends to assess only L1-L2 or professional-novice writing differences, but fails to examine discipline-specific tendencies. Such tendencies, however, deserve more explicit attention, as they may bring to light the various rhetorical purposes for which students from various disciplines employ these nouns (e.g. Ravelli 2004).

This paper aims to address some of the aforementioned gaps through an in-depth multifaceted analysis of the use of two shell nouns (i.e. *way* and *problem*) in 135 texts (391,962 words) from the 6.5 million-word *British Academic Written English Corpus*. The sample under analysis comprises only texts produced by third-year English native students, as these are more likely to show a greater understanding of the ways of meaning of their discipline. The research focus here is on four of the most popular disciplines among UK undergraduate students (*Higher Education Statistics Agency*): Biological Sciences and Engineering, from the natural sciences, and Business and Sociology, from the social sciences. As regards the two lemmas explored in this paper, *way* and *problem* feature among the ten most frequent shell nouns in Flowerdew & Forest's (2015: 86) corpus of academic journals, textbooks and lectures. This paper, therefore, sets out to examine the extent to which two highly frequent shell nouns in professional academic discourse are more or less primed for particular disciplinary uses in final-year undergraduate native student writing.

The overall evidence studied here consists of 521 concordances tagged on the basis of eight variables spanning formal, syntactic, semantic and textual features of shell-noun phrases: i) text type (e.g. essay), ii) formal structure (e.g. definite determiner + head noun), iii) semantic structure (e.g. Epithet + Thing), iv) syntactic function (e.g. Direct Object), v) participant type (e.g. Attribute), vi) Theme/Rheme, vii) direction of encapsulation (e.g. intersentential cataphora) and viii) antecedent type (e.g. sentence). For the sake of higher descriptive detail, the analytical framework rests on Quirk et al.'s (1985) grammar for the structural variables, and on Systemic Functional Grammar (Halliday & Matthiessen 2004) for the semantic and textual ones.

A corpus-based analysis of reporting verbs in nonnative graduate student papers in Applied Linguistics

Yasemin Bayyurt (Boğaziçi University)
Selahattin Yılmaz (Yıldız Technical University)

As one of the building blocks of academic discourse, citing others' work enables intertextual connections to build a successful argument, and by citing others' work, writers create not only a link between their research and previous related studies, but also their place in the related field of study (Swales, 1990; Hyland, 1999; Charles, 2006). Of the broad array of perspectives from which the citation practices have been researched, reporting verbs are one of the most widely researched and challenging features of citation, especially for novice and nonnative academic writers. Although several studies shed light on the use of reporting verbs (Thompson & Yiyun, 1991, Thomas & Hawes, 1994; Hyland, 1999, 2000, 2002), all these studies focused on expert writing. However, while the complexities in selecting appropriate reporting verbs in an academic discourse could pose challenges for novice writers, this challenge becomes even greater with being nonnative as they have to be both familiar with academic conventions of the target discourse community and proficient in the language that they write in (Flowerdew, 1999; Lang, 2004; Bloch, 2010). Therefore, investigation of the academic writing of the graduate students is of major importance in exploring L2 academic writing practices and improving English for Academic Purposes (EAP) pedagogy (Gilquin et al., 2007). Furthermore, to the best knowledge of the researchers, there is no study that addresses the issue in the Turkish context. In line with the literature and the research gap, the current study aims to explore use of reporting verbs in 50 research papers written by Turkish graduate students enrolled in a graduate program of applied linguistics at a state university in Turkey. The papers were written as partial fulfillment of the requirements for three doctoral courses Issues in Foreign Language Education Planning, World Englishes, and Program Evaluation in Language Education. The corpus of the study is composed of only the literature review sections, in accordance with Soler-Monreal & Gil-Salom (2011) that other-reference is most typically found in these sections. For the initial analyses of 6 papers, Antconc 3.4.3 was used to extract concordance lists of main verbs in citations by searching dates in brackets, as well as the words given in the framework of functions of reporting verbs by Thomson and Yiyun (1991) and Hyland's (1999). All extracted sentences were checked to assure that they function as other-reference, and can be grouped under the categories offered by the frameworks. Results were compared to similar studies that analyzed research articles (Hyland, 1999) and student papers (Ädel & Garretson, 2006). The pilot study indicated that Turkish graduate EFL writers tend to overuse reporting verbs to neutrally report the research procedures and outcomes in other sources and underuse the discourse acts and cognition acts. Since soft disciplines are known to be highly discursive and reliant on context (Hyland, 1999). This could be a seen as a weakness of the students' writing. Furthermore, evaluative function of reporting verbs was also found to be limited. Finally implications for further research and EAP writing pedagogy in relation to the teaching of reporting practices were discussed.

Exploring disciplinary variation in English Language and Literature: A multidimensional analysis of PhD theses

Vander Viana (University of Stirling)

The present paper aims at examining disciplinary variation between two under-researched academic fields, namely, English Language and Literature. Although both fields are well-known for their textual analyses, Language and Literature researchers have not fully engaged in investigations of their own textual products (MacDonald, 1990). This imbalance is confirmed by a review of the studies in Discourse Analysis and English for Academic Purposes, which have largely focused on the texts produced and/or circulated in medical, scientific and technological contexts (Biber et al., 2002; Flowerdew, 2002).

Disciplinary variation is here approached through an analysis of PhD theses, a second gap to be filled with the present research. Although written academic registers have received more attention from researchers than spoken ones (Biber, 2006, 2010; Biber, Conrad, Reppen, Byrd, & Helt, 2002; Flowerdew, 2002), a focus on an analysis of journal papers is easily noticeable (Biber, 2006; Biber et al., 2002; Dudley-Evans, 1999; Thompson, 2000). It is true that some general publications in genre and academic discourse (e.g. Hyland, 2009; Swales, 1990, 2004) dedicate a section/chapter to doctoral theses, and that some studies examine this genre in a more detailed way (e.g. Bunton, 1998; Thompson, 2001). However, as has been argued elsewhere (Bunton, 2002; Dudley-Evans, 1999; Paltridge, 2002; Shaw, 1992; Swales, 1990; Thompson, 2000), PhD theses still require extensive exploration from linguists.

In order to fulfil the research aims described above, the present investigation abides by the principles in Corpus Linguistics. Two specialized corpora were compiled to represent PhD theses in English Language and Literature. The corpora contain 40 theses produced in the United Kingdom between 2000 and 2009, totalling 2.9 million words. They were probed in a corpus-based, top-down way. The analytical procedure followed Biber's (1988) multidimensional model for the study of register variation, which has been productively adopted in a number of other investigations (e.g. Atkinson, 2001; Biber, 1995; Biber & Finegan, 2001a, 2001b; Conrad, 1996, 2001; Nesi & Gardner, 2012).

The results show both convergence and divergence in the dimensions of variation in English Language and Literature PhD theses. Convergence is found in the observation of high levels of informativeness and explicitness in theses from both fields. Language and Literature doctoral graduates share the same need of meeting the standard patterns of academic practice, and of avoiding contextual references to the here and now of when the research was conducted and/or when the thesis was written. However, the analysis also reveals that Language and Literature PhD theses inhabit discursive worlds of their own. Language theses are less narrative, more argumentative and more abstract than Literature ones. Given that "the need for the understanding of language is above all an essential aspect of understanding human ways of being, feeling, thinking and doing" (Hasan, 2011, p. xiii), this investigation on academic discourse opens new vistas to the fields of English Language and Literature and claims for the centrality of disciplinary variation in the analysis of academic discourse.

Investigating speech act verbs to describe register variation

Melanie Andresen (University of Hamburg)

Academic language and everyday language share a certain part of the lexicon. However, words are not always used in exactly the same way (Ehlich 1999). For learners of academic language this fact is highly relevant as they might already know the everyday use of a word while the specific usage in academic language still has to be acquired. Furthermore, learners of a foreign language might have to acquire both varieties at the same time. Consequently, what we need for a proper didactic preparation of academic language is a close comparison of academic language with a register of everyday language like journalistic language.

In this study, I focus on the use of speech act verbs in academic German. According to Harras & Proost's (2005, 319) theory, in which they use the term 'communication verbs', this type of verbs refer to situations that consist of 'a speaker, a hearer and an utterance which, in the prototypical case, contains a proposition.' This class of verbs is highly relevant to academic language as academic writing requires the authors constantly to quote and summarise what other researchers have written. Speech act verbs are not only used for introducing a quotation or summary but also for classifying and evaluating statements (Hyland 2004, Fandrych 2004).

In order to describe the differences in the use of speech act verbs between the two registers, I compiled two corpora: One consisting of 101 German online journal articles from educational studies and, for comparison, a subsection of the German newspaper corpus DeReKo1. They were lemmatized and PoS-tagged to enable a lemma search for the seven speech act verbs zeigen ('show'), darstellen ('depict'), beschreiben ('describe'), sagen ('say'), nennen ('name'), diskutieren ('discuss') and behaupten ('claim'). I searched both corpora for the verbs in question and annotated the resulting sentences manually for aspects of grammar, semantics and function. Altogether, 722 sentences were annotated for the academic corpus and 733 for the journalistic corpus.

The research indicates that there are significant differences in the use of speech act verbs between the corpora on several language levels. While some verbs are much more common in one of the registers, others show differences in meaning between the two corpora. In some cases, these semantic differences are also associated with specific grammatical patterns, e.g. what kind of objects the verb takes. Furthermore, also functional differences are visible. For instance, the verb sagen ('say') is mainly used for reported speech in the journalistic corpus and for hegding in the academic corpus.

This research shows that corpus studies of register variation can help identify aspects of language that might be challenging for language learners. The results of this research could be used to make students aware of differences between everyday language and academic language. In addition, the results suggest that the use of speech act verbs might generally be a suitable indicator for registers.

Information density in scientific writing: Exploring the SciTex corpus

Stefania Degaetano-Ortlieb (Saarland University)
Hannah Kermes (Saarland University)
Ashraf Khamis (Saarland University)
Elke Teich (Saarland University)

The linguistic evolution of scientific writing is characterized by two major motifs: *specialization* and *conventionalization*. The assumption is that as scientific domains become more specialized, particular meanings become more predictable in these domains and call for denser encodings that minimize redundancy while maintaining accuracy in transmission. Specialization is manifested linguistically by densification in encoding, something observed, for example, on single text instances from the language of physical science (Halliday 1988). Balancing the effects of specialization, conventionalization leads to greater linguistic uniformity, i.e. over time scientific texts show greater resemblance to one another and are more clearly distinguishable as scientific. Our main hypothesis is that the linguistic features realizing specialization and conventionalization serve to optimize information density in scientific writing. This hypothesis is based on recent work in psycholinguistics, which suggests that there is a correlation between variation in linguistic encoding and information density (see e.g. Aylett and Turk (2004); Levy (2008)). It is assumed that highly informative (i.e. informationally dense) parts of an utterance are less predictable and thus realized by more expanded linguistic forms, while less informative parts are realized by shorter, more reduced forms.

To empirically investigate information density in scientific writing, we use the SciTex corpus (see Teich and Fankhauser 2010; Degaetano et al. 2013) which covers nine scientific disciplines (computer science, computational linguistics, linguistics, bioinformatics, biology, digital construction, mechanical engineering, microelectronics, and electrical engineering). The corpus is annotated for structural information (such as sections (Abstract, Introduction, etc.), paragraphs, and sentences) as well as positional information (such as lemma and part of speech). To compare informationally dense vs. less informationally dense text, we consider *abstracts* vs. *research articles without their abstracts*, assuming that abstracts are more informationally dense than their research articles.

In terms of methods, we use (1) text classification and (2) calculation of cross-entropy rate. Text classification is performed by considering linguistic features possibly involved in optimizing information density (e.g. high/low standardized type-token ratio, high/low lexical density, complex/simple NPs, complex/simple clause structure, use/omission of relativizer, etc.), looking at how well abstracts can be distinguished from research articles by these features and which features mainly contribute to the distinction. Calculation of cross-entropy rate is based on Genzel and Charniak (2002), considering entropy at each token position. Particular tokens have higher entropy rates, showing peaks in entropy (e.g. lexical words), while others have lower entropy rates, pointing at troughs (e.g. function words). This helps to explore whether abstracts have a higher cross-entropy rate than research articles (i.e. show a higher amount of peaks in entropy, being thus more informationally dense). Here, we also consider the variation among scientific disciplines.

In the talk, we will present the methodology applied and the results from (1) text classification, which show that abstracts are indeed distinct from research articles in terms of linguistic features involved in information density, as well as from (2) cross-entropy calculation, which also points to distinctions between abstracts and research articles and differences across disciplines.

Use of topic models as a means to explore a corpus of academic English

Dominik Vajn (University of Birmingham)
Akira Murakami (University of Birmingham)

In this talk, we will introduce the use of topic models (Blei, 2012; Griffiths & Steyvers, 2004; Grün & Hornik, 2011) as a way to explore a corpus of academic English. Topic modeling is a suite of machine-learning algorithms that automatically identify "topics" in a given corpus according to patterns of word co-occurrence in individual texts (e.g., If word X is frequent in a text, word Y is also likely to be frequent in the same text). Certain sets of words co-occur frequently, and can be considered to form a topic. Each text is a composite of multiple topics of different probability, and a text with a high probability of a topic can be considered a key text of that topic.

The corpus we employ includes the full holdings of the Elsevier's journal *Global Environmental Change* over 20 years since its inception (1990/1991 – 2010). The corpus consists of four million words over 675 papers. We explore the chronological change of the journal through a topic model with sixty topics.

The strength of topic modeling is in that a word can be assigned to different topics depending on the words that it co-occurs with. For example, the word *level* was considered as one of the top 20 keywords in seven topics. Its use, however, varies across the topics. In Topic 23, for example, the word is used to refer to the physical sense, such as height or depths, as the following example demonstrates:

- In this study, coastal wetlands comprise saltmarshes, mangroves and associated unvegetated intertidal areas (and exclude coral reefs). Wetlands are sensitive to sea-level rise as their location is intimately linked to sea level. (2004_14_1_Nicholls_0.420021895146576)

On the other hand, in Topic 32, the word is used to refer to the rank on a scale, usually that of governmental or regional application:

- [L]ocal governments may feel they are left little option but to use their powers at the local level to respond to regional level concerns. (1995_5_4_Millette_0.489480090419058)

Similarly, in Topic 44, the word is used to express a degree of intensity or concentration:

- Under Baseline A, CO2 reaches an atmospheric concentration of 737 ppm, more than twice current levels. (1996_6_4_Alcamo1_0.824247355573637)

In this manner, it is possible to not only track the chronological frequency change of a word but also identify how different aspects of a word develop differently. For instance, the probability of Topic 44, which relates to discussions about greenhouse gases concentrations, significantly decreases over the years. This indicates that studies focussing on the impact of greenhouse gases are decreasing in the journal. On the other hand, Topic 32 gradually increases over the years, indicating an increase of the discussions focusing on the levels of regional or governmental application. Thus, although the frequency of the word *level* actually increases over the years, in the sense of the *levels of concentrations*, it actually decreases over time. The finding of this kind is not readily discoverable with traditional techniques in corpus linguistics.

Scholarly publication and writing development: Exploring interaction in research articles written by the same academic authors over time

Suthee Ploisawaschai (University of Exeter)

Academic corpora have proven useful for the investigation of different disciplinary discourses and contrastive rhetoric between novice and expert writers. However, little is known about the writing development of the same academic authors over time because most corpus studies focus on different writers or rely on a synchronic aspect of textual analysis to point out personal proclivities (e.g., Hyland, 2010). Therefore, a diachronic aspect of corpus analysis is needed to shed light on the issues of writing development in academic publication.

Based on Baker (2006)'s notion of relatively small specialised corpus (less than 200,000 words in size) and Leech (1991)'s suggestion that corpus is a carefully thought-out collection of texts required for a particular 'representative' function, I built a corpus of authentic published research articles in social sciences (175,589 words in size) to represent different points of the academic trajectory of three English-speaking full professors from their early career until now. Then, I adopted Hyland (2005)'s taxonomy of metadiscourse to explore the changing features of the three professors' interaction in academic publication over time. In this study, I adopted both synchronic and diachronic aspects of corpus analysis along with an interview with each professor to understand their writing development over time.

A synchronic textual analysis suggests that all three professors might have their own personal approaches to interaction in scholarly publication. To illustrate, one professor can be portrayed as an academic scholar who feels much more comfortable with boosters and prefers evidentials to code glosses in comparison to the two other professors.

By contrast, a diachronic textual analysis and a reflective interview with each professor indicate some important findings. To mention a few, for all three professors their recent papers contain a much higher frequency of evidentials when compared to their own earliest papers, signaling a growing significance of intertextuality and interconnected bodies of research for advancement of knowledge and argumentation in recent scholarly publication practices. Further, there is a gradually lower frequency of boosters and code glosses in their recent research articles, especially in one professor's case, to reflect a more conciliatory tone and a knowledgeable readership of the peer review panel.

These findings have suggested that a diachronic approach to corpus analysis can be useful for exploring how academic authors have developed and changed their interaction in writing over time. Therefore, timescale might need to be taken into account in contrastive rhetoric studies to better understand the differences between novice and expert writers, variations within disciplinary discourses, as well as personal proclivities and negotiations with research communities over time.

Comparing the usefulness of academic word list and academic vocabulary list: A corpus-based critical evaluation

David D. Qian (Hong Kong Polytechnic University)

At the heart of academic vocabulary studies, one of the parameters by which a given text or discourse can be judged academic lies in its profile against a reliable and valid list, either of individual words or of multi-word units. On one hand, such profiling will help inform the better development and measurement of English language learning materials and tests; on the other hand, more validity support should be lent to those newly generated lists so that they can be regarded as valid and reliable tools for application to various relevant contexts with confidence. As a result, the observed text or discourse can be thus labeled as "academic" by a quantifiable margin deriving from the profiling results. In the context of profiling spoken discourse features in academic settings for a TOEFL iBT validation project, using three academic spoken corpora of three million words, namely, the Spoken Sub-corpus of TOFEL 2000 Spoken and Written Academic Language (T2K-SWAL), British Academic Spoken English Corpus (BASE), and Michigan Corpus of Academic Spoken English (MICASE), it became clear that academic vocabulary forms an important part in such databases. In addition, existing research in other contexts also contends that vocabulary knowledge plays a significant role in various modes of communication, providing further support for our argument that knowledge of academic vocabulary facilitates academic communication and learning academic vocabulary forms an important learning task for EFL learners. Since the creation of the University Word List (Xue & Nation, 1984), a number of new academic vocabulary lists have appeared, which include the Academic Word List (AWL, Coxhead, 2000), Academic Formulaic List (AFL, Simpson-Vlach & Ellis, 2010), Phrasal Expressions List (PHRASE List, Martinez & Schmitt, 2012), and Academic Vocabulary List (AVL, Gardner & Davies, 2013). These lists were created following different approaches and using different databases, and therefore somewhat differ in their features and content of coverage. The present paper reports on an in-depth comparative study of AWL and AVL. The study first reviews the approaches and criteria that were adopted in creating AWL and AVL, then examines the contents of the two lists based on their stated and unstated criteria, and finally evaluates the usefulness of the two academic wordlists as regards their scopes of coverage as tools for evaluating academicality, or the density of academic vocabulary, in the three academic spoken corpora being investigated. The results of our analysis indicate that, while the lengths of the two lists differ greatly, the differences in their scopes of coverage of real academic vocabulary in BASE, MICASE and T2KSWAL are not as great as they would seem to be. We will report in detail our analysis, findings and recommendations for improvement.

Towards a Norwegian academic vocabulary list

Ruth Vatvedt Fjeld (University of Oslo)
Janne Bondi Johannessen (University of Oslo)
Kristin Hagen (University of Oslo)
Arash Saidi (University of Oslo)

We present the first attempt at creating a Norwegian Academic Vocabulary list (NAV) for the Norwegian Bokmål variety.

1. Academic Vocabulary

Gardner & Davies (2013:8) state that academic core words are "those that appear in the vast majority of the various academic disciplines" in contrast to the general high-frequency words "that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register" and in contrast to academic technical words "that appear in a narrow range of academic disciplines". The purpose of our research is to test and evaluate a method for identifying and extracting academic core words as defined above.

2. Academic Corpus

Prior to our endeavours no general academic corpus existed for Norwegian. We have assembled a set of academic texts in order to create an academic corpus of our own. To this end we used the University of Oslo digital publications archive (DUO), which consists of master's theses, doctoral dissertations, and journal publications at the University of Oslo. These documents have been downloaded in pdf format, converted to text, identified with respect to language (Norwegian Bokmål), and lemmatised with the Oslo Bergen Tagger. There are 9689 documents in the corpus totalling approximately 310 million tokens. The documents are from all eight faculties of the University of Oslo (Humanities, Education, Medicine, Social and Economic studies, Mathematics and Natural sciences, Law, Theology and Odontology). The corpus is arranged according to these and their respective departments.

3. Methodology

Our approach follows the method of Carlund et al. (2012), which is a modified version of the method utilised by Coxhead (2000, 2011). The method consists of three steps:

a. Reduced frequency: For each word in the corpus, the corpus is divided into a set of intervals based on that word's frequency. Then, the intervals that contain that word (at least once) are counted. This measure gives an indication of whether a word is spread out across the corpus, or if it is concentrated in a smaller section.
b. Range: To be certain that a word in the list is common to all the university, we used a selection method that removed words that had a reduced frequency of less than 15 per million tokens in each departmental section.
c. Removal of everyday words: Finally, using a stop list we remove words that have a high frequency in general language usage. These words have a high frequency across the corpus and a high reduced frequency, but cannot be considered be academic core words.

4. Further Work

The method we have used depends on the kind of stop lists used and the number of words they contain. The method described in Gardner & Davies (2013) does not use a stop list, so in the future we will compare with this method. We will then test the validity of the lists by examining the coverage measure on academic texts in comparison to other kinds of text.

Vocabulary test development with EAP specialized corpora

Magdolna Lehmann (University of Pécs)

The aim of the study reported here was to develop and validate vocabulary tests that meet the special lexical needs of students of English in a Hungarian university. The theoretical framework of the study integrates findings of latest research on reading comprehension, vocabulary testing and corpus linguistics. A wealth of studies support the prevalence of word knowledge over syntax in text comprehension (Laufer, 1997), a major task required of students at university. Even though there is no consensus on the minimal vocabulary size necessary for pursuing academic studies (Hazenberg & Hulstijn, 1996; Schmitt, 2010), vocabulary range versus depth has been shown to be predictive of achievement in reading, writing, language proficiency and general academic success (Morris & Cobb, 2003; Zareva, 2005), indicating how students are able to cope with the reading load in their courses. As words common in academic texts behave differently across disciplines (Coxhead, 2013), learners and testers need to focus on discipline-specific vocabulary items in higher education.

Therefore, the research questions addressed intended to explore what words are common and thus invaluable for students to know in readings in English Studies, what test format best fits the purposes of filtering students with inadequate vocabulary knowledge and how the developed test items work. It was assumed that besides a good knowledge of general low-frequency and academic words, being familiar with specific lexis rare in general English texts but frequent in the discipline of English Studies highly increases the potential of students in academic text comprehension. Consequently, these words should ideally form the basis of vocabulary testing in this particular context.

In a quantitative research design the data collection instrument, based on criteria of practicality, objectivity, suitability for computerized item analysis and ease of both scoring and administration, was a discrete point, receptive vocabulary size test. The innovative feature of the tests lies in the item-selection procedure: items were chosen based on a representative specialized corpus of texts used in our English Studies programme compiled specifically for this study, instead of adopting general-purpose tests of academic word knowledge. The participants were over 500 English majors in their first academic year, all native speakers of Hungarian. Data collection took four years on eight test occasions. The initial two test versions were revised and tested in two subsequent years after a thorough analysis of descriptive statistics, item characteristics, facility values, discrimination indices and IRT data on each test occasion in order to eliminate weak items. As a result of the process, two validated and reliable (α_1= .851; α_2= .828) 30-item parallel tests are reported here, which may be applicable by other institutions running English Studies programmes for the purposes of filtering students with inadequate lexical knowledge.