Paper Abstracts

Cristina Mayer Acunzo
São Paulo Catholic University – Brazil

The language of web registers – a multidimensional study of register variation in English and Brazilian Portuguese

In this paper we present a study of Internet text varieties in both English and Brazilian Portuguese aimed at comparing those registers and non-Internet registers along the dimensions of register variation set forth by Biber (1988) for English and carried out by Berber Sardinha, Kauffmann and Acunzo (forthcoming) in Brazilian Portuguese. Online communication is broadly used and studies on this language are still incipient. This research's purpose was, on the one hand, to answer the question of where today's online texts are in relation to the pre-Internet texts in the descriptions of register variation mentioned and, on the other hand, to draw a comparison among the Internet registers variant. A corpus of 12 registers for each language was designed and compiled following the recommendations introduced by Biber (1993) and the dimension scores for the Internet registers were mapped onto the existing dimensions of variation in both languages. The results of the addition of the registers to the existing dimensions for English will be shown and discussed in the paper presentation. The results will also show a comparison with both Titak and Robertson (2013) and Berber Sardinha (forthcoming), for similarities and differences with respect to how web registers stand in relation to each other and to extant pre-web registers.

Adnan Ajsic
Northern Arizona University

Effects of Corpus-Based Instruction on L2 Recognition and Recall of Signal Markers

Although signal markers (SMs) are a distinctive feature of academic prose (Biber, Johansson, Conrad & Finegan, 1999) and a crucial type of lexical discourse structure (Grabe, 2009), they have been neglected in the EAP literature on vocabulary instruction (e.g., Degand & Sanders, 2002; Jones & Haywood, 2004). SMs can take both single- and multiword form and Alali & Schmitt (2012) suggest that similar methods may be effective for teaching both single-word vocabulary and formulaic sequences. Römer (2011) notes that corpus-based approaches can be particularly effective in vocabulary instruction, but calls for pedagogic practice to match materials and methods with groups of learners. This study recruited a total of 69 advanced intermediate L1 Arabic and Chinese ESL learners in four intact IEP classes to test the effectiveness of different types of instruction on students' ability to recognize and recall SMs. Based on a specialized pedagogic corpus, ten single- and multiword SMs were selected and taught over a period of four weeks using three experimental treatment conditions (direct corpus-based, indirect corpus-based, and dictionary-based). The effects of treatment conditions were tested using five tasks in a pre-test, post-test study design. The results corroborate the previous findings that all explicit methods are effective for the teaching of text structure (cf. Grabe, 2009), but also suggest an aptitude-treatment interaction effect, whereby the learners' L1s combined with the treatment conditions and tasks to produce L1-related differences in effectiveness. A survey of learners' perceptions of treatment effectiveness showed a degree of congruence with the actual treatment effects.

Mohammed Albakry
Middle Tennessee State University & University of Connecticut

Telling by Omission: Hedging and Calibration in Academic Recommendation Letters

This corpus-based study explores some of the linguistic and discursive aspects of framing positive and negative information—mainly modals, evaluative adjectives, and mitigation strategies—in recommendation letters. The corpus is comprised of 114 letters of recommendation spanning three years of applications to an English PhD program with a total word count of approximately 46,000 words. Through quantitative and qualitative analysis informed by the appraisal framework (Martin &White, 2005) and the general category of doubt raisers (Trix &Psenka, 2003), the study found that about 79% of the letters are marked by praise and only 21% of these letters are characterized by caution and potential negative presentation. The results reveals consistent patterns in the way different types of modals and their associated collocates are used to hedge predictions as well as identifies the discursive frames of the most common mitigation strategies in presenting potentially negative information about applicants. The study illustrates the need to combine both corpus-based and qualitative methods for a more robust and fine-grained analysis of evaluative language and attempts to draw out the implications for interpreting subtle doubt raisers, absence of expected information, and the conventions of graduate applicants' calibration in the humanities.

Laurence Anthony
Waseda University

AntPConc: A Freeware Multi-Platform Parallel Concordancer

This paper describes a new, standalone, freeware, multiplatform, parallel concordance called *AntPConc*. This software is fully Unicode compliant and thus can work with mixed language corpora including Western languages and Asian Languages such as Japanese, Chinese, and Korean. It has also been designed to work with corpora that adopt right-to-left writing systems, such as Arabic and Hebrew. As *AntPConc* is an offline software tool, it gives users the flexibility to easily work with small and large Do-It-Yourself (DIY) corpora. It can also work smoothly with standard parallel corpora that are available for download.
To date, there has been a very limited number of software options for corpus linguistics who are interested in analysing parallel corpora in an offline environment. In the currently available mainstream parallel corpus tools, loading a parallel corpus can be a complex task. Small mistakes in the choice of settings or font type can result in the corpus not loading correctly and thus producing spurious results. In this paper, I will show how *AntPConc* simplifies the process of loading parallel corpora by first 'guessing' the appropriate character encoding of the files and then producing an internally indexed version of the files utilizing the more standard UTF-8 encoding where necessary. This resulting indexed version of the corpus can be exported as a single file, distributed, and loaded directly into the software in a single step. Other features of *AntPConc* that simplify working with parallel corpora will also be introduced and demonstrated in the presentation.

Francisco Javier Barrón Serrano, Cynthia M. Murphy, Jennifer Roberts & Eric Friginal
Georgia State University

Semantic and lexical analyses of K-12 parental documents for refugees: A comparative corpus study

The current study details a comparative corpus-based (critical) discourse analysis of corpora containing educational documents distributed to parents and guardians of K-12 children in public education. The exploratory corpus (*N* = 87; 227,726 words) is comprised of parent-directed educational documents collected from four public schools in a small city (population 7,500) located in the southeastern United States with a high percentage (31.8%) of foreign-born residents. It is estimated that over 60 languages are spoken within the total 1.1 square miles of this city. The comparison corpus (*N* = 125; 445,015 words) contains parent-directed educational documents collected from a sampling of K-12 schools across the United States. Text types collected for both corpora include student handbooks, inception materials, meal information, school-to-parent correspondence, school newsletters, and disciplinary material. Word frequency, n-grams, and multiword sequences with variable slots are investigated using AntConc, a freeware concordance, and kfNgram, a phraseological search engine. Linguistic Inquiry and Word Count (LIWC) is also used to conduct a comparative semantic analysis of the two corpora. Preliminary lexical and semantic analyses indicate quantifiable linguistic differences between the two corpora, suggesting that the language of the local corpus may target immigrant parents specifically and position them in a way that hinders their potential participation at home and in their communities. Results carry implications for language policy in public education in general and for policies related to K-12 immigrant parent/guar_dian correspondence in particular. Some implications for adult immigrant English language pedagogy are suggested as well.

Tony Berber Sardinha & Marcia Veirano Pinto

São Paulo Catholic University

Dimensions of variation across television registers

The Multi-Dimensional (MD) approach to register variation has been applied to many different registers, a number of which are from the entertainment industry, such as cinema and television. The MD analysis of TV content has been limited to a few individual programs, such as the sitcom “Friends.” Thus, a gap exists in the literature with respect to analyses of the broad range of content shown on TV. The main goal of this project is to help fill this gap by carrying out a comprehensive analysis of TV registers, including sitcoms, cartoons, news, sports, commercials, soap operas, and talk shows. A corpus was collected from closed captioning services streaming American television programs shown in English on both terrestrial and cable channels. The corpus was tagged using the Biber Tagger and later post-processed using the Tag Count program, which identified further features; ultimately, more than 200 different linguistic characteristics were identified. This paper reports an ‘additive’ MD analysis, in which the TV registers are mapped onto the existing English dimensions. Dimension scores for each text on each of Biber’s 1988 dimensions were computed, indicating where each TV register is placed along the dimensions. This allowed for the comparison of TV registers with the existing registers previously described by Biber (1988), thereby revealing how television programming is similar to or different from other registers in English. At the same time, this allowed for comparisons among the TV registers themselves. The paper will present these results in detail.

Tony Berber Sardinha

Sao Paulo Catholic University

Looking at cultural shifts in English over time: A Multi-Dimensional perspective

The Multi-Dimensional approach to corpus analysis is a powerful method for discovering underlying patterns of co-occurrence in language in use. It has been applied successfully to a range of corpora, both synchronically and diachronically. In this paper, we try to extend its reach by looking at the possible relationship between collocation use and time periods in an attempt to verify to what extent collocation can reveal socio-historical cultural trends. The main goal of this paper is to discover large-scale, lexically defined, diachronic representations of particular national cultures—namely, the US, the UK, and Brazil. The corpus consisted of a sample of Google Books published in English between 1850 and 2008 (159 years), totally approximately 450 billion words. The method consisted of obtaining the 4-grams for the entire sample and then for each focus word (American, Brazilian, British), grabbing its neighboring words in the 4-gram (its collocates), the year in which the n-gram occurred, and the 50 most frequent collocates. These frequencies were normed and run through a factor analysis. Five factors were extracted for both ‘American’ and ‘Brazilian’ and four for ‘British’. These were interpreted, suggesting strong relationships between collocation and time periods. To illustrate, the first factor for ‘American’ accounted for 97.4% of the variation in collocation and suggested a change between 1940 and 1950, with the pre-WW2 era characterized by such collocations as American + people/nation/citizens/flag whereas in the post-war era, ‘social’ collocations prevailed like American + society/century/economy/dream. The paper presents the different factors and their interpretation.

Silvia Bernardini & Adriano Ferraresi

University of Bologna

Language variation and institutional academic English: a study on phraseology

University registers of an institutional kind−e.g. course syllabi, university brochures−are increasingly attracting scholarly attention (Biber 2006). Research so far has focused on native texts, yet it has been suggested that "in order to understand the use of English in present-day academic communities, it is vital to look at English as a lingua franca" (Mauranen 2010). Indeed, universities in non-English speaking countries worldwide also use English to communicate with their stakeholders, trying to stand out in the global educational market. In this paper a corpus of online course syllabi is used to investigate phraseological patterns in native and non-native texts produced by European universities. The latter are sampled based on the language family of their official L1s, i.e. Romance (e.g. France), Slavic (e.g. Poland) and Germanic (e.g. Denmark). Drawing on Durrant and Schmidt (2009), we extract contiguous *pre-modifier* + *noun* sequences from the non-native and the comparable native (British) subcorpora. Deriving frequency data from *ukWaC*, we classify word sequences in three sets: 1) frequent vs. infrequent/unattested combinations, “strong” vs. “weak” collocations 2) based on t-score and 3) based on Mutual Information. Finally, we compare the degree to which the varieties represented in the corpus rely on different types of combinations. Results point to a significant overuse of infrequent combinations and underuse of strong collocations in Romance and Slavic countries, while differences between native and Germanic non-native texts are less marked. The paper discusses these results and their relevance for research on institutional academic English and native/non-native use of phraseology.

References
Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
Durrant, P. and N. Schmitt (2009). "To what extent do native and non-native writers make use of collocations?". *International Review of Applied Linguistics in Language Teaching* 47(2), 157–177.
Mauranen, A. (2010). "Features of English as a lingua franca in academia". *Helsinki English Studies* 6.

Patricia Bértoli
Rio de Janeiro State University (UERJ)

Lexical Bundles in Brazilian Students' academic writing teaching

This paper aims at presenting the results of a research on the use of lexical bundles, "simple sequences words" (Biber et al., 1999), present on essays of Brazilian students of English for Academic Purposes before and after exposed to specific activities. The initial research corpus of students' essays (about 45,000 words) was contrasted to LOCNESS (Louvain Corpus of Native English Essays) in order to verify the concomitant use and frequency of bundles (Biber et al., 2004; Dutra; Berber-Sardinha, 2013, Bertoli-Dutra, 2013). The research methodology included the collection of a study corpus in two different moments. The first collection was followed by the extraction of 3, 4 and 5-word bundles. After that, lexical bundles in the study corpus were verified as existing or not in the native speakers' corpus. The analysis pointed to the overuse of malformed lexical bundles such as "*in another countries*", in the Brazilian students' essays. Most and least frequent and malformed lexical bundles were selected and were used as the basis for the development of classroom hands-on activities where students were led to notice and practice bundles differences in use. A second collection of essays was made and the contrast of pre and post activities essays showed that although their lexical constructions still presented some rooting to lexical simplicity and influence by their first language, students improved their use of well-formed lexical bundles.

Vaclav Brezina & Dana Gablasova
Lancaster University

Epistemic markers in the Trinity-Lancaster spoken learner corpus: Effect of L1 background and task

Indicating epistemic stance (certainty and uncertainty) is an essential part of natural communication. This complex pragmatic phenomenon deserves our attention especially in relation to advanced learner language where it can show how successful learners are in natural discourse interaction and meaning negotiation (cf. Kärkkäinen 1992; Aijmer 2002). While several studies addressed the use of epistemic stance in learner writing, only little attention has been given to learner speech. This study therefore focused on the use of adverbial epistemic markers (AEMs) such as *probably* or *certainly* (cf. Biber, 2006) by advanced L2 speakers. In particular, we investigated to what extent the use of AEMs is affected by learner differences (L1 background) and by contextual factors (different communicative tasks).

The study was conducted using a corpus of advanced learner speech, which consists of transcribed dialogues between learners and examiners. The corpus contains 0.5M words and includes spoken production from 133 non-native English speakers from six L1 backgrounds performing different tasks (both monologic and interactive).

The findings show that the production of AEMs differs significantly according to both the L1 background and the type of task. As for the L1 background, Chinese, Mexican and Spanish users of English used considerably more AEMs than speakers of other L1s. With respect to the effect of the task, as expected, the monologic task elicited the smallest number of AEMs. On the other hand, by far the most AEMs were elicited by the task which required the learners to take responsibility for maintaining the flow of the conversation.

Dan Brown
Northern Arizona University

Exploring syntactic complexity: Variation in language use across writing task types

Applied linguists have identified language complexity as an important measure of L2 proficiency and development, relying on measures of coordination or subordination within T-units. Several researchers (e.g., Larsen-Freeman, 2009; Ortega, 2003; Palloti, 2009) have argued against any clear agreement on a definition of syntactic complexity that can reliably differentiate texts in a way that is sensitive to context (e.g., proficiency, register, task type). Recently, corpus informed approaches have shown great potential in better understanding language complexity by differentiating the patterns of syntactic features that may best represent complexity across registers (Biber, Gray & Poonpon, 2011) and proficiency levels (Lu, 2011; Parkinson & Musgrave, 2014). These studies have used a corpus informed approach to identify the noun phrase, in favor of T-unit, to distinguish complexity in academic writing. The present study seeks to add to this line of research by exploring word lists to identify potential differences in syntactic patterns across writing task types (genres), and then investigating those differences to determine which differentiate across genres of learners' writing. A corpus of two hundred texts written by 13 Thai university-level EFL learners are analyzed for syntactic and lexical features across four different writing task types (personal, response, comparison, and process writing). Noun phrases, verb morphology, and lexical diversity (among other targets) are analyzed. Preliminary findings show that writing task type impacts the syntactic choices in learners' production. A better understanding of this influence can help identify task-appropriate indicators of written complexity to inform judgments of L2 writing quality and development.

Maggie Charles
Oxford University Language Centre

Personal EAP corpora: What do independent users do?

This paper reports on a new data set of 40 corpus users who were surveyed about consulting their corpus independently a year after completing a corpus course. The respondents had been introduced to corpus use as part of an EAP writing programme in which they built individual personal corpora of research articles in their field. The purpose of the study was to investigate the extent and nature of independent corpus use by examining the types of consultation and attitudes to corpus work that prevailed over the longer term. It was found that 29 students (73%) had used their personal corpus for extended periods (3 months to over a year). Two groups of independent users were identified: 25 (63%) frequent users (once per week or more) and 15 (37%) infrequent users (once per month or less). Frequent users were more likely to have made modifications to their corpus by adding, deleting and cleaning files, thus showing greater commitment to their resource. They were also more positive about the effect of corpus use on improving their writing and tended to consider that their search techniques had improved. Both groups were equally likely to use the corpus to check items in their writing, but frequent users were more likely to look for new items to use; they tended to sort concordance lines more often and to use more context searching. This paper reports further on the practices of independent users and discusses the implications for encouraging students to use corpora outside the classroom.

Joseph Collentine and Yuly Asención-Delaney
Northern Arizona University

The Discourse Function of the Subjunctive in Foreign-Language Learners of Spanish

The Spanish subjunctive plays a central role in the Spanish foreign-language (FL) curriculum (Collentine, 2013). Although recent research studies the pragmatic and linguistic variables accounting for learner variation with the subjunctive in controlled conditions (Geeslin & Gudmestad, 2008), little research has studied the discursive features that learners associate with the subjunctive in writing. This would provide insights into the discursive functions that learners assign the subjunctive. We provide a corpus-based analysis of the lexico-grammatical features predicting the use of subjunctive with FL learners at the second-, third-, and –fourth years of university-level Spanish instruction. We employ a 200,213-word learner corpus of the learners' unedited written Spanish, identifying 890 subjunctive instances. To understand the subjunctive's discursive function while accounting for the likelihood that the learner data contained numerous topic shifts, we focused on a window of 20 words surrounding each subjunctive instance. Within that window we tabulated 38 variables known to be associated with Spanish learner production (Asención-Delaney & Collentine, 2011), which were reduced to 16 through a stepwise discriminant analysis. The discriminant analysis yielded two significant discriminant functions distinguishing the three levels of learners. The data suggest that learners increasingly associate a combination of hypothetical and informationally rich (i.e., semantically dense) features with the subjunctive as they progress from the second to the fourth year, employing it towards the production of encyclopedic language more and more over time. Additionally, the third-year learners employed the subjunctive in contexts entailing a combination of hypothetical and narrative features.

Works Cited

Asención-Delaney, Y., & J. Collentine. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32, 299-322

Collentine, J. (2013). Subjunctive in second language Spanish. In K. Geeslin (Ed.) *The Handbook of Spanish Second Language Acquisition* (pp. 270-287). Chichester, UK: John Wiley & Sons.

Geeslin, K., & A. Gudmestad. (2008). Comparing interview and written elicitation task in native and non-native data: Do speakers do what we think they do?" In J. Bruhn de Garavito & E. Valenzuela (Eds.) *Selected Proceedings of the Hispanic Linguistics Symposium*, (pp. 64-77). Somerville, MA: Cascadilla.

Susan Conrad
Portland State University

Situating Engineering Writing in the Multi-Dimensional World of English Discourse

This paper describes a study that applies Biber's (1988) dimensions of variation in English to writing in engineering. The study focuses on civil engineering (CE), a field where professional advancement depends on writing skills. The study is part of a larger corpus-based project sponsored by the National Science Foundation, in which linguists and engineers are

collaborating to improve the teaching of writing to undergraduate CE majors.

Using registers from the corpus compiled for the larger project, this study seeks to situate academic research articles, practitioner writing, and student writing from CE along a range of registers of English in order to understand the challenges that novice civil engineers face in gaining discourse competence. Specifically, the study addresses the following questions:

- To what extent do professional registers of CE writing share the linguistic features of academic writing generally (as found by Biber, 1988)?
- To what extent do the registers written by CE practitioners (e.g. design reports, site visit observations, plan sheet notes) differ from research articles written by academic faculty?
- When students write papers for assignments that are designed to mimic practitioner registers, to what extent does the student writing differ from practitioner registers?

Along certain dimensions, the CE registers are very similar to each other. Along others, differences exist and student weaknesses are clear. For example, on Dimension 3, Explicit vs Situation-dependent Reference, practitioner registers differ from both academic prose and the student writing in being more situation-dependent. Students neglect to express the precise space and time information typical of practitioner writing, using features more typical of academic writing.

Vanessa Conte Herse
California State University, Long Beach

Hedging in Court: A Corpus-based Study of Gender Effects on Testimony Language

This study measured the frequencies of two types of spoken constructions, hedge phrases (e.g. sort of, occasionally) and hedge clauses (e.g. I guess, I know), in witness testimonies to see if gender or examination type had quantitative effects on hedge production. A 40,476-word corpus of natural speech from 21 court cases was designed to contain a balanced sample of case topics, speaker education levels and occupations, and trial dates. The corpus contains data almost evenly split between males and females and among two examination types, direct and cross. In a direct examination, the witness is guided with questions to explain their evidence, whereas in a cross-examination, witnesses undergo a more aggressive questioning meant to devalue their claims. These variables were chosen because of their potential to express power dynamics in the courtroom. A factorial MANOVA found a significant main effect for gender on hedge production overall ($p = .021$), but none for examination type, $p > .05$. No interaction effect was observed, $p > .05$. Additionally, the production of hedge phrases, one or two-word adverbs or quantifiers and no pronouns, was significantly different ($p = .035$) among males ($N =15$) and females ($N = 15$) yet the production of hedge clauses, first-person-singular pronouns plus verbs, showed no difference, $p > .05$. These findings suggest that men and women use varying syntax to express the truth in testimony and propose that gender in the regulated linguistic domain of the court may influence how dynamics of power are maneuvered through language.

Viviana Cortes
Georgia State University

Analyzing the semantic prosodies and preferences of lexical bundles in research article introductions

The study of lexical bundles, sequences of three or more words that occur frequently in a register (Biber, Johansson, Leech, Conrad, & Finegan, 1999), has become the focus of many corpus-based studies in the last decade.
This presentation reports the findings of a study which analyzed the use of lexical bundles in a one-million word corpus of research article introductions. Those bundles were analyzed in terms of their communicative functions using a research article introduction move scheme (Swales, 2004). This analysis showed a strong connection between lexical bundles and the moves they help communicate. While some lexical bundles were used to trigger moves, other bundles were used as "comments," adding information to the discourse used to trigger moves (Cortes, 2013).
The focus of this presentation will be on this second group of bundles, comment bundles, and their relationship to the contexts in which they are used, analyzing the semantic prosodies and semantic preferences of these lexical bundles (Xiao & McEnery, 2006). This analysis showed that most 4 and 5-word comment bundles could be easily identified as relating to a particular semantic prosody (positive, negative, or neutral). In addition, a taxonomy that reflects the semantic domains frequently referred to in these contexts was designed and used to categorize the semantic preferences of the most frequent comment bundles.
The proposed presentation will introduce various pedagogical applications of the findings of this study, implications for the study of formulaic language in this academic register, and suggested paths for future research.

Eniko Csomay & Viviana Cortes
San Diego State University & Georgia State University

Lexical Bundles in Cyber-texts

Lexical bundles are sequences of word combinations frequently occurring in a register. To be considered a bundle, a four-word combination, for example, has to occur at least 20 times in a corpus and in more than 5 texts. Bundles vary in their length, and although they are not complete grammatical units, they can be classified into grammatical groupings (Biber et al. 1999). They are examined for the most typical functions they serve in texts (Biber, Conrad & Cortes 2004) as well as the most preferred position of the various functions within the structure of discourse (Csomay 2012). Bundles have been found in multiple languages (Ventura, Cortes & Biber 2007) and in different registers (Biber & Barbieri 2012). While a number of studies report on these aspects of bundles in 'traditional' registers and in multiple languages, relatively few scholars have examined and reported on lexical bundles in 'new' registers including electronic texts.

The present study reports on lexical bundles found in a one-million word corpus of cyber-texts collected from five internet-registers: pop-culture news, advertising, forum requests for advice, blogs, and tweets (Connor-Linton 2012). First, we report on the lexical bundles of four to eight words found in these five registers, and examine their functions. We compare their functions with those found in non-electronic registers (e.g., referential bundles, discourse organizers, and stance markers) and highlight those that we found specific to these electronic registers (e.g., descriptive, narrative). Second, we present data on the distributional patterns of four-word bundles and their functions in each register. Preliminary findings indicate that stance bundles are most frequent in forum texts as referential bundles are in pop-culture news. Finally, we report on how the bundle functions are positioned in these registers with varying text-lengths.

Mark Davies
Brigham Young University

Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE)

In this paper, I provide an overview of the new GloWbE corpus – the Corpus of Global Web-Based English, which was released in 2013. GloWbE is based on 1.9 billion words in 1.8 million web pages from 20 different English-speaking countries. Approximately 60% of the corpus comes from informal blogs, and the rest from a wide range of other genres and text types.

Because of its large size, as well as because of its architecture and interface, the corpus can be used to examine many types of variation among dialects of English, which might not be possible with other corpora, such as the much smaller International Corpus of English (ICE). These include variation in lexis, morphology, (medium- and low-frequency) syntactic constructions, variation in meaning, as well as discourse and its relationship to culture.

Elif Demirel
Karadeniz Technical University

Lexical and Grammatical Variation in Scholarly Writing: a Multidimensional Comparison of Published Native and Non-native Research Articles

In the global academia, the number of scholarly publications written in English by Turkish researchers indexed in the Arts & Humanities Citation Index has been increasing (Al, Şahiner and Tonta 2006). Additionally, social sciences journals published in Turkey by universities and other scientific organizations have started to accept and publish English publications alongside Turkish publications. Despite the increased tendency of Turkish researchers to publish in English both in Turkey and globally, there has been little research comparing Turkish researchers’ scholarly writing with native-speakers’ scholarly writing published in English. This research attempts to identify the dimensions of variation in social sciences research by comparing journal articles published in English by Turkish researchers in social sciences journals published in Turkey with articles written by native speakers and published by social sciences journals in the USA. For the comparison, two parallel corpora were compiled and tagged, namely the TAC (Turkish Academic Corpus) and the AAC (American Academic Corpus) and compared by using the Multi Dimensional Analysis (Biber, 1988) method. After the initial factor analysis,13 variables were eliminated and the remaining 40 variables were used to run the second factor analysis. As a result of the analyses, a 3 factor solution was settled as optimal and it was observed that there exists variation not only between native and non-native texts but also between various subject within the broad area of social sciences.

Sandra C. Deshors & Stefan Th. Gries
New Mexico State University & University of California, Santa Barbara

EFL vs. ESL: not necessarily a continuum

The study of learner language (often based on the ICLE corpora and related resources) and the study of speakers of indigenized varieties (often based on the ICE corpora) are two fast growing areas of corpus-linguistic research. Two trends are currently shaping the development of those two areas: First, the recognition that more rigorous methodological approaches are urgently needed: with few exceptions, existing work is based on simple decontextualized over-/underuse frequency counts that fail to unveil complex non-native linguistic patterns. Second, the collective effort to bridge an existing “paradigm gap” (Sridhar & Sridhar 1986) between the EFLand ESL research areas.

This paper contributes to those ongoing developments by offering a multifactorial analysis of seventeen lexical verbs used in alternating dative constructions in speech and writing by German/French learners and Hong Kong/India/Singapore English speakers. Our goal is two-fold. First, we exemplify the advantages of using hierarchical mixed-effects modeling. This kind of modeling not only allows us to control for speaker and verb-specific effects, but also for the hierarchical structure of the corpus data, which virtually no existing study accounts for.

Second, we contribute to the hotly debated theoretical question of whether EFL and ESL represent discreet English variants or a continuum. While our results pinpoint the linguistic factors that influence non-native speakers' syntactic choices, they also clearly show that EFL speakers behave differently from ESL speakers. This result bears significant implications for the development of the field as it questions the validity of the current trend to conflate EFL and ESL research.

Doug Biber, Jesse Egbert & Lize Terblanche
Northern Arizona University, Brigham Young University & Northern Arizona University

Dimensions of Variation in English Web Registers

We currently have little information about the text categories -- the ‘registers’ -- found on the web. Although there have been many attempts to classify web documents into register categories (Rosso, 2008; Titak & Roberson, 2013) these have mostly been based on small corpora coded by a single rater. When multiple raters have been employed, inter-rater reliability has been quite low (Sharoff et al., 2010; Rosso & Haas, 2010). To address these limitations, Biber and Egbert (2013) developed a comprehensive situational framework for web register categories recognized by end users. Based on this framework, they developed a computer-adaptive online survey that allows end users to classify web documents into eight general registers (e.g., narrative, interactive discussion, opinion) and 54 sub-register categories (e.g., opinion: opinion blogs, editorials, reviews, advice). Biber and Egbert (to appear) applied this instrument to analyze the range and distribution of registers found on the web. A random sample of 53,000 URLs was coded by a team of 750 raters (through Mechanical Turk), with each URL being classified by four independent raters. In this paper we report on the next stage in this project: to describe linguistic variation across registers and sub-registers on the web using Multi-Dimensional analysis (see e.g. Biber 1988, 1995, 2006). We will use textual examples from the corpus to show key patterns of web register variation. The findings will offer important insights into the language of the internet and facilitate more principled uses of the web as a data source for linguists.

Pierfranca Forchini
Università Cattolica del Sacro Cuore Milano

A Multi-Dimensional Analysis of the legal arena: movie vs. real trials

This paper sets out to examine the linguistic similarity between naturally occurring conversation and movie conversation. For this purpose, real trials, which are usually considered interactive for their adversarial nature (Pridalová 1999) and “the closest approximation to everyday speech of all public legal discourses” (Williams 2005:24), are compared to movie trials. Using corpus-driven criteria (Francis 1993, Tognini-Bonelli 2001, Biber 2009), the comparison is made via Multi-Dimensional Analysis (Biber 1988) on data retrieved from a new sub-corpus of the *American Movie Corpus* (AMC, Forchini 2012), namely, the *American Movie-Trial Corpus* (AMTC), and from the *American Real-Trial Corpus* (ARTC, a corpus purposely-built for the present analysis). The findings show very little linguistic and textual variability of the two investigated domains and, thus, confirm that the linguistic similarity of movie and naturally-occurring conversation is also present at a more specialized level. Hence, it is suggested that movie language could be used as a remarkable source for representating and learning not only the general usage of face-to-face conversation, as recently documented (Forchini 2012, 2013a, 2013b), but also the more specialized features of courtroom discourse. These findings, which confute the claim that “it is beyond dispute that the cinematic portrayal of the American legal system and its personnel is far removed from legal reality” (Machura & Ulbrich 2001:118), also add value to the role of corpora in teaching which is often emphasized by numerous authoritative linguists (Hunston 2002, Mauranen 2004, Sinclair 2004, Reppen 2010, *inter alia*).

Dana Gablasova & Vaclav Brezina
Lancaster University

Is the core vocabulary stable across British and American English? American English supplement to The New General Service List

Lexical diversity across different varieties of English has been widely studied (e.g. Algeo 2006). On the other hand, the issue of lexical stability of English vocabulary has received only little attention. Despite this fact, this issue has important implications for English pedagogy because it directly influences the choice of vocabulary taught in different EFL/ESL contexts (cf. Nation & Chung 2009). In a recent study (Brezina & Gablasova 2013), we used the corpus method to establish a new English vocabulary baseline – the *New General Service List (new-GSL)* – which is intended to help both researchers as well as English language teachers and learners with identifying common English vocabulary. The *new-GSL* includes 2,500 most frequent words that appear in current English texts regardless of the topic/genre.

In this study, we investigate the overlap between the British and the American variety (AE) of English and introduce the *AE Supplement* to the *new-GSL*. In this research, the *new-GSL,* was compared to three American English corpora *Brown, AM06* and *COCA.* The results show that there is a large overlap (over 80%) between the *new-GSL* and the wordlists based solely on AE corpora. The resulting AE Supplement consists of less than 350 items.

References

Algeo, J. 2006. British or American English?: A handbook of word and grammar patterns. CUP.

Brezina, V., Gablasova, D. 2013. Is There a Core General Vocabulary? Introducing the New General Service List. Applied Linguistics.
Nation, P., & Chung, T. 2009. Teaching and testing vocabulary. In The handbook of language teaching. Wiley-Blackwell.

Dee Gardner and Mark Davies
Brigham Young University

Determining the Technical Vocabulary of Academic English: A Corpus-Based Analysis

This paper presents our new *Technical Academic Vocabulary List* (TAVL), derived from a 425-million-word corpus of contemporary English. The list is subdivided into nine primary academic disciplines: (1) Education, (2) Humanities, (3) History, (4) Social Science, (5) Philosophy-Religion-Psychology, (6) Law-Political Science, (7) Science-Technology, (8) Medicine-Health, and (9) Business-Finance. We first explore the reasons why frequency-based lists of technical (discipline-specific) vocabulary are important for education. We next provide detailed explanations of the robust methodology used to identify the technical academic vocabulary within the larger corpus, and how technical academic words were distinguished from core academic words shared by most disciplines. We then present case studies demonstrating how well our TAVL discriminates between random academic texts (not in our corpus) that are representative of the nine disciplines noted above. We conclude with our recommendations for extending TAVL to various instructional settings, and discuss our new web-based interface that allows users to find and interact with technical words from TAVL that appear in any texts entered in the search window.

Anna Gates Tapia
Northern Arizona University

Key terminology in business introductory textbooks: Resources for presenting new words

Introductory textbooks are one of the key resources international first year university students have for learning the key terminology of their chosen field, yet results of the effectiveness of learning vocabulary through reading in context have been mixed (Nagy, Herman, and Anderson, 1985). A possible reason for inconclusive findings has been attributed to inconsistencies in the types of contexts used (Nation 2001). In fact, few implicit vocabulary acquisition studies have explored definition rich texts. These focus instead on contexts of varying degrees of target word guessability (Beck, McKeown and McCaslin, 1983) with no detailed descriptions of the features of the actual texts used.

Expanding upon definition grammars and corpus methods used in lexicography studies (Barnbrook, 2002; Sinclair, 1991), this paper explores the linguistic environment of key terminology presented in introductory business textbooks by describing the principal components of definition sentences and the nature of their combinations. Additionally, unlike most corpus studies of textbook vocabulary which strip texts of enhancement features such as bolding, glossing and italicizing, these elements have been tagged and considered in the analysis. Pedagogical implications of findings will be discussed.

Christer Geisler & Christine Johansson
Uppsala University

Noun Phrase Modification as an Indicator of Syntactic Development in Swedish L2 Learners of English

This paper investigates noun phrase modification in learner English. The study is based on student writing from the *Uppsala Learner English Corpus* (ULEC), which consists of essays from Swedish junior and senior high school students (see Johansson and Geisler 2009). In his classic study, Hunt (1966) argues that relative clauses and adjectival premodification of noun phrases serve as good indicators of syntactic development in L1 learners. Johansson and Geisler (2011) in their study of subordinate clauses in Swedish L2 learners of English show that relative clauses increase significantly over time. In the present study, we analyze the use and frequency of premodifiers from junior to senior high school. Preliminary results indicate that premodifiers occur infrequently in Grade 7 (junior high), as in (1), whereas senior high school students start to use more complex premodification, as in (2).

1) I think when you die you get a soul, but if you are *a bad person* you be a ghost. I have herd i soul before ore a ghost, it was when i was like 5 yers old then i herd someone go in the stiers and then i shout, but dosent answer. ULEC: Female student, aged 13, grade 7

2) I feel someones presence, hear sounds and ect. Then at those moments you do believe in ghosts and *other mystical and scary stuff*. But mostly I just don't really think about it.
ULEC: Male student, aged 16, senior high school (grade 10)

More advanced L2 learners use not only more premodifiers, but also more structurally elaborated ones, such as the coordinated adjectives in (2).

References

Hunt, K. W. 1966. 'Recent measures in syntactic development'. *Elementary English* 43: 732-739.

Johansson, C. and C. Geisler. 2009. 'The Uppsala Learner English Corpus: A new corpus of Swedish high school students' writing'. In A. Saxena and Å. Viberg (eds.), *Multilingualism: Proceedings of the 23rd Scandinavian conference of linguistics*, 181-190. Uppsala: Acta Universitatis Upsaliensis.

J ohansson, C and C. Geisler. 2011. 'Syntactic aspects of the writing of Swedish L2 learners of English'. In J. Newman, H. Baayen and S. Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, 138-155. Amsterdam: Rodopi.

Anne-Line Graedler
Hedmark University College, Norway

Presentative constructions in Norwegian L2 learner English

This paper presents research on the use of presentative constructions in Norwegian L2 English, addressing the following questions:

1) How often do learners produce inappropriate presentative constructions?
2) Is there a correlation between inappropriate use and L1 influence?
3) Are there contrasting patterns of usage across the spoken and written modes, and between free L2 writing and translation?

Presentative constructions with a dummy subject, e.g. *There are two cars in the driveway*, are traditionally given much focus in English L2 teaching in Norway, since ”Norwegian uses presentative constructions with a wide range of verbs (both passive and active), and

English often requires a different kind of structure" (Hasselgård, Lysvåg & Johansson 2012: 308). Presentative constructions in English vs. Norwegian have been studied in detail based on parallel corpora with texts from professional writers and translators (Ebeling 2000; Chocholoušová 2007); however, no research exists based solely on learner interlanguage (but cf. Palacios-Martínez & Martínez-Insua's [2006] study of Swedish learners' use of existential *there*).

The present investigation explores the use of presentative constructions extracted from three different L2 English advanced learner corpora: the Norwegian subcorpus of LINDSEI (Gilquin et al. 2010), the Norwegian component of ICLE (Granger et al. 2009), and NEST, a student translation corpus (Graedler 2013). The presentative constructions have been analyzed according to grammatical appropriacy, syntactic patterns and potential L2 influence.

References

Chocholoušová, B. (2007). Norwegian *det*-constructions and their translation correspondences in English and German: A contrastive corpus based study of dummy subjects. Master's Diploma Thesis, Masaryk University, Czech Republic.

Ebeling, J. (2000). Presentative Constructions in English and Norwegian: A corpus-based contrastive study. PhD dissertation, University of Oslo. Oslo: Unipub forlag.

Gilquin, G., Cock, S. D. & Granger, S. (eds.). (2010). *LINDSEI: Louvain international database of spoken English Interlanguage.* Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Graedler, Anne-Line. 2013. NEST – a corpus in the brooding box. *VARIENG* 13. http://www.helsinki.fi/varieng/journal/volumes/13/graedler/ [In Magnus Huber & Joybrato Mukherjee (eds.), Corpus Linguistics and Variation in English: Focus on Non-Native Englishes. *Studies in Variation, Contacts and Change in English*, Vol. 13.]

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

Hasselgård, H., Lysvåg, P. & Johansson, S. (2012). *English Grammar: Theory and Use*, 2nd ed. Oslo: Universitetsforlaget.

Palacios-Martínez, I. & Martínez-Insua, A. (2006). Connecting linguistic description and language teaching: native and learner use of existential *there*. *International Journal of Applied Linguistics*, 16:2, 213-231.

Sylviane Granger & Magali Paquot
Université catholique de Louvain

The *Louvain EAP Dictionary* (LEAD): A tailor-made web-based tool for non-native academic writers of English

In our presentation we describe the *Louvain English for Academic Purposes Dictionary* (LEAD), an integrated dictionary and corpus tool intended to help non-native speakers write academic texts in English (Granger & Paquot, 2010; Paquot, 2012). The LEAD contains a corpus-based description of c. 1200 academic words and phrases (with particular focus on their phraseology: collocations and lexical bundles) and highlights the main difficulties they pose to non-native writers as attested in a large corpus of learner texts. One particularly original feature of the dictionary is its customisability: the content is automatically adapted to users' needs in terms of discipline and mother tongue background. Another key feature is that the LEAD can be used as both a semasiological dictionary (from lexeme to meaning) and an onomasiological dictionary (from meaning/concept to lexeme) via a list of typical rhetorical or organisational functions in academic discourse (Pecman 2008). It is also a semi-bilingual dictionary (Laufer & Levitzky-

Aviad 2006): users who have selected a particular mother tongue background can search lexical entries via their translations into that language.

With the advent of electronic dictionaries, corpus data are making their way into the dictionary via new components such as example banks or corpus-query systems. Unlike most online dictionaries, however, the LEAD innovates by giving access to discipline-specific corpora rather than generic corpora. Collocations and lexical bundles are also illustrated with examples automatically extracted from discipline-specific texts, thus allowing users to visualise senses in a context close to their own working environment (Williams, 2003). The dictionary therefore caters for both general (EAP) and specific (ESP) needs.

Internal Representativeness and Specialized Corpora: The Influence of Topic on the Stability of Linguistic Findings in a Disciplinary Writing Corpus

Bethany Gray Jesse Egbert and Manman Qian
Iowa State University & Brigham Young University

Corpus representativeness is rarely empirically evaluated, despite widespread recognition of its importance in corpus building and analysis. Biber (1993) remains the primary study testing representativeness in corpora of relatively general registers. However, little is known about the representativeness of smaller specialized corpora, like those often used to investigate language use within and across academic disciplines. Researchers building specialized corpora attend primarily to 'external'/'situational' representativeness (McEnery et al., 2006; Biber, 1993), assuming corpus representativeness because the target registers are relatively well-defined and identifiable. Yet little is known empirically about the influence of external factors such as topic on the stability of linguistic findings from those corpora (i.e., 'internal'/'linguistic' representativeness; McEnery et al., 2006; Biber, 1993).

We investigate the influence of topic on internal representativeness in a specialized corpus of writing in a single discipline (applied linguistics) by evaluating the reliability of a range of linguistic features (passive voice, pronouns, grammatical complexity, and multi-dimensional analysis scores). The corpus is composed of 30-text sub-corpora: one 'general' corpus representing a range of topics, and a series of topic-focused sub-corpora1 (language testing, CALL, second language acquisition and language pedagogy, pragmatics, text analysis/discourse analysis, sociolinguistics, language planning and policy). First, reliability coefficients between the sub-corpora are used to evaluate the degree to which linguistic findings are stable across the sub-corpora. Second, reliability coefficients are calculated for a series of test corpora created by randomly assigning texts from the full corpus into sub-corpora for which topic is not controlled (sampling with replacement). The presentation ends with implications for the design of small, specialized corpora intended to represent disciplinary writing.

References

Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243-257.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. New York: Routledge.

Note

1. The issue of topic is complicated, and we acknowledge that there are overlap/relationships between these topic areas. The issue of overlap and the complexities of defining topic in the context of disciplinary writing will be discussed in the study.

Bethany Gray, Douglas Biber, and Joe Geluso
Iowa State & Northern Arizona University

Quantitative Measures for Characterizing Predictability and Variability in Discontinuous Lexical Frames

Research on the phraseological patterning of language has given renewed attention to discontinuous sequences, multi-word units consisting of a ‘frame’ surrounding a variable slot (e.g., in the * of, the * of the). Because frames are by definition sequences with a variable slot, frame research has also been concerned with characterizing discontinuous patterns according to how variable they are (the number of different words that occur in the variable slot), and how predictable the variable slot is (the frequency of the most frequent filler). Previous research has relied primarily upon type-token ratios, but other statistical measures (e.g., entropy, mutual information, proportions) have also been proposed to measure predictability and variability. At the same time, Biber (2009) has demonstrated that different quantitative measures used to characterize phraseological patterns reveal different kinds of associations.

Thus, in this study we explore a range of measures of predictability and variability by calculating type-token ratios, proportions of most frequent fillers, mutual information, entropy, and .p (delta p) for all 4-word frames (patterns 1*34 and 12*4) occurring at least 40 times per million words (c. 500 frames) in large corpora of conversation and academic writing (c. 4.5 and 5.3 million words respectively from the Longman Corpus of Spoken and Written English). We directly compare results based on these different measures, interpreting what each indicates about types of multi-word associations in linguistic terms. We discuss the implications of these findings for the use of such quantitative measures in frame research, demonstrating that different measures have the potential for revealing different types of frames, and different types of associations between frames and their fillers.

Stefan Th. Gries
University of California, Santa Barbara

The most underused method in corpus linguistics: Multi level and mixed effects models

For several decades, the statistical analysis of especially experimental data in psycholinguistics was characterized by the recognition that the data points collected in an experiment are not independent of each other because (i) each subject provides more than one judgment / reaction time/ … and (ii) each experimental stimulus is judged/reacted to more than once. At this point, the state-of-the-art statistical method for such data is mixed-effects modeling. Interestingly, corpus linguists stand to benefit from this tool even more than psycholinguists because (i) corpus data exhibit the same interrelatedness of data points (speakers/writers and lexical items) as psycholinguistic data; (ii) corpus data are usually much messier/noisier than psycholinguistic data; (iii) corpus data often come with a hierarchical sampling structure such as the ICE-GB's structure represented below.

Nested (from right into left)

Mode	**Register**	**Subregister**
spoken	dialog	private, public
	monolog	scripted, unscripted
	mixed	broadcast
written	printed	academic, creative, instructional non-academic, persuasive, reportage
	non-printed	letters, non-professional

Crucially, we know corpora come in these structures and that nearly every phenomenon will exhibit differences on the levels of speakers/lexical items and on the levels of the mode and/or the register and/or the subregister – however, while some corpus-linguistic studies now use MEM to address the former level of variability, there is virtually no work at all also accounting for levels of corpus organization using multi-level modeling. In this paper, I discuss particle placement in the ICE-GB – Mary gave up[DO smoking]vs. Mary gave[DO smoking]up – and show how we can straightforwardly study such corpus data correctly. I will show how the results of this approach are far superior to those of traditional regression modeling in terms of classification accuracy and R2, but especially regarding the precision and interpretation of the results.

Jack Grieve, Diansheng Guo, Alice Kasakoff & Andrea Nini
Aston University & University of South Carolina

Big Data Dialectology: Analyzing lexical spread in a multi-billion word corpus of American English

This presentation introduces a multi-billion-word dialect corpus of American English and describes an analysis of lexical spread in American English based on this dataset. The corpus consists of approximately 800 million geo-coded and time-stamped tweets totaling approximately 9 billion words, which were tweeted by users from across the United States during 2013. On average, the corpus contains 2 million tweets totaling 25 million words per day for each day of 2013. For each tweet, the precise longitude and latitude of the user when they tweeted is known. This corpus therefore makes it possible to obtain an unprecedented view of regional linguistic variation in American English. The corpus is especially useful for investigating lexical variation, which requires very large amounts of data. In particular, this presentation will describe a preliminary analysis of the regional spread of emerging words (e.g. *thottie, furloughed, plurnt*) in American English in real time, including outlining the computational methods used to extract emerging words from the corpus. Advanced techniques for spatial analysis and geographical visualization will be used to map the spread of new words across the United States over the course of a year for the first time.

Jack A. Hardy
Georgia State University & Emory University

Biology Discourse: A Multi-Dimensional Analysis

The purpose of this paper is to better understand the written registers of academic biology. College students often must navigate multiple registers both as consumers and producers. However, such students may be unaware of the differences and similarities between registers even within their major field of study (Moran, 2012). This corpus-based investigation complements genre-based studies of networks or systems of genres within a discourse community (e.g., Berkenkotter, 2001). Following Biber (1988), I describe a multi-dimensional (MD) analysis of important text categories in biology, using pre-established dimensions extracted from A-graded, upper-level student writing across the curriculum (Hardy & Römer, 2013). Multiple types of genres are explored, including a professional genre (research articles), pedagogical genres (textbooks, lab manuals, and lectures), and student-written genres (proposals, reports, response papers, research papers, and summaries). The student-written samples are from both upper-level university students (from MICUSP) and lower-level undergraduate students at a liberal arts college in the southeastern United States. This study shows the variation present within a single discipline and provides potential pedagogical implications.

References

Berkenkotter, C. (2001). Genre systems at work: DSM-IV and rhetorical recontextualization in psychotherapy paperwork. *Written Communication, 18*(3), 326-349. doi: 10.1177/0741088301018003004

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

Hardy, J.A. Jack Grieve, Diansheng Guo, Alice Kasakoff & Andrea Nini
Aston University & University of South Carolina, & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 8*(2), 183-207. doi: 10.3366/cor.2013.0040

Moran, K. (2012). *Exploring undergraduate disciplinary writing: Expectations and evidence in psychology and chemistry.* (Unpublished doctoral dissertation), Georgia State University, Atlanta, GA.

Brandy C. Judkins
Georgia State University

Revisiting the Dolch Word Lists: Corpus-Influenced Examination & Revision

This paper describes a corpus-influenced examination of the Dolch Word Lists. While the Dolch Word Lists are used in primary grades around the globe, advances in language analysis have not been utilized in examination of these lists. In fact, most re-evaluation of these lists began and ended in the 1970s. This study bridges this gap through identification of the rates of correspondence between the Dolch Word Lists and the Corpus of Contemporary American English. This study provides guidance for educators and materials designers in regards to which aspects of the Dolch Word Lists are accurate representations of contemporary English use across five registers: (a) newspapers, (b) magazines, (c) fiction, (d) spoken discourse, and (e) academic writing. In light of the age of the Dolch Word Lists and their frequent use in the primary grades, the author presents suggested revised Dolch Word Lists that reflect contemporary English use. The Dolch Word Lists were chosen for the (a) prevalence in primary grades pedagogical materials, (b) use in English medium primary grade classes throughout the world, and (c) public domain status of the lists. Given these reasons, revised and updated Dolch Word Lists are likely to be received well by materials designers and utilized by educators. Thus, the author hopes that educators and materials designers can use this information to engender more relevant and accurate educational experiences for English Language Learners in the primary grades.

Natalia Konstantinovskaia
UCLA

New functions of Japanese masculine sentence-final particles: a corpus study

Japanese feminine speech has attracted attention of many linguists, while masculine speech becomes the object of research much more rarely. Sturtz-Sreetharan (2004) examined Japanese masculine sentence-final particles *zo* and *ze*, which traditionally are considered to convey strong appeal (*ze*) and strong assertion (*zo*). Sturtz-Sreetharan came to a conclusion that **“**men use stereotypically masculine sentence-final particles infrequently” (275). Shibuya in her 6.5 hours of recording detected only three instances of *zo* (2004:122). However, there is lack of research of *zo* and *ze* in spoken and written corpora. The current study aims to fill this gap.Although based on the previous findings, the amount of zo/ze in spoken data is rather scarce, in the corpus of blogs they are utilized rather frequently. It contradicts the common notion that sentence final particles are commonly used for interactional purposes and, therefore, are restricted to conversations**.** In this research I aim to investigate the usages of *zo/ze* in blog corpus and compare them to conversational data**.** My hypothesis is that *zo* and *ze* have different functions in various types of discourse.I approach this by using three corpora: two corpora of conversations and one corpus of blogs. In particular, I examine the Japanese CallHome corpus that consists of 120 dialogs and Sakura corpus comprised of 18 face-to-face conversations. For the analysis of blogs, I use the Balanced Corpus of Contemporary Written Japanese. The obtained results show that *zo* and *ze* are often utilized by both male and female speakers. In many cases the goal is to construct an easy identifiable hyper-assertive character for comical effect, or to report male speech. Thus, current study demonstrates efficiency of corpus linguistics in description of the current functions of *zo* and *ze* in conversations and blogs. The results also contribute to the field of Japanese pragmatics and language acquisition.

Christian Koops & Arne Lohmann
University of New Mexico & University of Vienna

Interpreting discourse marker sequencing constraints: the case of English *so*

An important problem in discourse marker (DM) research is how to determine the status of DMs that remain phonologically identical to their sentence-level sources. For example, the English DMs *and*, *but*, and *so* are not easily distinguishable from coordinating or subordinating conjunctions. In this talk we argue that a promising solution to this problem is to draw on a largely unexplored behavior of DMs: their sequencing constraints relative to each other.

We first present quantitative evidence showing that two-part DM sequences, e.g. *and so* vs. *so and*, generally exhibit strong ordering preferences. The preferred order tends to be that which is predicted from the canonical order of the DMs' sentence-level counterparts. For example, DMs deriving from coordinators precede DMs deriving from subordinators (*and so* is significantly more frequent than *so and*). Next, we show that non-canonical orderings specifically reveal a form's DM use and thereby provide a window on its functional development.

Our analysis focuses on *so* in sequence with *and* or in sequence with *but*. All instances of the relevant ordering possibilities were extracted from the Fisher corpus (Cieri et al. 2004, 2005). We observe a sharp contrast between the function of *so* in each sequence. For example, *and so* is used to establish a local semantic connection between propositions in adjacent clauses (‘therefore’). *So and*, on the other hand, is used to manage superordinate topics as when speakers return to a larger question after a digression. In this sense, sequencing constraints formally reveal a DM's pragmatic scope expansion.

References

Cieri, Christopher, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004. *Fisher English Training Speech Part 1, Transcripts*. Philadelphia, PA: Linguistic Data Consortium.
Cieri, Christopher, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2005. *Fisher English Training Speech Part 2, Transcripts*. Philadelphia, PA: Linguistic Data Consortium.

Nicholas A. Lester
University of California, Santa Barbara

The comings and goings of *come* and *go* (and *move*)

Much work has been devoted to investigating the English deictic motion verbs come and go over the past several decades. These studies have primarily focused on unpacking the pragmatic principles governing the choice and acceptability of deictic verbs in context (e.g., Fillmore, 1997). However, no empirical study has yet examined the effects of these deictic properties on (1) optional path-encoding tendencies or (2) 'non-deictic' uses. I close this gap in the literature by innovating a
procedure to uncover how deictic and non-deictic predicates differ in path structure as instantiated by S(ource)-P(ath)-G(oal) PPs and directional adverbials.

I extracted all inflectional variants of COME(n=1,548), GO (n=1,810), and MOVE (n=507) from the Brown corpus (Francis & Kucera, 1979), without distinguishing deictic from non-deictic uses. I annotated all tokens for five variables: source preposition, path preposition, goal preposition, directional adverbials (e.g., up in He went up to the third floor), and overall path schema (e.g., G for He went up to the third floor). I then analyzed this data by combining a novel extension of the
Multiple-Distinctive Collexeme Analysis (MDCA; Gries & Stefanowitsch, 2004; Gilquin, 2006) with a hierarchical cluster analysis (HCA). The results reveal that come and go are distinct from move in their path encoding, which suggests that deictic pressures have altered the general clause-level behavior of come and go. I also introduce a post-hoc procedure to explore what variable levels contributed most to the groupings generated by the HCA. The results of this final step provide a
fine-grained outline of the path-encoding tendencies of come and go. I discuss these findings in terms of the constructional templates that unify all deictic and non-deictic uses of come and go, respectively, with a special focus on how and with what effect deictic features have been abstracted into generalized (pragmatically neutral) schemas. I also address implications for the acquisition of motion deixis.

References

Fillmore, C. J. (1997). Lectures on deixis. Stanford: CSLI.
Francis, W. N., & Kucera, H. (1979). Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, Rhode
Island: Department of Linguistics, Brown University.
Gilquin, G. (2006). The verb slot in causative constructions: Finding the best fit. Constructions, 1(3), 1, 46.
Gries, S. Th. & Stefanowitsch, A. (2004). Extending collostructional analysis: a corpus-based perspective

on ‘alternations.’ International Journal of Corpus Linguistics, 9(1), 97-129.

Don Miller
California State University, Stanislaus

Lexical Diversity, Sophistication, and Error in Generation 1.5 Writing

Over the past decade, a great deal of attention has been given to the goal of understanding the unique academic literacy needs of an increasing proportion of university students: long-term U.S. resident English language learners, commonly known as Generation 1.5 students (e.g., Forrest, 2006; Harklau, 2003; Roberge, 2002; 2009, Singhal, 2004). Among the exceptionally complex network of skills required for successful academic writing is the ability of writers to accurately use a variety of sophisticated, “academic” vocabulary (e.g., Academic Word List vocabulary, Coxhead, 2000). Indeed, previous research has demonstrated the contribution of such vocabulary to assessment of overall writing quality (Laufer & Nation, 1995; Olinghouse & Leaird, 2009). Thus, a focus on the lexical diversity and sophistication employed by these developing writers, as well as the types of vocabulary-related error that they produce, will deepen our understanding of the complex the nature of the academic literacy gap that they face.

This research talk summarizes a comparative analysis of vocabulary use (i.e., lexical diversity, lexical sophistication, and vocabulary-related error) by native English speakers and Generation 1.5 writers on a university-level writing proficiency exam. Implications for educators who work with Generation 1.5 writers will be discussed.

Katherine Moran
Georgia State University

Discovering the incongruities in undergraduate writing in chemistry and psychology through the methodological integration of multidimensional analysis and qualitative interviews

This presentation will report the results of a study using multidimensional analysis (Biber, 1988), along with interviews with instructors and students, and a taxonomy of writing tasks and course syllabi to develop a more complete picture of undergraduate disciplinary writing in chemistry and psychology. Multidimensional analysis, based on the four dimensions formulated by Gray (2011) (academic involvement and elaboration vs. informational density, contexualized narration vs. procedural discourse, human versus non-human focus, and “academese”), is used to describe the linguistic differences between what undergraduates read in their major courses and the writing they produce in the upper division. The taxonomy of writing assignments gives an additional lens through which to interpret the results of the multidimensional analysis. These results are considered in conjunction with the analysis of student and faculty interviews which focus on writing expectations and writing experiences.

Results indicate that student writing tends to be more informationally dense than what students are reading, is more inclined towards procedural discourse rather than narration, and demonstrates less explicitly empirical language. There are also distinctions between the linguistic behavior of each discipline. While psychology shows more features of involvement, narration, an explicitly empirical stance, and has a clear human focus, chemistry prefers language that is informationally dense, procedural, less explicitly empirical, and has a non-human focus. In addition, the findings show that undergraduates write most late in their academic careers, that professors and students often have misaligned ideas of writing expectations, and that expectations can be idiosyncratic even for similar assignment types.

Cynthia M. Murphy & Scott Crossley
Georgia State University

Co-occurring linguistic features in an L2 writing corpus: Insights into prompt effects and successful writing

This study uses a Multidimensional Analysis to reveal patterns of co-occurring linguistic features in nonnative English speakers' writing across three parameters: prompt (*N* = 4), text type (expository writing vs. letter writing), and holistic score. The corpus used in this study comprised 800 English essays written by L1 Cantonese high school students in response to one of four prompts (Milton, 2000). All essays were assigned scores by trained raters. The corpus was analyzed using Coh-Metrix (Graesser et al., 2004), a computational tool that measures a text's characteristics in relation to linguistic, discourse, and conceptual representations. Results reveal that sets of co-occurring linguistic features related to word information and negation reliably distinguished essays according to writing quality, indicating that better writers exhibited greater lexical sophistication and less negation over all. Results also show that patterns of co-occurrence related to lexical and semantic overlap distinguished essays by prompt topic, suggesting that certain topics may elicit greater lexical and semantic overlap. Finally, the results indicate a text type dimension that divided essays according to co-occurring linguistic features related to word frequency, with expository writing demonstrating far less frequent language than letter writing. This finding suggests that task type may influence the production of linguistic features. Results carry implications for L2 assessment and L2 writing pedagogy.

Amanda C. Murphy
Università Cattolica del Sacro Cuore

Evidence of rewriting: food bundles in different text types

As part of a larger project on the language of food, and taking current theoretical models of rewriting into consideration (e.g. Cucchi and Ulrych 2009) this paper investigates 'food bundles', defined as "a lexical bundle referring to food", in the different registers found in the transcription of the film *Julie and Julia* (Ephron 2009), in an episode lasting 30 minutes of *The French Chef,* Julia Child's television series (4000 words), and in a corpus of food blogs (154,000 words). The TV series was based on Child et al.'s classic cookbook "Mastering the Art of French Cooking" (1961), and the film took inspiration both from the cookbook and from Julie Powell's accounts of how she cooked her way through the book, which can still be read in online extracts from her blog 'the Julie and Julia project'. The complex genesis of the film, in particular, is reflected in multiple modes, registers and levels of narration. The paper investigates food bundles in several registers, including face-to-face conversation, monologues, television programs both of an instructional and ludic nature, and blogs. The forms of the lexical food bundles, retrieved by corpus-driven

criteria, are first illustrated, and their functions subsequently analysed, with particular regard to the interpersonal function (Fischer 2013), seen in terms of involved production (Biber 1988) and "positively marked emotional terms" (Diemner and Frobenius 2013), as a conflation of the linguistic expression of evaluation (Hunston 2011) and of emotion (Bednarek 2008).

Bednarek, M. 2008. *Emotion Talk Across Corpora*. New York: Palmgrave Macmillan.
Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: CUP.
Child, J., Beck, S. & Bertholle, L. (1961). Mastering the Art of French Cooking. Volume 1. New York: Alfred A. Knopf.
Cucchi, C. and M. Ulrych, 2009. "Translation, rewriting and recontextualization: forms of mediated discourse" in M. Bertuccelli Papi, A. Bertacca, S. Bruti eds., Threads in the Complex Fabric of Language, Pisa, Felici Editore, 139-170.
Diemer Stefan/ Frobenius, Maximiliane. 2003."When making pie, all ingredients must be chilled. Including you. Lexical, syntactic and interactive features in online discourse- a synchronic study of food blogs" in Culinary Linguistics (Gerhardt,C. Frobenius, M. and Ley, S.eds.). Amsterdam: John Benjamins.
Ephron, N. (2009). Julie & Julia. USA: Easy There Tiger Productions & Scott Rudin Productions
Fischer, Kerstin 2013. The addressee in the recipe in Culinary Linguistics (Gerhardt,C. Frobenius, M. and Ley, S.eds.). Amsterdam: John Benjamins Publishing Company, 103-117.
Hunston, Susan 2011. Corpus Approaches to Evaluation. Phraseology and Evaluative Language. London: Routledge.

Susan Nacey
Hedmark University College, Norway

Error versus creativity: The metaphorical language of Norwegian L2 English learners

This paper explores the dividing line between 'difference' and 'deficiency' in the written language of Norwegian EFL learners by focusing on the complex concept of metaphorical creativity and its identification (see e.g. Nacey 2013: 157-203). Creativity merges the known with the familiar; metaphor —according to cognitive theorists— links disparate semantic domains to illuminate a less familiar (often abstract) concept in terms of a more familiar (more concrete and/or embodied) concept (see e.g. Steen 2011). The products of the creative process are new, and in some sense extraordinary. The prototypical metaphor —according to the traditional view— is vibrant and novel, provoking new insight (see e.g. Black 1981). Metaphor and creativity would thus seem to go hand in hand. Indeed, L2 language users, who per definition have access to two or more languages, may also produce manifestations of 'bilinguals' creativity' resulting from the 'mixing' of languages (Kachru 1985; Kumaravadivelu 1988, p. 313 and 316).

This corpus-based study examines all occurrences of metaphorical language in roughly 20,000 words of argumentative texts written by advanced Norwegian students of English that meet (at least) one of three oft-mentioned criteria of creativity: novelty, significance (i.e. the deliberate 'crafting' of language), and appropriateness (i.e. intelligibility) (Boden 2004: 43; Cameron 2011; Kövecses 2010: 664; Pitzl 2009, 2012; Semino 2011; Steen 2008);

the potential role of the L1 is also investigated. The overarching goal is an evaluation of these criteria as valid measures of creativity, in an attempt to tease apart "what looks like a mistake but is in fact poetry" (McArthur, cited in Rubdy & Saraceni 2006: 23) in L2 learner language.

David Oakey
Iowa State University

A Comparison of Recent Corpus-Derived Phraseological Lists for Pedagogy

It has long been recognized that phraseology comprises a significant proportion of English usage, and that proficient learners successfully acquire phraseological forms and their meanings and develop an awareness of the appropriate registers in which to use them. In the field of English for Academic Purposes (EAP) several recent studies have produced lists of phraseological forms, backed by large-scale corpus evidence, which are intended for use in EAP pedagogy. New terms have been coined for these forms and existing terms have been appropriated: 'lexical bundles' (Biber et al. 1999; 2004; Hyland 2008; 2012), 'collocations' (Durrant 2009; Ackermann and Chen 2011), 'academic formulas' (Simpson-Vlach and Ellis 2010), 'multi-word constructions' (Liu 2012), and 'phrasal expressions' (Martinez and Schmitt 2012). Confronted with so many lists and so many terms, however, the teacher or materials designer may understandably find it difficult to select a particular list or combine items from different lists into their course syllabus and materials.

This paper consequently aims to clarify for EAP practitioners this recent work on phraseological items by reviewing and comparing these recently published lists. It uses a comparative framework which draws on "Eastern European" lexicography (Aisenstadt 1981; Howarth 1996), "Empirical Firthian" lexicology (Stubbs 2001), and "Usage-Based" cognitive linguistics (Gries 2008) to explore syntactic, pragmatic, semantic, lexical, and methodological reasons for the differences between them. It then discusses how serious these differences are likely to be in practice for EAP learners, and makes suggestions to assist EAP teachers and materials developers in selecting items for inclusion in the syllabus.

Aurore Paligot
F.R.S. – FNRS and University of Namur

Signing styles of French Belgian Sign Language (LSFB): investigating the role of audience and interaction

Recent technological advances paved the way for the creation of an increasing number of sign language (SL) corpora, thus bringing the focus around to the description of natural language use. While numerous studies have shed light on the influence of external and internal factors on SL variation (e.g. Lucas and Bayley, 2010), the description of SL variation according to situational contexts – phenomenon which is referred to as "register" or "stylistic" variation (Biber, 1998; Schilling-Estes, 2002) – has been little studied to date (Zimmer, 1989; Quinto-Pozos and Mehta, 2010).
This paper is part of a PhD project that aims to describe register variation in French Belgian Sign Language (LSFB). Comparing the productions of two deaf signers in four different settings (an online video, a conference, a course and a dialogue), we examine the influence of the audience and the interaction on the usage levels of phonological features that are related to formality. As showed by Paligot and Meurant in a previous study (2013), some features (e.g. location symmetry of two-handed signs) are associated with formal settings

while others (e.g. sign dropping) are characteristics of informal situations. Based on a multivariate analysis conducted with Rbrul (Bayley, 2002), this paper determines how the use of these features varies across settings with different audiences and amounts of interaction. This study contributes to give better insights into the organization of registers of LSFB along the formal to informal continuum and reveals some phonological variables that are shared by SL in general.

Magali Paquot
FNRS Université catholique de Louvain

An integrated approach to phraseology in EFL learner writing

Phraseological studies of English as a Foreign Language (EFL) learner writing have been particularly numerous and have generated a wealth of interesting results (Paquot & Granger, 2012). However, they have usually presented a one-sided view of the field – adopting either a traditional or frequency-based approach to phraseology (Granger & Paquot, 2008). Studies that have adopted the latter approach also focus either on co-occurrence or recurrence phenomena. There is clearly an overlap between the three approaches but recent research has failed to integrate them into a sound analytical framework.

The main objective of this presentation is to compare the three approaches on the same learner corpus dataset and investigate how they can be used to develop an integrated approach to phraseology in foreign language learning. I report on a threefold study of word combinations in learner writing: (1) a 'phraseological' analysis (in its most traditional sense) that distinguishes between free combinations, collocations and idioms; (2) a co-occurrence analysis that uses the t-score and the MI to identify statistically salient collocations; and (3) a macro analysis centred on the routine aspects of learner language that investigates the frequency, structure and function of lexical bundles.

The learner corpus data used come from the *Varieties of English for Specific Purposes dAtabase* (VESPA, https://www.uclouvain.be/en-cecl-vespa.html) and consist of 123 research papers written by French EFL learners in the context of BA and MA linguistics courses (c. 370,000 words).

Seonmin Park
Northern Arizona University

Methodology for a Reliable Academic Vocabulary List

Vocabulary is one of the crucial factors for students' academic comprehension (Anderson, 2008; Grabe, 2004 & 2009; Laufer, 1992; Nation, 2001; Qian, 2002). Thus, researchers (Coxhead, 2000; Garnder & Davis, 2013) have created vocabulary lists such as Academic Word List (AWL) and Academic Vocabulary List (AVL) which other researchers and material developers of English as a second language (ESL) and English for specific purpose (ESP) have implemented. Although the academic vocabulary lists have been used widely for research and language teaching, few studies probe into the reliability of the lists. Therefore, this study investigates the effect of word selection criteria on the reliability of an academic vocabulary list. Three questions are addressed: 1) which set of criteria including ratio, range, and discipline measure and discipline dispersion would extract a stable academic vocabulary list? 2) would all criteria be necessary to build a reliable vocabulary list? 3) would textual dispersion influence reliability of a vocabulary list? A 1.7-miliion-word corpus is created with 180 academic articles across six disciplines. The corpus is divided into two sub-corpora and sixteen criteria sets are applied to each sub-corpus for vocabulary list creation. The replicability of the vocabulary lists is examined to find a set of criteria extracting the most reliable vocabulary list. The results show that discipline dispersion and textual dispersion are decisive factors in the reliability of an academic vocabulary list. The study indicates that word selection criteria should be carefully considered in order to produce a reliable and replicable academic vocabulary list.

Geoffrey G. Pinchbeck
University of Calgary

An L1 adolescent learner corpus**:** academic success and lexical competence in grade 12 expository writing

This presentation will examine the relationship between vocabulary use and academic success in mainstream, academic-track grade 12 English Language Arts (ELA) classrooms. There have been recent calls for academic language to be given a more prominent role in mainstream public educational planning across the curricula in Canada and in the U.S. This project is impeded in part by the limitations of existing vocabulary assessment tools. Working towards the development of an academic lexical syllabus component within the mainstream K-12 secondary curricular framework, we hope to refine and operationalize the construct of the lexical component of general academic language within Canadian secondary education settings. We are in the process of compiling a >2,000,000-word, grade-12-student written corpus from a large random sample of essays from a government-administered diploma ELA exam. Lexical indexes of frequency and diversity will be aligned with those of reference corpora of adult British and American English as well as an American K-12 textbook and reader corpus. Vocabulary profiles will then be compared to the following associated data: 1) official provincial exam essay scores (holistic rubric scoring), 2) writing error data using a detailed coded rubric, and 3) student high-school transcripts. Using a regression approach, we identify a domain of mid-frequency vocabulary that explains significant and unique variance of both essay quality and general academic success. We present how this research might be used to develop tools to monitor English academic literacy development for diagnostic purposes and to inform a strategic K-12 academic language pedagogy.

Jed Sam Pizarro-Guevara
UC Santa Cruz

The distributional information of relative clauses in child-directed Tagalog

This present study is a part of a larger study on the acquisition of relative clauses (RC) in Tagalog. In order to have an ecologically valid assessment of children's RC-comprehension, a 30,537-word corpus from 25 children storybooks was constructed and used to determine distributional information of the structure. Relative clauses involving transitive verbs were extracted and annotated based on the identity of the nominal contained therein. Results indicated that out of the 267 RCs involving transitives, 110 (41.20%) were attenuated forms, that is, either as null or as pronominals; 139 (52.06%) were lexical nouns; and 18 (6.74%) were "other" (i.e., sentential complements). While these findings would suggest that the nominals inside the RCs are almost equally attenuated, as they are lexical, it is important to note the effect of the genre of the source texts. Storybooks tend to use over-lexicalization. Nonetheless, these findings are pivotal for the ensuing experimental study because it directed the experiment by showing the distributions of embedded NPs in RCs, so that the stimuli were representative of what children hear, or at least, attempted to maximize the items' authenticity, thereby eliminating as a confound the mismatch between the distributional patterns of their input and what they were tested on.

Ute Römer, Audrey Roberson, Nick Ellis & Matthew Brook O'Donnell
Georgia State University, University of Michigan & University of Pennsylvania

Combining learner corpus and experimental data in studying L2 learner knowledge of verb-argument constructions

There has been a recent increase in studies that highlight the value of combining corpus and experimental evidence in the study of linguistic phenomena (e.g. Ellis and Simpson-Vlach, 2009; Gilquin and Gries, 2009; Wulff, 2009). These studies tend to utilize corpora of native speaker output in combination with speaker judgments collected in experimental settings and demonstrate that different types of data can present converging evidence which helps strengthen research hypotheses. The present paper discusses how experimental data and corpora which capture second language learner (rather than native speaker) output complement each other in providing insights into learner knowledge of 20 different verb-argument constructions (VACs), including the 'V *about* n' construction (as exemplified by *she thinks about chocolate a lot*). Advanced learners' (L1 backgrounds German and Spanish) dominant verb-VAC associations are examined based on evidence retrieved from the German and Spanish subcomponents of ICLE (the International Corpus of Learner English) and LINDSEI (the Louvain International Database of Spoken English Interlanguage) and verbs collected in lexical production tasks in which participants complete VAC frames, such as, 'he ___ about the...' (with e.g. *talked*, *thought*, *wondered*). Native speakers, L1 German and L1 Spanish participants were asked to generate as many verbs as possible for a given frame over a span of 60 seconds. We compare findings from the different data sets and consider the strengths and limitations of each in relation to questions in usage-based language acquisition and Construction Grammar, thus demonstrating the value of linking corpus and experimental evidence. [247 words]

References

Ellis, Nick C. and Rita Simpson-Vlach. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5 (1): 61-78.
Gilquin, Gaëtanelle and Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5 (1): 1-26.
Wulff, Stefanie. 2009. Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory* 5 (1): 131-159.

Margo Russell
Portland State University

Linguistic Features in ELL and L1 University Student Writing: Revisiting Hinkel's Findings

Academic writing is a difficult skill to master regardless of a student's first language. To better prepare ELL students for success in mainstream content courses, instructors need to know more about the characteristics of their students' writing. Knowing which linguistic

features to target for intervention allows instructors to help students produce writing that is appropriate for the academic written register.
Two corpora of 30 research essays each, one of undergraduate L1 writing and the other of advanced ELL writing produced in an intensive English program, were compared for the frequency and function of 13 linguistic features previously found in different frequencies between L1 and ELL essays (Hinkel, 2002). The Mann-Whitney U test corroborated Hinkel's findings that L1 essays contained significantly higher frequencies of features associated with academic writing: modal *would,* perfect aspect, passive voice, reduced adjective clause, *it*-cleft, and type/token ratio. Contrary to Hinkel's findings, ELL essays did not contain higher frequencies of the features typical of conversation.

An analysis of how each significantly different feature was used in the essays revealed that ELL students were still acquiring grammatical accuracy and the uses appropriate for the academic written register. A surprising finding was that L1 students struggled with using modal *would* appropriately for academic writing*.*

To help students raise their awareness of these features, teachers should lead students in identifying grammatical uses and differentiating between uses which are standard to academic writing and those which are appropriate only in conversation. Two sample activities illustrate how to implement these recommendations.

Mike Scott
Aston University

Mapping Dickens

This paper tackles an issue in studying key words (KWs) concerning the stretches of text they apply to and are key in. It has been clear for years that some KWs are global and others localised or bursty (Katz, 1996) in their applicability to a text but the way they pattern and how they relate to each other, forming what one might think of as KW constellations is of particular interest. In the case of Dickens' novels it has also long been known that characters are signalled by mannerisms or appearance so that the reader may easily recognise them after many intervening pages. For example in *Bleak House* when Lord Dedlock is in scene, his gout usually makes its appearance; in *Great Expectations* 'the Aged' is linked with Wemmick, Mr Pancks in *Little Dorrit* is usually compared with or even represented as a steam engine and Major Bagstock In *Dombey and Son* always refers to himself in the third person as *Joe B* or variants thereof. In this study we consider a number of Dickens novels and examine ways of plotting and linking such related KW bursts. A Dickens novel is characteristically a patchwork where the focus moves from theme to illustration to any one of numerous sub-plots: the focus in this paper is on mapping out the changes and the linkages between pieces of the patchwork.

Katz, S. 1996. Distribution of Common Words and Phrases in Text and Language Modelling, *Natural Language Engineering* 2 (1): 15-59.

Stephen Skalicky & Scott Crossley

Georgia State University

Satirical irony in Amazon.com product reviews

Satire is a form of humor that uses irony as a means to construct an incongruous reality between what is said and what is meant (Simpson, 2003). While related forms of language, such as sarcasm, have received scholarly attention (e.g., Campbell & Katz, 2012), satire is still relatively understudied. Thus, better understanding satire linguistically would complement current theoretical definitions. This study addresses this need by performing a computational, linguistic analysis of written satire in an online corpus of satirical Amazon.com product reviews.

The corpus consists of 750 product reviews taken from Amazon.com. Half of these reviews were identified as satirical while the other half were identified as non-satirical. Based on previous findings from computational analyses of sarcasm, the corpus was analyzed for measures of lexical sophistication, grammatical functions, and the semantic properties of words using the Linguistic Inquiry and Word Count (LIWC) tool and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). Our purpose was to test if a measurable difference exists between satirical and non-satirical texts in our corpus using these linguistic indices.

Initial multivariate analyses of variance demonstrated that a significant difference exists between satirical and non-satirical product reviews. Follow up discriminant function analyses demonstrated the satirical texts are more specific, less lexically sophisticated, and contain more words associated with negative emotions and certainty. These results suggest that satire relies on specific, measurable linguistic strategies when compared to non-satire. The results also contribute to a data-driven definition of satire.

Shelley Staples
Purdue University

Triangulating Data in Corpus-Based Discourse Analysis: Using Corpus, Assessment, and Interview Data to Better Understand a Discourse Domain

Quantitative corpus-based approaches to discourse analysis can be usefully triangulated with other data (e.g., assessment and interview data) and qualitative methods to better understand the characteristics of a discourse domain. Such mixed methods approaches are becoming more popular in corpus-based studies (see e.g., Conrad, 2013; McGrath & Kuteeva, 2012) but are still relatively rare. This presentation will focus on a case study involving three methods used to analyze nurse-patient interactions. This mixed-method approach can be applied to other research areas.

First, a quantitative *and* qualitative corpus-based analysis is used to compare the linguistic features used by international and U.S. nurses in their interactions with patients. Second, I examine the relationship between the linguistic features used by nurses and two assessments of effectiveness (a patient satisfaction measure and an interpersonal skills assessment). Finally, interviews with international nurses are used to better understand the results from the first two approaches.

The quantitative corpus-based analysis reveals key differences in the use of linguistic features by international and U.S. nurses (e.g., backchannels, 1st person pronouns, pitch range). Analysis of the assessment data provides evidence that there is a relationship between the use of particular features (e.g., 1st person pronouns) and more effective interactions. Finally, the qualitative discourse analysis and nurse interviews offer insight into the different patterns of use, suggesting that the variation across nurse groups and more/less effective interactions is related to different approaches to patient care. Taken together, this triangulation of the data offers a richer understanding of the discourse domain.

Geoffrey T. LaFlair, Jesse Egbert & Shelley Staples,
Northern Arizona University, Brigham Young University & Purdue University

Comparing Oral Proficiency Interviews to Academic and Professional Spoken Registers

The use of oral proficiency interviews (OPIs) to measure speaking ability has been rationalized by the argument that they mirror aspects of interactive spoken discourse. Although discourse analysts have examined OPIs since the 1990s (see, e.g., He & Young, 1998), few studies have quantitatively examined the linguistic features of OPIs in relation to conversation and other spoken registers. This study investigates linguistic features used in one such OPI, the Michigan English Language Assessment Battery (MELAB) speaking assessment. The corpus used for this study comprises a sample of 100 MELAB OPIs from 2013 that were administered by 20 different CaMLA certified examiners and rated after each interview. We compare the frequency of lexico-grammatical features used in this corpus to their frequency in the Longman Corpus of American Conversation, nurse-patient interviews and spoken university encounters. These registers reflect the major purposes for taking the MELAB OPI: academic entry and professional certification (e.g., nursing). The lexico-grammatical features we examine (e.g., personal pronouns, modals, and discourse markers) have been found to be important characteristics of spoken discourse (see, e.g., Biber, 2006; Biber, Johansson, Leech, Conrad, & Finegan, 1999). We expect that there will be important differences in the use of lexico-grammatical features between the MELAB OPI and other spoken registers. These findings contribute to our understanding of the linguistic characteristics of OPIs and the extent to which test taker speech reflects interactive spoken discourse.

Lize Terblanche
Northern Arizona University

Navigating rocky terrain: Measuring complexity in L2 student writing

Complexity in L2 student writing has been analyzed using different measures, mostly on the clausal rather than the phrasal level. It is commonly assumed that grammatical complexity develops as students progress in their L2, but the linguistic features associated with the process are not yet fully understood. For this paper, I will analyze grammatical features that are hypothesized to be indicative of a developmental index in L2 writing proposed by Biber, Gray, and Poonpon (2011). The purpose is to test whether L2 students at high versus low proficiency levels use complex grammatical structures in different ways. I will use a hypothesis-testing approach to analyze complexity features in a small corpus of placement essays (54,000 words) written by freshmen at a private midwestern university. The dependent variables are essays at high and low proficiency levels. The independent variables are 20 linguistic features that are grouped into three grammatical types: nonfinite dependent clauses, finite dependent clauses, and dependent phrases. The complexity features will be automatically tagged (Biber, 1988, 1995) and the frequencies per text will be normed. Next, means and standard deviations will be calculated per feature. To test the statistical significance of the results, *t* tests will be used. The results will be compared to Biber et al.'s (2011) findings for academic writing, as well as Parkinson and Musgrave's (2014) results for L2 graduate student writers.

Paul Thompson, Susan Hunston, Akira Murakami, Dominik Vajn & Douglas Biber
University of Birmingham & Northern Arizona University

Talking over the academic garden fence: a multidimensional perspective on interdisciplinary research discourse

This paper reports first results from a large-scale corpus investigation of interdisciplinary research discourse. Research into disciplinary discourses has a relatively long history, and most such research presupposes disciplinary differences, identifying contrasts between the cultures and the discourses of separate physical and social sciences (e.g., Hyland, 2000; Biber, 2006). These comparisons reflect traditional institutional distinctions. However, with the recent growth in research that crosses disciplinary boundaries, the question arises as to whether the discourse of such interdisciplinary research is similar to or different from that of traditional disciplines.

To address this question, the present study, part of a larger ESRC-funded project (ES/K007300/1), employs Biber's (1988) multidimensional (MD) analysis. The corpus used for this study consists of all the research papers published in 11 international journals over a period of 10 years (2001-2010). Access to the full holdings of the journals has been provided to the project team by the scientific publisher, Elsevier. Five of the journals are mono-disciplinary, and the rest are interdisciplinary. The aim of the MD analysis is to examine the differences between mono-disciplinary and interdisciplinary discourse from the viewpoint of "dimensions" of variation of linguistic features. The analysis tests whether interdisciplinary research discourse is quantitatively different from mono-disciplinary research discourse.

In the second part of this paper, we examine more closely the full holdings (1990-2010) of one highly successful interdisciplinary journal, Global Environmental Change, and report observations on how an interdisciplinary field, as a discourse object, emerges in measurable ways over time.

Nicole Tracy-Ventura, Kevin McManus & Rosamond Mitchell
University of South Florida, University of York & University of Southampton

The development of lexical diversity during study abroad: Introducing the new LANG-SNAP longitudinal learner corpus

This presentation will introduce a new longitudinal learner corpus that was compiled as part of a large-scale project on the L2 acquisition of French and Spanish before, during, and after a 9- month stay abroad. The learner corpus (approximately 625,000 words) is comprised of oral and written data collected six times over 20 months, including three visits whilst abroad. Participants were university students of L2 Spanish (n=27) and L2 French (n=29) spending their third year of a four-year degree in either France, Spain or Mexico. The same data were also collected from native speakers (n=10 for each language). Oral and written data were formatted in CHAT for use with the CLAN program (see MacWhinney, 2000). In addition to presenting the design of the corpus, we present a case-study demonstrating how the new longitudinal learner corpus can be used to investigate language learning during residence abroad. We investigate development of lexical diversity using D (Malvern & Richards, 2002; computable in CLAN) in three different registers: 1) an oral interview, 2) an oral picture-based narrative, and 3) a written argumentative essay.

Participants completed all three activities at each data collection cycle allowing us to investigate change in scores over time as well as compare learners' performance across registers. Results demonstrate that both sets of learners show significant gains in lexical diversity in their oral production from early on in their stay abroad whereas development in written production is slower to develop. Register differences were also apparent at each data collection cycle.

Alfredo Urzúa
San Diego State University

Exemplification and reformulation in learner writing at different levels of proficiency

In academic writing, elaboration strategies such as exemplification and reformulation constitute key rhetorical functions that enhance coherence and convey a writer's sense of audience (Hyland, 2007). Not surprisingly, then, English language teachers spend considerable time reminding students to elaborate their ideas and provide specific examples to support their arguments while composition textbooks highlight the importance of elaboration. Given this situation, researchers are paying increased attention to learners' exemplification and reformulation strategies, with findings indicating that their writing tends to exhibit a limited repertoire of expressions, e.g., learners overuse the expression for example while underusing alternative expressions (Gilquin, Granger, & Paquot, 2007). However, little information exists on how these metadiscoursal expressions evolve (or not) as students develop their linguistic and writing skills.

This presentation reports on a corpus-based analysis of exemplification and reformulation strategies used by college-level English language learners. Using quasi-longitudinal data from the ULCAE corpus, a context-specific learner corpus of written English, quantitative and qualitative data on the use of exemplification and reformulation expressions (e.g., for instance, in other words) are analyzed and compared across four levels of proficiency (from low-beginning to high-intermediate) in an ESL/EAP program housed at a mid-size, public university located along the US-Mexico border. The data consist of 348 essays (approximately 250,000 words), generated by mostly Spanish-speaking students, reflecting different writing tasks (from five-paragraph essays to problem-based writing and research reports). The presenter discusses the findings in light of the learners' English language curriculum and textbook information, as well as implications for instruction.

Elaine Vaughan & Brian Clancy
University of Limerick & Mary Immaculate College

The devil is in the detail: Using corpora to investigate spoken language varieties

Comparing spoken corpora of different language varieties affords insights into not only the lexico-grammatical features of those varieties, but also their pragmatic systems (e.g. O'Keeffe & Adolphs 2008). The most frequent items in wordlists tend to be 'small' items, pronouns, determiners and the like. Questioning further the ways in which high-frequency functional items are used, particularly if they occur in differing proportions in different corpora, can provide insights both intuited and unexpected about language varieties. Hence, 'the devil is in the detail'. This paper focuses on corpora of Irish English, notably the Limerick Corpus of Irish English, a one-million-word corpus of primarily casual conversation, in order to launch a comparative investigation into the varietal nuances of items which fall into traditional deictic categories, such as temporal (e.g. now) and spatial (e.g. there) deixis. Clancy & Vaughan's (2012) investigation of now highlighted a pragmatic function in clause-final position which occurred more frequently in the Irish than British datasets used. Now was found to additionally function in Irish English as a pragmatic marker, softening the impact of negative evaluations or judgements, and as a deictic presentative, akin to the French violà. This paper investigates the linguistic behaviour of another stalwart of the higher reaches of corpus frequency lists, there. Similarly to now, a nuanced investigation of there unearths a potential varietal idiosyncrasy. We know that there functions existentially and as a spatial deictic marker. However, corpus findings from LCIE also suggest a distinct function for there: while it does function existentially and spatially, it also has what appears to be a temporal function.

Clancy, B. and Vaughan, E. 2012. "It's lunacy now" A corpus-based pragmatic analysis of the use of 'now' in Irish English." In Migge, B. and Ní Chosáin, M. (eds.) New Perspectives on Irish English. Amsterdam: John Benjamins, 225-245.
O'Keeffe, A. and Adolphs, S. 2008. Using a corpus to look at variational pragmatics: response tokens in British and Irish discourse. In Schneider, K.P. and Barron, A. (eds.) Variational Pragmatics. Amsterdam: John Benjamins, 69-98.

Elaine W. Vine
Victoria University of Wellington

'that': usage and pedagogy

'that' is important pedagogically for at least two reasons: 1) it is a very high frequency word, ranking 7th in the BNC, and 2) it is category-ambiguous, i.e. the one word-form has different grammatical uses. For example, 'that' is used as complementiser, demonstrative pronoun, demonstrative determiner, relative pronoun, adverb and conjunction in the Wellington Corpus of Spoken NZ English.

Research questions:
What are the patterns and frequencies of use of 'that' in general English, learner English and ELT coursebooks?
To what extent are the corpus findings reflected in pedagogical applications in coursebooks?

I report on comparisons of the forms and functions of 'that' in corpora of British and NZ English, learner English, and ELT coursebooks. I have found variation in patterns of use within and across corpora, for example, 'that' is used most frequently as complementiser in all corpora except spoken NZ English and spoken English of L1 young adults. However, in the New Headway coursebook series, 'that' is taught explicitly as demonstrative pronoun in Book 1, and as demonstrative determiner in Book 2, yet in the explanations of both, 'that' is used as relative pronoun, which is not taught explicitly until Book 3. The most frequent use of 'that' in the general and learner corpora – as complementiser – is not taught explicitly in the New Headway series, yet it is used in all six books in the series. Such findings give rise to a concluding discussion of issues around the relevance and application of corpus findings in pedagogical contexts.

Xiaoying Wang & Yunhua Qu
Zhejiang University

A Corpus-driven Study: Chinese Lexical Frames in Conversation and Academic prose

While past few decades has witnessed a number of researches in continuous sequences of words (Biber et al., 2004; Ellis, 1996; Howarth, 1998; Moon, 1997; and Wray, 2002), discontinuous sequences, as a more prevailing lexical pattern, should gain more attention. Gray and Biber (2013) first explored English lexical frames which consist of words and a variable slot in academic prose and conversation. However, almost no researches have systematically investigated Chinese discontinuous patterns in different registers.

Building on previous studies in other languages, this paper aims to fill the research gap in Chinese through direct computational analysis based on Zhejiang University Corpus of Spoken and Written Mandarin Chinese (which contains 1,000,000 Chinese words and 13 sub-registers). Following Gray and Biber's (2013) classification of frames, this paper tries to provide answers to these questions: (1) how do 4-word frames distribute in academic prose and conversation? (2) how is the variability degree of frames in different registers? (3) what are the functions of the frames in discourse?

The findings turn to be different from what have been found in English. The results indicate some impressive features of lexical frames in Chinese: (1) frames found in spoken and written Chinese are significantly less than English frames; (2) there are more frames discovered in spoken register; (3) the function-word frames are least prevalent in both registers; (4) frames demonstrate a lower level of variability in spoken texts. The findings should shed light on Chinese phraseological studies and provide more evidence to the difference between isolating language and inflectional language. The results should also imply different importance and function of lexical frames in Chinese.

Heidi Wright
Northern Arizona University

Stance Features within Stand-Alone Literature Reviews and Research Articles: An Interdisciplinary Register Analysis

Research articles (RAs) and stand-alone literature reviews (i.e., reviews published separately from a research article, hereafter LRs) are important, related genres (or sub-registers) within the register of academic prose (Swales, 2004). While both present the "facts" of new and previous research, scholars debate how overtly persuasive or evaluative these texts are. To date, studies of the register features tied to evaluation or "stance" (Biber & Finegan, 1988) within RAs exist for many disciplines. However, the few empirical studies of LRs focus primarily on genre conventions rather than grammatical features and limit themselves to a single discipline. To address this gap in the literature, I conducted an interdisciplinary register analysis of stance features in RAs and LRs. The corpora created contained 80 RAs and 68 LRs taken from 4 disciplines: applied linguistics, psychology, geology/planetary sciences, and education. Using a series of ANOVA's with sub-register, discipline, and the presence/absence of a methods section as independent variables, the corpora were examined for the prevalence of 19 grammatical markers of stance. Findings revealed that LR's and RAs share a common set of prevalent stance features. With regard to differences by sub-register, LRs exhibited statistically significant differences and higher means for only 6 stance features. In contrast, the ANOVA by discipline revealed statistically significant differences for 10 stance features with significant differences often occurring between the natural and social sciences. The presence of a methods section further complicated the picture, suggesting that many factors influence the quantity and types of stance features academic texts exhibit.

Stefanie Wulff
University of Florida

Are L2 learners sensitive to register differences? A case study of complementizer variation in German and Spanish L2 English

This paper examines the variable realization of the complementizer THAT in English object-, subject-, and adjectival complement constructions as in (1).

(1) a. Maria thought (that) Nick likes chocolate.
b. The problem is (that) Nick doesn't like chocolate.
c. I'm glad (that) Maria likes chocolate.

While native speakers' choices have been researched intensively, comparatively little is known about what drives L2 learners' decision to realize or omit the complementizer. More specifically, no study to date has examined whether learners are sensitive to register differences like native speakers have been argued to be (Torres Cacoullos & Walker 2009).

The present study therefore presents a contrastive corpus-based analysis of THAT-variation in native English speakers and German and Spanish L2 English. 9,445 were retrieved from native and intermediate-advanced level learner English spoken and written corpora (the International Corpus of English; the International Corpus of Learner English; and the Louvain International Database of Spoken English Interlanguage). 12 predictors were coded for, crucially including L1 background and register. A minimal adequate binary logistic regression model (Log-likelihood=5059.45; df=28; p=0) predicts all speakers' choices very well (Nagelkerke's R2=.55; C=.88; prediction accuracy=80.63%). The results suggest that (i) processing-related factors most strongly impact native speakers' and learners' choices alike; (ii) Spanish learners are more conservative regarding complementizer omission than German learners; (iii) both learner groups exhibit knowledge of target-like verb-complement type associations; and (iv) both learner groups indeed display sensitivity to register differences. These and other results will be discussed through the lens of usage-based construction grammar.

References

Torres Cacoullos, R. & J.A. Walker. 2009. On the persistence of grammar in discourse formulas: a variationist study of that. Linguistics 47.1:1-43.