Teaching and Language Corpora

Eleventh International Conference

Lancaster University

20th to 23rd July 2014

Abstract Book

Plenary presentations

A view to the future in corpus tools development

Laurence Anthony
Center for English Language
Education (CELESE)
Waseda University, Japan
anthony@waseda.jp

In a conference on teaching and language corpora, it is easy to forget that there is another central component of any corpus study, i.e., the corpus software tool (or tools) that we use to view and analyze our data. Without a corpus tool, a researcher, teacher, or learner would be lost in the depths of hundreds or thousands of texts and millions or billions of words of data.

Today, not only are corpora getting larger, but research on corpora is becoming ever more complex, and the users interacting with corpus data have broadened to include researchers, teachers, and learners. As a result, the design of corpus tools has become an increasingly important factor in the success of any corpus approach. However, the limited interaction that corpus linguists have with software developers makes effective tool development a challenging subject.

In this plenary talk, I will first explain why it is vitally important that we understand the nature of corpus tools and how they impact on our view of corpus data. I will then briefly review the history of corpus tools, looking at some of the most popular desktop and online tools to date, and discussing their strengths and weaknesses. Next, I will consider the future of corpus tools development, looking at the role of programming in corpus linguistics education and suggesting a practical approach to software tools development that mirrors the way tools are developed in other fields, such as physics. Finally, I will introduce some ongoing tool development projects that are freely available and exemplify the approach I describe.

Corpus linguistic investigations of construction usage and construction learning

Nick Ellis
University of Michigan
ncellis@umich.edu

Usage-based approaches believe that we learn language over the episodes of our communicating using language. Our linguistic ability emerges as a result of our cognitive learning mechanisms analysing this experience. Relevant research therefore requires the study of (1) the regularities of usage, (2) the regularities of acquisition, and (3) the regularities of construction knowledge. Corpus Linguistics provides relevant evidence.

1 Usage

The usage of Verb-Locative and Verb-Object-Locative English verb-argument constructions (VACs) is investigated in large corpora in terms of grammatical form, semantics, lexical constituency, and distribution patterns. VAC type-token frequency follows Zipfian scale-free patterns, as does the degree distribution of the corresponding semantic networks. This suggests that language form, language meaning, and language usage might come together across scales to promote robust induction by means of statistical learning over limited samples.

2 Usage in Learning: Child language acquisition

Analysis of the distribution of VACs in English child-directed speech (CDS) and child language in CHILDES corpora is also shown to be Zipfian, and measures of VAC-verb contingency showed VACs to be selective in their constituency. Language acquisition follows the leads of CDS usage.

3 Usage in Mind: L1 and L2 knowledge

VAC processing is sensitive to statistical patterns of usage. Native speakers of English generated V slot-fillers in 40 sparse VAC frames such as ‘he __ across the....’. Multiple regression analyses predicting the frequencies of types generated show independent contributions of (i) verb frequency in the VAC, (ii) VAC-verb contingency, and (iii) verb prototypicality in terms of centrality within the VAC semantic network. VAC processing involves rich associations, tuned by verb type and token frequencies and their contingencies of usage, which interface syntax, lexis, and semantics.

These results suggest that:

- Language usage is highly patterned in ways that support learning.
- Language acquisition is guided by this patterning.
- Language users have rich implicit statistical knowledge of these patterns.

Data driven learning in teacher training: tackling the challenge

Agnieszka Leńko-Szymańska
University of Warsaw
a.lenko@uw.edu.pl

Corpora have long been recognized as a valuable resource in language pedagogy. Numerous books, journal articles and conference presentations have advocated a variety of corpus applications: from more faithful descriptions of the target language and learners' needs, through the creation of more adequate materials for language teaching and learning, to the use of corpora by teachers and learners themselves. Indeed, it can be safely said that corpora are no longer solely a topic of the academic debate but they have found their way into real-life education. They are present in writing dictionaries (e.g. *Longman Dictionary of Contemporary English. New Edition* 2003), and reference grammars (e.g. Biber et al. 1999), as well as in designing courses and language materials (Mascull 1995; McCarthy et al. 2005). Yet, despite the encouragement from several researchers (Johns 1991) corpus data are still rarely used by teachers and learners in language classrooms (Römer 2010; Boulton 2010).

The unwillingness of language teachers to exploit corpora in their work can be caused by a number of real and perceived obstacles. However, the problem which is probably at the heart of teachers' reluctance to exploit corpora in language instruction is their lack of knowledge about how large linguistic databases can be used in the classroom (Mukherjee 2004; Römer 2009, 2010). Graduates of language departments and teacher training institutions might have heard about or even encountered corpora during their linguistic education. In some cases, they might have even used corpora regularly in their language or linguistics classes (O'Keefe and Farr 2003; Götz & Mukherjee 2006; Amador Morenot et al. 2006; Chambers 2005; Farr 2008; Heather & Helt 2012). However, this experience does not automatically imply that they know how to apply corpora in their teaching. Teachers may find it difficult to select items which are suitable for data-driven learning and relevant to their students, to develop effective corpus-based activities and to integrate them with other classroom techniques and procedures.

The aim of this presentation is to argue the importance of explicit teacher training in the potential of corpora for classroom use. This training should go beyond the skills in operating corpus tools and in interpreting the results of corpus explorations, but it should also focus of purely pedagogical issues related to the role and place of data-driven tasks in teaching and learning a foreign language. The presentation will review recent books directed to language teachers and promoting the use of corpora in language education (O'Keeffe et al. 2007; Reppen 2010). It will also survey few available accounts of institutionalized teacher training courses devoted to corpus applications in language pedagogy (Breyer 2009, Hüttner et al. 2009, Hather & Helt 2012). The talk will also present a teacher-training course on the use of corpora in language education offered to graduate students at the Institute of Applied Linguistics, University of Warsaw. The design, the syllabus, the progression and the outcomes of the course will be outlined; examples of student teachers' reflections on corpus-based activities will be summarized and corpus-based teaching activities developed by teacher trainees will be presented and discussed. Finally, the results of two questionnaires distributed to the participants after two editions of the course will be examined. The students' response reveal their attitudes to data-driven leaning and their reactions to the course itself. The conclusion will outline the implications for teacher training which could effectively promote data-driven learning among future teachers.

References

Amador Moreno, C.P., O'Riordan S. and Chambers, A. (2006) Integrating a corpus of classroom discourse in language teacher education: the case of discourse markers. *ReCALL, 18* (1): 83-104.

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

Boulton, A. (2010) Data-driven learning: Taking the computer out of the equation. *Language Learning, 60* (3): 534-572.

Breyer, Y. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning, 22* (2): 153-172.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning and Technology, 9* (2): 111-125.

Farr, F. (2008) Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness, 17* (1): 25-43.

Götz, S. and Mukherjee, J. (2006) Evaluation of data-driven learning in university teaching: a project report. In: Braun, S., Kohn, K. and Mukherjee, J. (eds.) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. English Corpus Linguistics, 3. Frankfurt: Peter Lang. 49-67.

Heather, J. and Helt, M. (2012) Evauating corpus literacy training for pre-service langauge teachers: Six case studies. *International Journal of Technology and Teacher Education, 20* (4): 415-440.

Hüttner, J., Smit, U., and Mehlmauer-Larcher, B. (2009) ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis. *System, 37*: 99-109.

Johns, T. (1991) Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal, 4*: 1-16.

Longman Dictionary of Contemporary English. New Edition (2003) Harlow: Pearson Education Limited.

Mascull, B. (1995) *Collins COBUILD Key Words in the Media*. London: HarperCollins Publishers.

McCarthy, M., McCarten, J. and Sandiford, H. (2005) *Touchstone Student's Book* 1. Cambridge: Cambridge University Press.

Mukherjee, J. (2004) Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In Connor, U. and Upton T. (eds.) *Applied corpus linguistics: A multidimensional perspective*. Amsterdam, NewYork: Rodopi. 239-250.

O'Keefe, A., & Farr, F. (2003) Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly, 37* (3): 389-418.

O'Keeffe, A., McCarthy M. and Carter, R. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Reppen, R. (2010) *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Römer, U. (2009) Corpus research and practice: What help do teachers need and what can we offer? In: Aijmer, K. (ed.) *Corpora and Language Teaching.* Amsterdam/Philadelphia: John Benjamins, 83-98.

Römer, U. (2010) Using general and specialized corpora in English language teaching: Past, present and future. In: Campoy-Cubillo, M.C., Belles-Fortuño, B. and Gea-Valor. L. (eds.) *Corpus-based Approaches to English Language Teaching*. London: Continuum, 18-35.

Taking stock, framing plans

Paul Thompson
University of Birmingham
p.thompson@bham.ac.uk

TaLC has come a long way, both in terms of distance and in terms of development . The first conference was held in Lancaster twenty years ago. Since then there has been a conference every two years, moving from its starting point here (1994, 1996), to Oxford (1998) thence to Graz (2000), Bertinoro (2002), Granada (2004), Paris (2006), Lisbon (2008), Brno (2010) and Warsaw (2012), before making its way home for this 2014 gathering. The TaLC conferences have been an inspiration and a focal point for many researchers and teachers who share concerns about introducing corpus resources, insights and tools into language teaching and teaching about language. My own introduction to TaLC came in 1996, when I was embarking on a PhD with corpus linguistics at its heart, whilst employed as a full-time EAP lecturer. TaLC soon became an addiction, an opportunity to discuss, enthuse and learn about new ideas in how corpora can be exploited in, and how they can inform, teaching about language, and I attended the next five conferences after that.

In this talk, I will look back over the twenty years of TaLC, reviewing the publications that came out of those events, and discussing the trends and concerns of papers in that period, with a particular eye on direct uses of corpus tools and resources in first and second language teaching. While there have been many encouraging signs of growth and dispersion of corpus uses in language education, this seems to remain still at a restricted level; in Rogers' (1962) terms, there are plenty of innovators (enthusiasts) emerging and even some early adopters, but large-scale diffusion of the innovation remains a long way off, with no sign yet of early majorities. There is a growing body of evidence that indicates that carefully designed and integrated corpus investigation activities can contribute richly to learning about language in a range of classroom settings, and, given that, it is time now to see how corpus tools, resources and insights can be introduced into language education on a much broader scale.

References

Rogers, Everett (1962). *Diffusion of Innovations*. Glencoe: Free Press.

TALC in action: 10 years on

Yukio Tono
Tokyo University of Foreign Studies
yukio.tono@gmail.com

It was in 2003 that I worked with NHK (Japan Broadcasting Centre) and supervised a series of corpus-based TV English conversation programmes. It was a huge success and the word "corpus" became a buzz word in Japan. In 2008, I was invited to give the first plenary at TALC8 in Lisbon and shared this news with TALCers (Tono 2011). Since then, the corpus fever has gone, but there has been a growing awareness that corpus-based research will shed light on various aspects in foreign language teaching and learning. In this talk, I will mention three major research projects I have been working on. One is the compilation of corpora of NHK foreign language learning programs. They collect all the past skits and model dialogues used in the English program on NHK, for which we tag the data not only for basic morphological analysis, but also for verb complementation patterns, pragmatic and functional roles of the sentences, and the situations in which each sentence occurs. This resource is called the NHK English Database based on the CEFR. We also developed a specialized interface for teacher education purposes, called LEAD. Teachers can search the database for particular functions of language, which can be closely associated with 'can do' descriptors used in the CEFR. This resource will provide not only the educational corpus tuned to Japanese learners of English, but also excellent materials for teacher training, with which teachers can build their own teaching materials.

The second area I have been working on is the development of the CEFR-J, an adaptation of the CEFR in Japanese contexts. I will report on the aims and the process of development of this CEFR-based framework in Japan and its impact. I will also discuss a process of Reference Level Descriptions (RLDs) for English. Using EFL course book corpora and learner corpora, we have been attempting to extract 'criterial features' for the given CEFR level, which attracts much attention now in L2 profiling research. Intensive use of machine learning is also a unique feature of my project.

Finally, I will discuss the on-going project of compiling the CEFR-based can-do performance corpus and error tagging with association rule mining. This new type of learner corpora and error tagging scheme will enrich the information about how learners can do with language across a range of tasks and help describe the Interlanguage processes in a more dynamic way.

References

Tono, Y. (2011). TaLC in action: recent innovations in corpus-based English language teaching in Japan. *New Trends in Corpora and Language Learning*, Ana Frankenberg-Garcia, Lynne Flowerdew, and Guy Aston (eds.), Continuum., 3-25.

Parallel session presentations

Teaching collocations and lexical phrases: A data driven learning approach

Hailah Alhujaylan
University of Essex
hsaalh@essex.ac.uk

L2 learners' struggle with formulaic language even at advanced levels is an attested issue in several ESL and EFL contexts. As multiword units are notoriously difficult for learners, they attracted researchers' attention to investigate whether multiword units can be deliberately learned. In this study, and after compiling a learner corpus of Saudi students written output, a number of verb-noun collocations and lexical phrases (e.g. at the expense of, take account of) appeared to be problematic. Therefore, they were selected to be taught via concordance-based and dictionary-based tasks. The concordance lines were retrieved from Collins WordBanks corpus. The concordance lines were in the keyword in context (KWIC) format, and they were truncated to fit onto the page, but not edited in anyway. Dictionaries were used for comparison with concordance lines because they provide an obvious point of comparison (Frankenberg-Garcia, 2005b; Yoon & Hirvela, 2004; Boulton, 2010). One of their main advantages is the list of examples they provide (Cobb, 2003). For the collocation worksheets, the entries were taken from *Longman Collocations Dictionary and Thesaurus* was used, and for the lexical phrases worksheets, *Oxford Idioms Dictionary for Learners of English* was the source of entries.

It is generally agreed that learning and retention of the various aspects of new vocabulary depends on the amount and the quality of learners' attention and processing to the new information. Consequently, Analyzing teaching techniques is without doubt a necessity to find out which vocabulary teaching activities can best help learning. According to Nation and Webb (2011) the best known and the best-researched way of analyzing vocabulary teaching activities is Laufer and Hulstijn's (2001) involvement load hypothesis. Thus, the design of the teaching materials is guided by the involvement load hypothesis. The involvement load hypothesis postulates that learning vocabulary is conditional upon three factors in tasks: need, search and evaluation. Since learners working under the concordance-based instructional condition will have to exert a more cognitive effort in decoding the new vocabulary information, it is suggested that their learning gains will be more durable. There were two experimental groups and the study implemented a counter-balanced design. Each group received both instructional conditions by learning half of the items using the corpus-based worksheets, and the other half with dictionary-based worksheets. The situation was reversed in the two experimental groups. In this way, no group or language item receives special treatment, and each can serve as control for the others. Learners' receptive and productive knowledge were measured by means of pre-tests, posttests and delayed tests. The results showed that learners in general learn better under the concordance-based treatment. Learning gains were not significantly better than dictionary-based instructional condition in the case of collocations, but they were significantly higher for the lexical phrases.

References

Boulton, A. (2010). “Data-Driven Learning: Taking the Computer Out of the Equation”. *Language learning*, 60(3), 534-572.

Cobb, T. (2003). “Do corpus-based electronic dictionaries replace concordancers?” In B. Morrison, G. Green, & G. Motteram (Eds.), *Directions in CALL: Experience, experiments, evaluation* (pp. 179–206). Hong Kong: Polytechnic University.

Frankenberg-Garcia, A. (2005b). “A peek into what today's language learners as researchers actually do”. *International Journal of Lexicography*, 18(3), 335–355.

Laufer, Batia, & Hulstijn, Jan H. (2001). "Incidental vocabulary acquisition in a second language: The construct of task-induced involvement". *Applied Linguistics*, 22(1), 1-26.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.

Parkinson, D., & Francis, B. (2006). *Oxford idioms dictionary for learners of English*. University Press.

Pearson Education, Limited. (2013). *Longman Collocations Dictionary and Thesaurus*. Pearson Longman.

Yoon, H., & Hirvela, A. (2004). “ESL student attitudes toward corpus use in L2”. *Journal of Second Language Writing*, 13(4), 257–283.

Advanced learners' errors in correcting Machine Translation output: comparative corpus-based analysis

Bogdan Babych
University of Leeds
b.babych
@leeds.ac.uk

Anne Buckley
University of Leeds
a.buckley
@leeds.ac.uk

Svitlana Babych
University of Leeds
s.babych
@leeds.ac.uk

1 Introduction

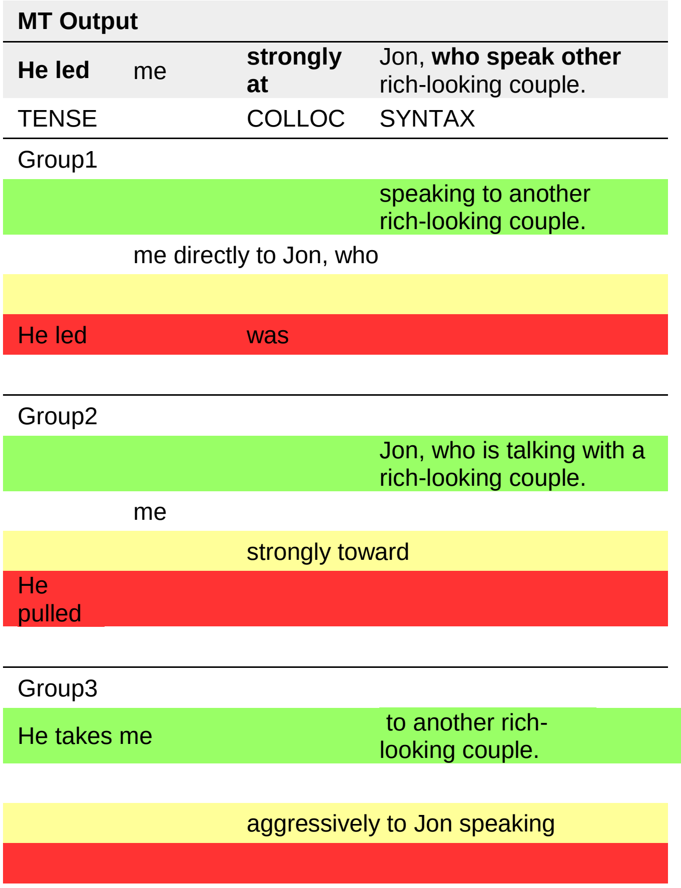
This paper describes an experiment on representing, annotating and analysing errors made by language learners who correct the output of Machine Translation (MT) systems. In our previous work (Babych et al., 2012) we presented the method of using error correction in advanced stages of language learning and translation training, where negative linguistic evidence is automatically generated by rule-based MT systems. MT output usually contains the original message, but with its fluency disrupted on the lexical, collocational or stylistic levels. By correcting MT errors the students are refining their skills in producing idiomatic and stylistically appropriate texts, with acceptable usage patterns, terminology, synonyms, collocations and lexico-grammatical constructions appropriate for the situation and linguistic context. This high proficiency level is particularly important for trainee translators (Kuebler, 2011; Aston, 1999). Our students critically review the MT output, discuss potential solutions in a group and/or with the tutor, and check their decisions by doing corpus-based research. In our method MT is used not simply as a useful dictionary alternative, but for systematically generating negative linguistic evidence (cf. Landure & Boulton 2010). Even though using ill-formed L2 can be counterproductive in the initial stages of foreign language learning (Somers, 2004), in the advanced stages negative linguistic evidence is useful, since students are aware of contrastive differences between languages and consciously take control over developing their productive skills in autonomous learning.

In this paper we describe the corpus format we use to represent MT output errors, which are categorized and aligned with corresponding students' (successful and unsuccessful) error corrections, and also classified and compared to the initial set of MT errors. Further we present an analysis of different error types in the corpus. The results inform the way in which we apply the proposed method in our 1-semester MA module English for Translators taught for Translation Studies students at the University of Leeds.

2 Error representation format and categorisation scheme

Students receive MT-generated texts as homework and do corrections in 3 or 4 groups on Wikis on the VLE. For the following class their corrections are annotated with the following colour coding: Green - excellent solution; White - acceptable solution; Yellow- could be improved (e.g. meaning correct but not very idiomatic); Red - wrong solution.

For the corpus we align initial MT errors with students' correction solutions submitted by each of the groups, as shown in Figure 1.



	MT Output			
	He led	me	**strongly at**	Jon, **who speak other** rich-looking couple.
	TENSE		COLLOC	SYNTAX
	Group1			
🟩				speaking to another rich-looking couple.
⬜		me directly to Jon, who		
🟨				
🟥	He led		was	
	Group2			
🟩				Jon, who is talking with a rich-looking couple.
⬜		me		
🟨			strongly toward	
🟥	He pulled			
	Group3			
🟩	He takes me			to another rich-looking couple.
⬜				
🟨			aggressively to Jon speaking	
🟥				

Figure 1: Alignment of MT errors and solutions

We annotate students' and MT errors using an error categorization scheme inspired by (James, 1998). The scheme is based on linguistic levels of errors (morphological, syntactic, lexical), but also takes into account the frequency of different error types. For example, collocation errors are a type of lexical error, but we annotate them separately because this category is very frequent. Table 1 shows examples of annotated error types.

Error type	Example
COLLOC	***Exaggerated** lipstick (=too much lipstick)*
LEXICAL	*Put hand on her shoulder with a **possessiveness** (=possessively)*
PREPOS	*Ten yards **at** my left (=on my left)*
TENSE	*Before I can stop him he **led** me to Jon (=leads)*
SYNTAX	*I **have not to him spoken** yet (=I have not spoken to him)*
ARTICLE	*You can make **better decision** (=a better decision)*
MORPH	*We are open for **negotiation** (=negotiations)*

Table 1: Examples of error types

3 Corpus-based error analysis

When working in groups, students can either miss, or successfully identify but incorrectly change, or finally – successfully correct MT errors. Table 2 shows percentages of such cases for each of the error types in MT output.

	not found	*changed (wrong)*	*corrected*	*Total*
COLLOC	18.9%	39.2%	41.9%	44.0%
LEXICAL	43.5%	30.4%	26.1%	13.7%
PREPOS	20.0%	33.3%	46.7%	8.9%
TENSE	50.0%	8.3%	41.7%	7.1%
SYNTAX	0.0%	52.0%	47.7%	26.2%
Total	19.60%	38.7%	41.7%	100.0%

Table 2: Students' initial correction of MT errors

It can be seen from the table that while students always identify syntax errors, in around 20% of cases they do not see that there is a problem with a collocation or a preposition; when MT errors are correctly identified, about 50% of students' initial changes are correct.

Finally, we annotated error relation patterns in MT and student texts. We recorded which types of MT errors resulted in which types of the student errors, and which errors were corrected or emerged from the correct structures ('0' symbol in Table 3 shows absence of errors):

Pattern	Percentage
COLLOC>COLLOC	17.3%
COLLOC>0	10.9%
SYNTAX>0	10.9%
LEXICAL>LEXICAL	10.0%
0>COLLOC	5.5%
LEXICAL>0	5.5%
0>TENSE	4.5%
TENSE>TENSE	4.5%
PREPOS>PREPOS	3.6%
SYNTAX>COLLOC	3.6%
0>PREPOS	2.7%
PREPOS>0	2.7%

Table 3: Frequent error relation patterns (MT>Students)

It can be seen from Table 4 that the most frequent patterns are COLLOC>COLLOC – non-identified or wrongly corrected collocation error, followed by corrected collocation and syntax errors; in 5.5% of cases a new collocation error was introduced. This highlights the fact that collocations remain one of the most serious challenges for advanced language learning. Other patterns for newly introduced student errors are 0>TENSE and 0>PREPOS, which shows that these types of problems also have high priority for advanced learners.

References

Aston, G. 1999.Corpus use and learning to translate. *Textus* 12: 289-313.

Babych, B., Buckley, A., Hughes, R & Babych, S. 2012. Machine Translation technology in advanced language teaching and translator training: a corpus-based approach to post-editing MT output. In: Proceedings of of TALC 2012 : Teaching and Language Corpora Conference. Warsaw, Poland on 12th - 14th July 2012.

James, K. 1998. *Errors in language learning and use. Exploring error analysis*. London and New York: Longman.

Kübler, N. 2011. Working with corpora for translation teaching in a French-speaking setting. In *New trends in corpora and language learning*,A. Frankenberg-Garcia, G. Aston & L. Flowerdew (eds.), 62-79. London: Continuum.

Landure, C., & Boulton, A. 2010. Corpus et autocorrection pour l'apprentissage des langues. *Asp* 57: 11-30.

Somers, H. 2004. Does machine translation have a role in language learning? In *Proceedings of UNTELE 2004: L'Autonomie de l'Enseignant et de l'Apprenant face aux Technologies de l'Information et de la Communication – Teacher and Learner Autonomy vis-a-vis Information Communication Technology*, Compiègne, France, 28.

Textual patterns and text types: using connectors for automated genre classification

Svitlana Babych
University of Leeds
s.babych@leeds.ac.uk

The knowledge of a text type, or genre, is a useful concept for many FLT tasks. The main reason is that genres are dependent on communicative situations or contexts (such as writing an official letter vs. a letter to a friend, talking to a colleague at work or at a formal job interview) and form a part of the formal schemata. These contexts are characterised by a set of highly conventional textual patterns at the lexical, syntactic and rhetorical levels. While native speakers can much more easily link the text types to appropriate language patterns and vice-versa, this task is very difficult for FL learners: even high proficiency in grammar and the lexicon of a foreign language does not help if a learner is not aware of language patterns appropriate for specific text types and genres used in the given contexts and communicative situations. Therefore, there is a need to develop a systematic methodology to support language learners in developing conscious awareness and skills of recognising and using textual patterns, based on insights that discourse analysis has provided into text types and the relationships between texts and their contexts (McCarthy, 1991).

Textual patterns are systems of linked linguistic resources in the micro- and macro-structure which endow texts with new functions via their relations to each other. Certain patterns tend to occur frequently in particular settings: for instance, temporal connectors in relation to verbs with a certain tense are characteristic for a narrative genre. Other common types of textual patterns are 'problem/solution' relations, which are frequent in advertising texts and in texts reporting technological advances, and the 'general/specific' relation found in encyclopaedias and other reference texts.

I focus on detecting those textual features which indicate the text structure and text genre and which also can help FL students to better understand the general organisation of a text depending on its type. I consider *conjunction* as a type of text cohesion (Halliday and Hassan, 1976) to be the most appropriate starting point for discourse analysis for this purpose. Functioning to mark semantic relations between parts of the text, conjunctive elements - in following referred to as *connectors* - signal the logical text structure. A constellation of connector types in each text forms its *conjunction profile.*

This paper investigates whether conjunction profiles (as a type of language patterns) can be efficiently used to detect different genres automatically.

In this experiment I tested on a large scale my *connector-text-type hypothesis* that textual connectors are useful features to predict genre characteristics of texts.

A corpus of news texts was collected from the websites of major Ukrainian, Russian and English newspapers, around 250MW each. A large proportion of texts were automatically labeled, based on the tokens extracted from their URL addresses and file names. The labels, which roughly classified the texts into different genres, were *story, review, comment, blog* and *interview*. The total number of labelled texts is 32215 for English, 73356 for Russian and 89044 for Ukrainian.

For each of the texts in my corpus I also automatically extracted its conjunction profile, using my multilingual connector classification scheme described in (Babych, 2012). Then I ran a machine-learning experiment that related features of conjunction profiles (such as the constellations of conjunction types) with the categories for genres.

Counts or binary indicators of the presence/absence of connector classes in those documents were used as input features, and the task was to predict the correct genre label for each text. These connector-based features and labels were submitted to Weka (Hall et al., 2009) – an open-source machine learning toolkit. A supervised machine-learning algorithm (SVM) ran on this data, learnt and re-applied an automatic classifier that can predict genre labels.

The numbers of labeled texts were balanced: if a particular class was too large, the number of instances presented to Weka was limited to 2000, which allowed the system to use the same order of instances for the majority of the classes.

To evaluate the accuracy of genre classification, part of the corpus was used for training on the labeled data, and then the classifier was testing using the standard 10-fold cross-validation procedure.

The results show that conjunction profiles make good predictions of genres. The best accuracy on automatically labeled data was over 75% for English and Russian and over 90% for Ukrainian. This is well above the random assignment baseline of 17% (for 6 genre labels). The accuracy depends on the consistency of genre labeling across different sources in each of the languages, but the results clearly indicate that the connector-text-type hypothesis is valid and connector profiles of text are linked to genres of the texts and can support reading strategies in learning these text types. Text genres are associated with configurations of textual connectors. The link between the discourse structure signaled by the connectors and genres explains the need to consciously develop understanding of contrastive features of genres and the text structure.

References

Babych, S. 2012. “Exploring patterns of textual cohesion in multilingual corpora and their application for teaching language and translation”. *In: Teaching and Language Corpora Conference TALC 2012.* Warsaw, Poland.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. 2009. “The WEKA Data Mining Software: An Update”. *SIGKDD Explorations*, 11(1).

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

McCarthy, M. 1991. *Discourse Analysis for Language Teachers.* Cambridge: Cambridge University Press.

Adverb use in L2 writing

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

1 Introduction

The focus for the present study is L2 writing and for this we analyse samples of argumentative writing in English by French, Polish and Swedish speakers taken from the International Corpus of Learner English (Granger et al. 2002), along with some equivalent essays written by native speakers taken from the Locness Corpus. Using this data, we can carry out a comparative analysis and can contrast the patterns of usage of the different groups of learners and make comparisons with the English native speakers (NS).

Previous studies of the underuse of overuse of adverbs or other expressions is typically been based on a contingency table comparing the frequency of occurrence of the two elements, taking into account the number of words in each of the corpus. The software producing the calculation acts like a centrifuge, grinding up the text structure to release the words, which can then counted. This can be called the "bag of words approach" because the learner corpus and reference corpus are treated as collection of words rather than texts that consist formally of sentences and paragraphs and functionally contain ordered discourse or rhetorical structures. Bolton et al. (2002:172) describe the bag of words method as "fundamentally flawed", at least in the context of examining connectors.

Whether or not such an approach is fundamentally flawed, it can be said to be a rather blunt instrument. This study aims to extend previous studies on, for example, the overuse/underuse of adverbs, by including position information.

2 Method

The essays were annotated using the CLAWS7 POS tagset and the writing was then analysed using WordSkew, which allows the user to determine the frequency of words, phrases or POS tags across portions of different textual units: sentences, paragraphs, or the complete text. The “portions” can either be calculated in relation to equal division of the unit --- first 10%, second 10%, etc --- or as absolute positions such as first word in the sentence, first sentence in the paragraph etc. Thus it is possible to search for positions “1 2 3 Other #1 #” where # stands for last position and #1 is the penultimate position. The software gives the results as histograms (and tables). Clicking on a particular bar of the histogram reveals the concordance lines for that position. In this study a search for the position of a linguistic expression within a sentence will process around 10,000 sentences for each L1.

3 Sample results

The graph in Figure 1 shows the distribution of simple adverbs in sentences. The general configuration is remarkably similar for the different groups and follows a max-low-high-falling trajectory. The adverb use then decreases slowly towards the end of the sentence. Compared with a general count of adverb use by total words or total sentences, the positional analysis provides a clearer picture of overuse and underuse and here we can see that the distinctions among the different groups is most pronounced in positions, 1 and 3, with position 1 being the most markedly divergent. The Polish writers use general adverbs the most frequently in sentence-initial position but appear to drop below French amd Swedish writers in the frequency of adverb use in position 3.

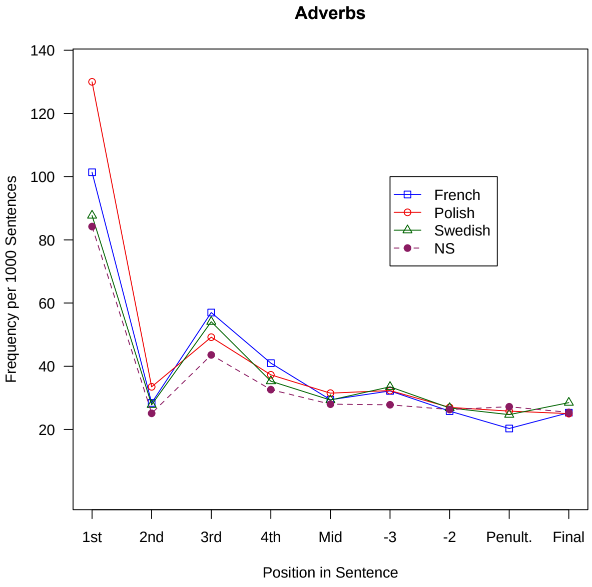


Figure 1: Adverb frequency across sentences

Figure 1 also reveals that the distribution of adverbs in the NS writing is less skewed. The highs are not so high and the lows not so low.

Taking another example, we can track the use of time adverbs, as shown in Figure 2. In this case, we find, for example, that all the L2 writers overuse time adverbs in position 1. If we examine the particular adverbs used in sentence-initial position, we find the top-ranked words are *nowadays* (French and Polish), today (Swedish), and *now* (NS).

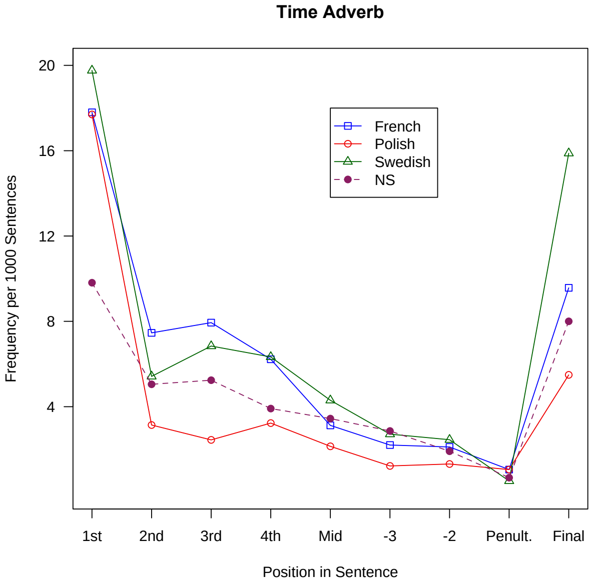


Figure 2: Time adverb frequency across sentences

References

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

Granger, S., Dagneaux, E. and Meunier, F. (eds.) 2002. *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

“Nevertheless” and “however” in learners’ writing: Aspects of use and variation

Lina Bikelienė
Vilnius University
l.bikeliene
@gmail.com

Ekaterina Zaytseva
University of Bremen
zaytseva
@uni-bremen.de

Ever since their advent learner corpora have served as an invaluable tool for providing an empirically-based description of learners’ language use. Corpus-based evidence on the advanced/ upper-intermediate interlanguage (IL) is of equal potential value for Second Language Acquisition Researchers and EFL or EAP practitioners.

One of the features repeatedly pointed out in Learner Corpus Research (LCR) as being characteristic of highly proficient EFL learners has been “non-nativeness” (Granger 1998:13), i.e. learners’ non-idiomatic choices in terms of general frequency of individual linguistic items, as well as lexico-grammatical and syntactic patterns thereof. LCR studies of advanced interlanguage have contributed to a better awareness of those areas of language use which remain problematic even for advanced learners. One of such difficult issues has been the use of connectors used to mark a certain relation between discourse units (e.g. resultive connectors *therefore*, *hence*, contrastive/ concessive *however*, *nevertheless*, listing connectors *moreover*, *in addition*, etc.) (cf. Altenberg and Tapper 1998; Bikelienė 2008; Granger and Tyson 1996; Narita and Sugiura 2006; Paquot 2008, 2010). The LCR studies have reported quantitative as well as qualitative differences between learners and native speakers as to preferences for individual connectors and usage patterns thereof in writing (e.g. Narita et al. 2004; Tang and Ng 1995).

The approach that has been taken in the LCR in general and in the LCR studies on linking devices in particular, has been the comparative/ contrastive approach, i.e. learner language production has been compared with that of English native speakers, where the latter has been used as a kind of yardstick against which features of learner writing have been characterized as non-native-like. Valuable as it has been for identifying features of non-idiomaticity in L2 language use, the comparative/contrastive methodology, however, does not offer a comprehensive list of possible explanations of learners’ choices. Thus, for example, only limited evidence has been provided on possible reasons behind learners’ non-idiomatic use of connectors in writing, such as register unawareness, transfer of L1 usage patterns, etc. (e.g. Altenberg and Tapper 1998; Crew 1990; Paquot 2008, 2010; Fei 2006; Babanoğlu 2012). The majority of LCR studies of written interlanguage to date deal with learners’ production of essays and not with writing of other genres, like research papers, summaries, etc. (cf. however, Paquot et al. 2011; Römer 2009; Wulff and Römer 2009). It remains unclear to what extent the LCR findings on the L2 use of connectors revealed so far, can be generalized to various genres/ text types, L1 backgrounds, and task settings.

Meanwhile, a variationist perspective on advanced interlanguage considering a possible influence of variables (e.g. genre/ text type, native language, task setting, etc.), has a potential to provide missing information on (hidden) systematicity in learners’ linguistic behaviour. Yet, studies combining comparative/ contrastive and variationist approaches to learner language are still scarce (see, however, Ädel 2008; Granger 1996; Paquot 2008, 2010; Paquot et al. 2011; Wulff and Römer 2009).

This study combines contrastive/ comparative and variationist frameworks to investigate the use of concessive/ contrastive connectors “nevertheless” and “however” in writing of L1 and L2 novice academic writers and addresses the following research questions:

1. To what extent is learners’ use of *however* and *nevertheless* different or similar to that of native speakers across several genres/ text types?
2. Is there variation in L2 language use as to the use of these linking devices in writing?
3. Is variation in learners’ use of *however* and *nevertheless* determined by genre/ text type and learners’ native language as two plausible variables?

The analysis draws on a combination of several L1 and L2 corpora. The L2 writing will be represented in the International Corpus of Learner English (ICLE) (Granger et al. 2009) and the *Corpus of Academic Learner English* (CALE). Several comparable native English corpora will be used in order to provide contrastive evidence on the use of the linking devices at hand: the Louvain Corpus of Native English Essays (LOCNESS) (Granger, 1996), the Michigan Corpus of Upper-level Student Papers (MICUSP) (Römer and Brook O'Donnell 2011), and the British Academic Written English corpus (BAWE) (Alsop and Nesi, 2009). Preliminary findings indicate differences in the use of *however* and *nevertheless* by learners and native speakers and point to variation in L2 written language use.

References

Ädel, A. 2008. "Involvement features in writing: do time and interaction trump register awareness?" In G. Gilquin, S. Papp and M. B. Diez-Bedmar (eds.) *Linking up contrastive and learner corpus research*. Amsterdam, Atlanta: Rodopi.

Alsop, S. and Nesi, H. 2009. "Issues in the development of the British Academic Written English (BAWE) corpus". *Corpora* 4 (1): 71-83.

Altenberg, B. and Tapper, M. 1998. "The use of adverbial connectors in advanced Swedish learners' written English". In S. Granger (ed.), *Learner English on computer*. Harlow: Addison Wesley Longman Limited.

Babanoğlu, P.M. 2012. *A corpus-based study on Turkish EFL learners' written English: The use of adverbial connectors by Turkish learners*. Unpublished PhD thesis. Çukurova university. Available online at: http://library.cu.edu.tr/tezler/8714.pdf.

Bikelienė, L. 2008. "Resultive connectors in advanced Lithuanian learners' English writing". *Kalbotyra* 59 (3): 30-37.

Crew, W. 1990. "The illogic of logical connectives". *ELT Journal* 44 (4): 316-325.

Fei, D. 2006. "The effect on the use of adverbial connectors in Chinese EFL learners English writing quality". *CELEA Journal* 29 (1): 105-111.

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast. Text-based cross-linguistic studies*. [Lund Studies in English 88]. Lund: Lund University Press.

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

Granger, S. 1998. "The computerized learner corpus: a versatile new source of data for SLA research". In S. Granger (ed.) *Learner English on computer*. Addison Wesley Longman: London and New York.

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

Narita, M., Chieko, S. and Sugiura, M. 2004. "The use of linking adverbials in the English essay writing of Japanese EFL learners". In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004).*

Narita, M. and Sugiura, M. 2006. "The use of adverbial connectors in argumentative essays by Japanese EFL college students". *English Corpus Studies* 13: 23-42.

Paquot, M. 2008. "Exemplification in learner writing: a cross-linguistic perspective". In S. Granger and F. Meunier (eds.) *Phraseology in foreign language learning and teaching*. Amsterdam: Benjamins.

Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. United States: Continuum Publishing Corporation.

Paquot, M., Hasselgård, H. and Ebeling, S.O.. 2011. "Writer/reader visibility in learner writing across genres: A comparison of the French ad Norwegian components of the ICLE and VESPA learner corpora". *Paper Presented at the International Conference 'Learner Corpus Research 2011', September 2011, Louvain-la-Neuve, Belgium.*

Römer, U. 2009. "English in academia: Does nativeness matter?" *Anglistik: International Journal of English Studies* 20 (2): 89-100.

Römer, U. and O'Donnell, M.B. 2011. "From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP)". *Corpora* 6 (2): 159-177.

Tang, E. and Ng, C. 1995. "A Study on the Use of Connectives in ESL Student's Writing". *Perspectives Working Papers* 5(1): 38-45.

Wulff, S. and Römer, U. 2009. "Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive". *Corpora* 4(2): 115-133.

Assembling the data on data-driven learning: A meta-analysis of design issues and outcomes

Alex Boulton
CNRS & University of Lorraine, France
alex.boulton@univ-lorraine.fr

Tom Cobb
Université du Québec à Montréal, Canada
cobb.tom@uqam.ca

Corpus linguistics is, by its very nature, an applied field, with potential relevance in any area which deals with text in one form or another. One of the earliest such applications was in language teaching, where it is commonly associated with the work of Tim Johns (1990). He coined the phrase *data-driven learning* (DDL), which is still commonly used as a cover term for any use of language corpora or associated corpus linguistic tools or techniques for L2 users. The last 20-odd years have seen considerable output of academic papers in the area, outlining possible activities, describing actual courses, explaining and justifying the rationale, and so on. The obvious question, of course, is: “Does it work?”

There are several ways of addressing this, the most obvious being to conduct an original study to collect new data on specific questions. As individual studies accumulate, however, some kind of overview becomes necessary to gain a broader picture of the field as a whole. Considerable work in applied linguistics over recent years has sought to find ways to make such research synthesis as systematic as possible, and can generally be divided into two broad streams: narrative synthesis and meta-analysis, each with its own advantages and disadvantages (see the papers in Norris and Ortega 2006 for an overview).

The narrative synthesis is not unlike the traditional ‘literature review’ which features in the introductions to many academic papers. It differs in avoiding the narrow focus of a specific topic, and attempts a systematic trawl of all relevant publications, thus reducing the subjective selection of papers for consideration. But it potentially falls down on the rigour of the analysis itself, which can retain some of the problems inherent in literature reviews (e.g. Boulton 2010 on learning outcomes of DDL; Boulton 2012 on DDL in English for Specific Purposes).

The meta-analysis involves essentially the same systematic collection of papers, but focuses exclusively on quantitative studies, thus neglecting the value of qualitative studies (for an overview, see Richards 2009). It is thus less broad in coverage than the narrative synthesis, and the analysis is potentially reductive and simplistic, lumping together of all types of specificities of individual studies. Its advantage is that it allows a pooling of the quantitative data from all relevant studies available.

This paper presents a meta-analysis of DDL studies (cf. the preliminary work in Cobb and Boulton forthcoming). The work is still in progress, but to date we have collected 140 papers which seek to evaluate some aspect of L2 corpus use, of which 21 provide suitable quantitative data – minimally, means and standard deviations deriving from pre/post-tests and/or experimental/control groups. Work so far suggests a substantial effect size, currently standing at 1.42 standard deviations. Focus on a single effect size figure can be strategically or politically expedient (cf. Grgurović *et al*. 2013), but meta-analysts are keen to go beyond this to avoid a reductionist picture in such a complex area as language learning (cf. Larsen-Freeman and Cameron 2008). This paper thus seeks to situate the study, with the focus correspondingly not only on the outcome itself but also on the issues raised in collecting and selecting the studies for inclusion, as well as in analysing and sorting the resulting data.

Due consideration is given in particular to the deliberately broad **definition** of DDL, which we have taken to include all uses of corpora by non-native speakers. This seems to be compatible with Johns’ original vision, and since DDL clearly means a range of different things to different people, it seems sensible to begin with a broad sweep. We also discuss the **inclusion / exclusion** criteria in the selection process: we make no distinction between papers appearing in prestigious peer-reviewed journals and elsewhere – smaller journals, book chapters, conference proceedings, as well as ‘grey’ literature in the form of unpublished doctoral theses (though we exclude research which has not been formally written up, such as unpublished conference presentations or slides). This should ensure that quality work published outside mainstream sources is not ignored, and that negative outcomes in particular are less likely to be overlooked (Oswald and Plonsky 2010). Some meta-analyses have introduced weighting systems, though we have initially attempted to avoid such *a priori* judgements.

Other issues arise from the pooling of quantitative data from highly varied studies – in other words, are the studies sufficiently similar that their results can be legitimately pooled at all? Though we argue that we are not comparing apples and oranges in the overall meta-analysis, the studies can usefully be grouped into different sub-categories for more in-depth understanding, and in one of two ways. In terms of **research design**, particular importance is accorded to the distinction between the *effectiveness* of a treatment (as measured by within-groups pre/post-tests: ES = 1.68; *d* = .84) and its relative *efficiency* (between groups: ES = 1.04; *d* = .73). In terms of **research questions**, it is also possible to derive a number of sub-categorisations allowing meta-analyses of subsets depending on more focused topics, thus allowing greater depth of understanding on more specific issues.

As in corpus linguistics, raw data and statistics are useful, but they need interpretation and contextualisation to become meaningful. A careful meta-analysis, with transparent inclusion criteria and sensitivity to individual differences between studies, provides one way of combining both, underlining the importance of effect size in relation to statistical significance (cf. Duff *et al*. 2007). As such, it allows us to go some way towards overcoming the fragmentation of the field and to provide some kind of evaluation of the state of research in DDL as a whole. In devising more focused subsets of studies, we are able to adopt a realistic evaluation (Pawson and Tilley 1997) and address not just the question of ‘Does it work?’, but how effective and efficient it might be in different forms for different learners for different purposes in different circumstances.

References

Boulton, A. 2010. “Learning outcomes from corpus consultation.” In M. Moreno Jaén, F. Serrano Valverde and M. Calzada Pérez (eds.) *Exploring new paths in language pedagogy: lexis and corpus-based language teaching*. London: Equinox, p. 129-144.

Boulton, A. 2012. “Corpus consultation for ESP: a review of empirical research.” In A. Boulton, S. Carter-Thomas and E. Rowley-Jolivet (eds.) *Corpus-informed research and learning in ESP: issues and applications*. Amsterdam: John Benjamins, p. 261-291.

Cobb, T. and Boulton, A. 2014. “Classroom applications of corpus analysis.” In D. Biber and R. Reppen (eds.) *The Cambridge handbook of corpus linguistics*. Cambridge: Cambridge University Press.

Duff, P.A., Norris, J.M. and Ortega, L. 2007. “The future of research synthesis in applied linguistics: beyond art or science.” *TESOL Quarterly* 41: 805-815.

Grgurović, M., Chapelle, C.A. and Shelley, M.C. 2013. “A meta-analysis of effectiveness studies on computer technology supported language learning.” *ReCALL* 25 (2): 165-198.

Johns, T. 1990. “From printout to handout: grammar and vocabulary teaching in the context of data-driven learning.” *CALL Austria* 10: 14-34.

Larsen-Freeman, D. and Cameron, L. 2008. *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Norris, J.M. and Ortega, L. (eds.) 2006. *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.

Oswald, F.L. and Plonsky, L. 2010. “Meta-analysis in second language research: choices and challenges.” *Annual Review of Applied Linguistics* 30: 85-110.

Pawson, R. and Tilley, N. 1997. *Realistic evaluation*. London: Sage.

Richards, K. 2009. “Trends in qualitative research in language teaching since 2000.” *Language Teaching* 42 (2): 147-180.

Collocations in the primary school EFL classroom

Martina Bredenbröcker
University of Paderborn
martina.bredenbroecker@upb.de

1 Introduction

Various research papers (cf. Meunier/Granger 2008, Aijmer 2009, Römer 2009, Reppen 2010, etc.) stress the importance of transferring research results from corpus linguistics to the wide field of language pedagogy. These studies mostly focus on learners older than eleven years of age, in secondary school EFL classrooms. However, to provide a solid foundation for **life-long language learning** as proposed in the Common European Framework of Reference for Languages (CEFR), it is necessary to start from an earlier age (e.g. from age 6 – the age when children begin school – onwards). Consequently, this work-in-progress report concentrates on young learners at primary school level.

2 Data

The Oxford Children’s Corpus (OCC) is the basis for an investigation of lexical structures relevant for young learners of English (Wild et al. 2012; Banerji, N. et al. 2013). It has been developed by Oxford University Press and is used by their lexicography team to inform the writing of dictionaries for children. It currently comprises 126 million tokens, including material written for 5- to 14-year-old children (e.g. fiction, websites, magazines), and over 86 million tokens from writing by children themselves (mainly short stories collected in a BBC competition held in 2012 and 2013). Using SketchEngine (Kilgarriff, A. et al. 2004), an online corpus query tool, a subcorpus (OCC-SUB) was compiled, which is currently at ca. 26 million tokens. OCC-SUB contains writing exclusively from primary school children and is therefore pertinent to the purpose of this study.

3 Methodology

Lemmatized frequency lists that are derived from the OCC-SUB will be contrasted with BNC frequency lists. The latter was selected because it contains mostly adult language, and it is comparable in size, mode and language variety. In a second step, five nouns and five lexical verbs from the top frequency ranks of both corpora were chosen for a study of their collocational behaviour. These verbs include *SEE, SAY, LOOK, GET, COME* and the nouns *DAY, MAN, TIME, FRIEND,* and *SCHOOL.* Using different statistical measures (MI-score, T-score) a basic set of collocations will be compiled. This is followed by a discussion whether the resulting set is applicable for every primary EFL classroom, regardless of the learner´s first language.

4 Example *SEE*

To give an example: *SEE* is ranked fourth in the frequency list of lexical verbs in the OCC-SUB, with 100,858 instances (3808.5 per million). In the entire OCC it is in fifth place with 425,832 hits (3354.2 per million).

Among the top five object nouns for the verb *SEE* in the BNC are *p.* (short for *page*), *chapter, page, figure (= diagram),* and *man.* There is only one overlap with the OCC-SUB, namely *man.* The other object nouns found in the children´s corpus *(light, face, thing, figure = shape of person,* and *mum)* describe objects and concepts taken from the childhood world of experience, i.e. basic aspects of life that are relevant for children at primary school level.

In particular, a look at the most frequent subjects highlights the extent to which adult language usage is different: adults use *SEE* commonly with inanimate nouns such as *year, century* and *world* whereas the OCC subjects are all animate and consist mostly of personal pronouns. A possible explanation for this is that the basic meanings of *SEE*, namely *perception by sight, watch* and *understand* presuppose not only animate but particularly human subjects. Obviously children acquire these meanings first.

A common collocation in both corpora is *to see things* as in: ‘Then I saw a green human thing coming towards me.’ [BBC-E-111392]

The MI-score for this collocation is at 29.29 and the T -score at 47.48[1], so it can be considered both a strong and a certain collocation. It should therefore be included in the basic vocabulary to be taught in EFL textbooks at primary school level.

5 Language-pedagogical consequences

The set of collocations will be used as a starting point for an investigation of language-pedagogical consequences. By applying tried and trusted vocabulary selection criteria, the chunk collection will be evaluated with regard to the needs of young learners. Among these criteria are – apart from frequency – learnability, availability, familiarity, coverage and regularity (cf. Nation 2001).

[1] MI-score cutoff: >3, T-score cutoff >2

Implications for different vocabulary presentation and new forms of motivating exercises which are suitable for children at that early age will also be discussed.

6 Conclusions

The study shows that there is a need for corpus-based language-learning materials specifically designed for very young learners. To support improved retention and more native-like use of collocations not only in primary but also in secondary schools and beyond, (basic) vocabulary should be taught in its collocational context wherever possible and appropriate.

References

Aijmer, K. (2009). *Corpora and Language Teaching*. Amsterdam: Benjamins.

Banerji, N., Gupta, V., Kilgarriff, A., Tugwell, D. (2013). ‘Oxford Children's Corpus: A Corpus of Children’s Writing, Reading, and Education’. In: Corpus Linguistics 2013 Abstract book <http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf> (accessed 9.9.13).

European Commission. (2003) ‘Action Plan on Language Learning and Linguistic Diversity’. <http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11068_en.htm> (accessed 9.9.13).

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). ‘The Sketch Engine‘. In: Williams, G., Vessier S. (Eds.) *Proceedings of the Eleventh Euralex Congress*, Lorient: UBS.

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

Nation, P. (2001). *Learning Vocabulary in another Language*. Cambridge: Cambridge University Press.

Reppen, R. (2010). *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Römer, U. (2009). ‘Corpus Research and Practice: What Help do Teachers Need and what can we Offer?’ In: Aijmer, K. *Corpora and Language Teaching*. Amsterdam: Benjamins.

Wild, K., Kilgarriff, A., Tugwell, D. (2012). ‘The Oxford Children's Corpus: Using a Children’s Corpus in Lexicography’. In: *International Journal of Lexicography*, ecs017v1-ecs017. Oxford: Oxford University Press.

Certainty and uncertainty in learner speech: An exploration of the use of epistemic markers in the *Trinity-Lancaster Spoken Corpus*

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

1 Introduction: Epistemic positioning and learner language

Certainty and uncertainty (epistemic stance) in language can be either explicitly marked (Biber 2006; Simon-Vandenbergen & Aijmer 2007) or merely implied by the pragmatics of the speech act (Holmes 1984). By indicating our epistemic stance, we simultaneously evaluate a proposition, position ourselves and align with the hearer (Du Bois 2007).

Epistemic stance is a complex pragmatic phenomenon which 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).

2 Method

The study is based on the advanced speech subcorpus of the *Trinity-Lancaster Spoken Learner Corpus*, which consists of transcribed dialogues between learners and examiners. The subcorpus comprises 0.5 million running words, 60 per cent of which come from the learners. The corpus is a unique tool for investigating the dynamics of learner spoken interaction in a semi-formal context. Table 1 below provides more details about the data used.

Tokens	**Students**	**Student speech (turns/ tokens)**	**Countries of origin**
521,199	133	19,785/ 313,752	China, India, Italy, Mexico, Sri Lanka, Spain

Table1: *Trinity-Lancaster Spoken Corpus* – advanced speech

This study focuses on adverbial markers of epistemic stance (AEMs) and their distribution in the corpus. The following is a list of 25 AEMs compiled from the literature (Holmes 1988; Biber 2006) which were searched for in the corpus:

actually, always, apparently, certainly, definitely, evidently, in fact, in most cases, in most instances, indeed, inevitably, kind of, maybe, never, no doubt, obviously, of course, perhaps, possibly, predictably, probably, roughly, sort of, undoubtedly, without (any) doubt

Table 2: Adverbial epistemic markers (AEMs)

These AEMs were further divided into seven different categories based on the meaning similarities (cf. Simon-Vandenbergen & Aijmer 2007). These categories can be seen in Table 3.

Semantic group	**AEMs**
EPISTEMIC CERTAINTY	undoubtedly, without (any) doubt, no doubt, certainly, definitely
ACTUALITY-REALITY	actually, indeed, in fact
OBVIOUSNESS	of course, obviously
TEMPORAL CERTAINTY	always, never
PROBABILITY	apparently, evidently, inevitably, in most cases, in most instances, predictably, probably
POSSIBILITY	maybe, perhaps, possibly
IMPRECISION	kind of, sort of, roughly

Table 3: Semantic categories of AEMs

Finally, the results were compared with the findings based on British informal speech from a subset of the *BNC* reported in Brezina (2012).

3 Results and discussion

Overall, out of the 25 AEMs that were investigated 21 occurred in the corpus but only 15 appeared with a frequency greater than 10. Table 4 shows top ten AEMs in the whole dataset.

AEM	**Frequency**
maybe	1096
kind of	540
always	496
actually	464
of course	324
perhaps	256
never	230
probably	211
sort of	153
in fact	116

Table 4: Top ten AEMs in the *Trinity-Lancaster Spoken Learner Corpus*

The results also indicate interesting differences in the use of AEMs between students and examiners. While the students preferred adverbs such as *maybe, kind of, of course* and *actually*, the examiners overused *sort of, certainly, perhaps, obviously* and *possibly*.

When the semantic categories were taken into consideration (see Figure1) an even more interesting picture emerged. Figure 1 reports the occurrence of AEMs grouped under seven semantic categories in students' and examiners' speech; in addition, it also offers figures based on the *BNC* conversation (cf. Brezina 2012) for comparison.

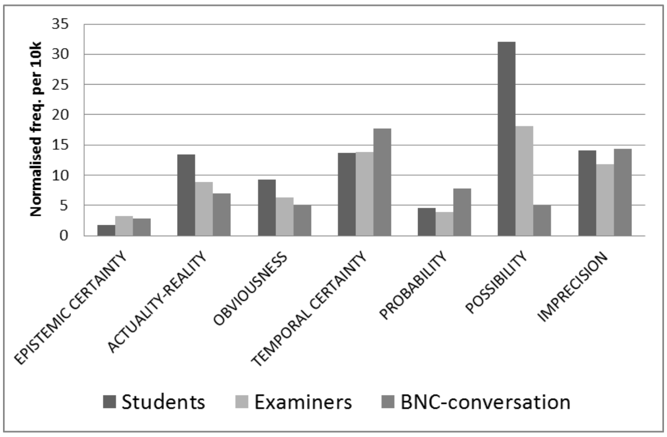


Figure 1: Semantic categories of AEMs in the *Trinity-Lancaster Spoken Learner Corpus* and *BNC* – conversation

As can be seen, the most common semantic group in the *Trinity-Lancaster Spoken Learner Corpus* is POSSIBILITY followed by TEMPORAL CERTAINTY and IMPRECISION. On the other hand, the least frequent of these is EPISTEMIC CERTAINTY. Figure 1 also shows large differences between student and examiner epistemic positioning. With the exception of EPISTEMIC and TEMPORAL CERTAINTY, students use the AEMs grouped under the remaining five categories more frequently than the examiners. This is especially noticeable with the semantic grouping of POSSIBILITY where the student overuse is significant also in comparison with the *BNC* baseline. On the other hand, we can see that examiners clearly overuse AEMs grouped under EPISTEMIC CERTAINTY.

Many of the observed differences can be explained by different roles of the speakers and their role-related epistemic positioning. The paper therefore explores the use of the AEMs in the prototypical contexts paying particular attention to individual differences between speakers. In addition, the paper investigates features of successful and unsuccessful use of these markers.

References

Aijmer, K. 2002. "Modality in advanced Swedish learners' written interlanguage." *Computer learner corpora, second language acquisition and foreign language teaching*, 55-76.

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Brezina, Vaclav. 2012. "Certainty and uncertainty in spoken language. In M. Pütz, J. Robinson and M. Reif (eds.) In search of epistemic sociolect and idiolect." *Variation in language and use*, 97-128. Frankfurt am Main: Peter Lang.

Du Bois, J. W. 2007. The stance triangle. *In:* Englebretson, R. (ed.) *Stancetaking in discourse: subjectivity, evaluation, interaction.* Philadelphia: John Benjamins.

Holmes, J. 1988. Doubt and Certainty in ESL Textbooks. *Applied Linguistics,* 9**,** 21-44.

Holmes, J. 1984. Modifying illocutionary force. *Journal of Pragmatics,* 8**,** 345-365.

Kärkkäinen, E. 1992. "Modality as a Strategy in Interaction: Epistemic Modality in the Language of Native and Non-Native Speakers of English".*Pragmatics and language learning*, *3*, 197-216.

Simon-Vandenbergen, A.-M. & Aijmer, K. 2007. *The Semantic field of modal certainty. A corpus-based study of English adverbs,* Berlin, Mouton de Gruyter.

The effect of topic on language use in the Cambridge Learner Corpus

Andrew Caines
University of Cambridge
apc38@cam.ac.uk

Paula Buttery
University of Cambridge
pjb48@cam.ac.uk

1. Overview

We report the findings of a research project investigating the effect of topic on language use in learner corpora. We tagged a subset of documents from the Cambridge Learner Corpus (CLC) with topic labels — one of 'commerce', 'narrative', 'personal' or 'society' — and assessed the differences in language use between each topic type. We found that topic had an effect on lexis, subcategorization frames and language functions, but relatively little effect on part-of-speech frequencies or grammatical relations.

2. Motivation

The results are important both from a research and assessment point of view. By the nature of how they are collected, learner corpora tend to be made up of a plethora of topics. Researchers need to be wary, therefore, of treating corpora as homogeneous in topic terms.

However, this assumption has nevertheless been made in some previous research (e.g. Hawkins and Buttery 2010; Hawkins and Filipovic 2012). In these studies, the dependent variable is proficiency level: observations are made on, for example, language use at level A2 versus B1[2], without overt attention to extraneous independent variables such as topic.

As for assessment, the question is whether different topics afford exam candidates equal opportunity to demonstrate their language proficiency. This is particularly relevant to those examinations in which there is a choice of questions across diverse topic types. One way to answer this question is in terms of 'opportunity of use'.

For example, Buttery and Caines (2012) showed that opportunity of use is affected by document length. This finding is important for learner corpus research (LCR) because documents tend to lengthen with increasing proficiency, thus causing a confound.

Document length is therefore one of the variables that needs to be controlled for. So too is document

[2] These level descriptors are taken from the Common European Framework of References (commonly referred to as CEFR levels; http://www.coe.int/t/dg4/linguistic/cadre1_en.asp).

topic. We will outline a number of steps which can be taken to safeguard against interference from these variables in LCR.

The use of large-scale corpora to underpin second language acquisition research is now well-established. Broadly speaking, LCR seeks to reveal what learner language 'looks like' at different proficiency levels, from various perspectives including lexical, syntactic and discursive. We believe that the outcomes and impact of such research will be optimised if all variables other than the dependent one are controlled for.

3. Findings

In summary, we worked on a subset of exams from the year 2009 in the CLC and investigated the interrelation between topic and the occurrence of a range of lexico-syntactic features:

- We were able to classify documents into topic types with high accuracy based on vocabulary alone.
- There was no significant effect of topic on the frequency of nouns, adjectives, verbs or adverbs, though there was an effect of verb and adverb diversity.
- There was no significant effect for topic with regard to the 'grammatical relations' between words, though there was some effect on 'subcategorization frames' (i.e. verb argumentation patterns).
- We showed that certain 'language functions' — ways of imparting information, structuring discourse, etc — are significantly affected by topic.

4. Implications

These findings might be of relevance to exams which feature a number of different topics. Take the example of an exam in which candidates are given a choice of 'personal' and 'narrative' type questions. We found that narrative texts tend to feature more verbs and adverbs than personal texts, but fewer adjectives and nouns — a difference which means that candidates opting to answer a narrative rather than personal question have a greater opportunity to demonstrate knowledge of verb groups and verb argumentation patterns and less opportunity to demonstrate language features based around the noun.

Above all, where they have not been controlled for in corpus design, such differences in opportunity of use as influenced by topic need to be accounted for in research and allowed for in assessment.

Acknowledgements

We thank Cambridge English Language Assessment for sponsoring this research.

References

Buttery, P.J. and Caines, A.P. 2012. "Normalising frequency counts to account for 'opportunity of use' in learner corpora". In: Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research.* Amsterdam: John Benjamins.

Hawkins, J.A. and Buttery, P.J. 2010. "Criterial features in learner corpora: theory and illustrations". *English Profile Journal,* 1 (1): e5.

Hawkins, J.A, and Filipovic, L. 2012. *Criterial features in L2 English: specifying the reference levels of the Common European Framework.* Cambridge: Cambridge University Press.

A linguistically informed, systematic, and in-depth approach to developing data driven language learning materials

Krassimira Charkova
Southern Illinois University
sharkova@siu.edu

Denitza Charkova
Plovdiv University
“Paisii Hilendarski
d.charkova@gmail.com

1. Rationale

The Data Driven Learning approach has been around for almost three decades (Johns, 1986; 1991) and a substantial number of English language teaching resources have incorporated corpus data with a focus on inductive language learning. However, pedagogical and empirical literature shows that DDL has revealed mixed results in terms of learner benefits (Braun, 2007; Boulton, 2009, Cresswell, 2007; Widdowson, 1998).

The reasons for these mixed effects of DDL are numerous and of different nature. Yet, we believe that there are three important reasons why DDL activities may not always be as effective as desired. The first reason is the lack of a good linguistic focus in such activities, i.e. language learners are often asked to discover patterns of random nature in one single exercise, such as noun-article uses, verb-noun agreement, adjective-noun collocations, etc.

Another reason is that such exercises usually end at the discovery stage without offering further opportunities for practice and use of the target lexical and grammatical items.

A third reason, as already described in McCarthy (2008), is that teachers are not prepared to be sophisticated consumers of existing materials and confident users of corpus data for the purpose of developing their own materials.

2. Principles and Phases of the Training

In this presentation, we will describe and illustrate our approach of training English language pre-service teachers of how to use corpus data in order to develop language teaching materials in a linguistically informed, systematic, and in-depth way.

Contrary to some beliefs about the implementation of DDL (e.g. Johns, 1991), where both teachers and learners discover language patterns at the same time, our method puts primary importance on teachers’ preliminary work in identifying the most appropriate language data for a target lexical or grammatical item. In other words, we believe that teachers not only should know the answers before the students, but moreover, should use these answers to develop language learning materials in the most engaging and effective way.

In our training module, we follow a six-phase process of teaching teachers how to develop corpus-based materials with a specific lexical or grammatical focus.

Phase One: Teachers learn how to use their metalinguistic knowledge of English in order to identify morphological, lexical and grammatical targets in reading passages or listening scripts which are good candidates for DDL. We emphasize on the fact that not all words or grammatical structures are worth the time to explore through DDL.

Phase Two: Once they have learned how to identify relevant morphological, lexical and grammatical targets, teachers are trained how to conduct corpus-based searches in order to generate the most pertinent data for a particular morphological, lexical or grammatical target. At this stage, they are encouraged to experiment with several different searches in order to find the best data for the target structure.

Phase Three: Having generated the concordance data, teachers learn how to examine the data critically in order to identify the most common patterns of morphological, lexical and grammatical uses. They are encouraged to validate their analysis through consultation with relevant reference materials. They also receive practice in editing concordance data in order to prepare data sets with an optimum number of the best examples, where inappropriate words, unnecessary symbols, and redundant information have been replaced or removed.

Phase Four: Teachers learn how to use the edited concordance data to develop learning materials which incorporate three connected stages:

*The Analysis Stage***:** The activities in this stage are based on principles of three language teaching approaches, including DDL, the Cognitive Approach, and the Lexical Approach. Students are given corpus data generated and edited by their teacher in order to discover patterns of use and meanings of morphological, lexical or grammatical targets. Students categorize their conclusions in summary tables and graphic organizers.

The Practice Stage: The purpose of this stage is to provide relevant contexts for students to apply the patterns and rules they have discovered in the Analysis Stage. The activities in this stage use novel corpus-based examples written in complete sentences and expanded contexts.

The Use Stage: In this stage, students are given the opportunity to use the morphological, lexical and grammatical collocations in creative individual, pair or team work.

Phase Five: Teachers present their corpus developed materials and receive structured peer and instructor feedback.

Phase Six: Teachers revise and edit their materials to implement the peer and instructor feedback.

References

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19 (3), 307-328, Cambridge University Press.

Boulton, A. (2009a). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, *35*(1), 81-106.

Cresswell, A. (2007). Getting to ‘know‘ connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 267-287). Amsterdam: Rodopi.

Johns, T. (1986) Microconcord: a language-learner’s research tool. *System*, **14**(2): 151–162.

Johns, T. (1991) Should you be persuaded – two samples of data-driven learning materials. In:Johns, T. and King, P. (eds.) Classroom concordancing. Birmingham University: *English Language Research Journa*l, 4, 1–16.

Widdowson, H. G. (1998). Context, community and authentic language. *TESOL Quarterly*, *12*, 705-716.

Using personal corpora independently: ‘too much effort’ or ‘a lovely friend’?

Maggie Charles

University of Oxford Language Centre

maggie.charles@lang.ox.ac.uk

Over the last two decades there have been many accounts of direct corpus use with language learners, particularly those studying academic writing in English (see Boulton 2010; Yoon 2011). To date, the majority of this work has evaluated the success of corpus consultation immediately on completion of an in-class corpus intervention. As Pérez-Paredes et al. (2013) note, fewer studies attempt to examine independent corpus use over the longer term. One notable exception is Yoon (2008), who tracks six students’ corpus use over a six month period; however, her study provides qualitative rather than quantitative data on continuing corpus use. Work by Charles (2013, in press) reports on independent corpus use by 40 students one year after their corpus course, finding that 70% of respondents were corpus users. However that study does not shed much light on the reasons behind use or non-use of the corpora.

To investigate this question further, the present paper reports on a new set of data from 72 students who responded to an on-line survey one year after using corpora on an EAP course. During this course, students built their own personal corpus from research articles in their field and used it to explore discourse functions in their discipline. The data reported here is part of the more wide-ranging survey and focuses particularly on factors that are likely to promote ongoing corpus use and those that may hinder or prevent take-up. The quantitative results are supplemented by qualitative data from interviews with two students, a corpus user and a non-user.

Respondents were first asked whether they had used their corpus at any time since the academic writing course had ended; 41 students (57%) had done so, while 31 (43%) had not. The two largest groups of users were those who were currently using their corpus (15, 37% of users) and those who had used their corpus in the past, but were not using it at the time of the survey because they were not doing any academic writing (16, 40%). Some users mentioned other factors which negatively affected their corpus use: preference for other resources, lack of time and lack of usefulness were each noted by 3 students (7%), while technical problems were important for just 1 user (2%). The findings for non-users presented a similar pattern: the biggest single reason for non-use was that respondents had not done any academic writing (11, 36% of non-users). Seven students (23%) preferred other resources and 6 (19%) did not find the corpus useful. Lack of time was cited by 4 (13%) and lack of experience by 2 (6%), while technical problems affected one student (3%).

The first conclusion that can be drawn from this data is that when students work independently, the need for writing resources such as corpora is likely to be sporadic. This has implications for the provision and content of corpus courses, as well as for their timing. Two specific factors that have a negative effect on continuing corpus use also stand out: lack of time and a preference for other resources, often perceived as quicker and more convenient to use. The importance of these factors is also underlined by student ratings of the potential disadvantages of their personal corpus: 41 students (58% of all respondents) considered lack of time and lack of convenience to be *very important* or *important*.

However the likelihood of take-up or rejection of personal corpus use also depends upon individual student concerns, as illustrated by the contrasting attitudes reported by two students. Ahmad[3] characterised his personal corpus as *'like having a lovely friend with you who can advise you any time you want.'* Piotr, however, concluded that personal corpus use *'just took too much time, too much effort'*. In accounting for this difference, it is noteworthy that Ahmad's self-reported writing needs were mainly lexico-grammatical. Thus it was worth his while to build up his corpus to over a million words and he incorporated corpus use as a proofreading tool within his pre-existing writing practices. Piotr, however, was mainly concerned with the coherence of his text. Since such issues of overall textual organisation lend themselves much less easily to corpus investigation, he saw little or no benefit in investing time and effort in corpus construction and use.

This paper presents and discusses further the data on factors affecting independent personal corpus use and draws out the implications for fostering corpus consultation among students.

References

Boulton, A. 2010. "Learning outcomes from corpus consultation". In M. Moreno Jaén, F. Serrano Valverde and M. Calzada Pérez (eds.), *Exploring new paths in language pedagogy: lexis and corpus-based language teaching*. London: Equinox, 129-144.

Charles, M. 2013. "Student corpus use: giving up or keeping on?" In A. Lenko-Szymanska (ed.) *TaLC10: Proceedings of the 10th international conference on teaching and language corpora*. Warsaw: Institute of Applied Linguistics, University of Warsaw. Available online at http://talc10.ils.uw.edu.pl/proceedings/.

Charles, M. in press. "Getting the corpus habit: EAP students' long-term use of personal corpora". *English for Specific Purposes*.

Pérez-Paredes, P., Sánchez-Tornel, M. and Alcaraz Calero, J. (2013). "Learners' search patterns during corpus-based focus-on-form activities". *International Journal of Corpus Linguistics* 17 (4): 482-515.

Yoon, C. 2011. "Concordancing in L2 writing class: an overview of research and issues". *Journal of English for Academic Purposes* 10: 130-139.

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

[3] Student names are pseudonyms.

Phrasal verbs or prepositional verbs: Which one is more difficult for Chinese EFL learners? A longitudinal-corpus-based study

Meilin Chen
Hong Kong Institute of Education
meilinchen8388@gmail.com

1 Introduction

Multi-word verbs (Quirk et al. 1985; Biber et al. 1999) are perceived as notoriously difficult for ESL/EFL learners because they are very often semantically non-compositional, polysemous, and stylistically complicated as they behave very differently in different registers (e.g. phrasal verbs are more commonly used in informal or spoken registers while prepositional verbs are highly frequent in different registers, especially in written or academic registers [see Biber et al. 1999]).

Previous empirical studies have repeatedly found that, regardless of their L1 background, learners tend to avoid using multi-word verbs when a single-word verb alternative is available (Dagut & Laufer 1985; Hulstijn & Marchena 1989; Laufer & Eliasson 1993; Liao & Fukuya 2004; Schmitt & Redwood 2011). However, learner corpus studies show a more complex picture of learners' use of multi-word verbs, with some learner populations using multi-word verbs very frequently in writing while others tending to use fewer in comparison with their native counterparts (e.g. Waibel 2007; Gilquin 2011; Chen 2013).

This study explored Chinese university students' acquisition of two sub-types of multi-word verbs from a longitudinal perspective, namely, phrasal verbs (e.g. *take up*, *turn out*, *take in*) and prepositional verbs (e.g. *deal with*, *come across*, *depend on*). The investigation focused on three aspects: 1) the number of multi-word verbs used in their writing; 2) the number of meanings in which one multi-word-verb form is used; 3) the appropriateness of multi-word verbs in the learner writing, i.e. whether the multi-word verbs used are appropriate in style.

2 Corpora

To fulfill the three aims mentioned in the previous section, a three-year longitudinal learner corpus was built by collecting argumentative essays (250-300 words) from a group of 130 English majors at a Chinese university. The data collection was carried out twice a year for three years; therefore, altogether six essays were collected from each student. The essays were collected under examination condition with no access to reference tools or additional materials.

The learner corpus was designed to be able to be divided into three sub-corpora, each representing a level of the learners' undergraduate studies. The first sub-corpora represents the learners' first year of studies, the second sub-corpora their second year of studies, and so on.

Apart from the learner corpus, two native expert corpora, i.e. the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) were also used to determine the stylistic behaviour of the multi-word verbs found in the learner corpus.

3 Research questions

Based on the aims of the study, three research questions were drawn to guide the data analysis. They are

1) Have the Chinese learners made any progress in acquiring phrasal and prepositional verbs, i.e. is there an increase in the number of phrasal and prepositional verbs used in their writing?
2) Do the Chinese learners show any progress in their semantic knowledge of phrasal and prepositional verbs, i.e. can they use a multi-word verb in more meanings in writing?
3) Can they use phrasal and prepositional verbs more appropriately in style as their study proceed?

The methods of data analysis are elaborated in the following section.

4 Methodology

To explore the phrasal and prepositional verbs in the learner writing, the learner corpus was first POS tagged by using the free online CLAWS 7 tagger provided by Lancaster University[4]. Next, all co-occurring of lexical verbs and particles or prepositions were extracted from the corpus using the Concord function of WordSmith Tools 5.0 (Scott 2008). However, the concordances automatically generated by WordSmith Tools included not only phrasal and prepositional verbs but also verb + particle free combinations or verb + prepositional phrases. For example, *act out* is a phrasal verb in "*These children **acted out** their anger in the previous game.*", but a verb + a prepositional phrase in "*...and they **acted** <u>out of concern</u> for anyone who may...*". A good example of verb + particle free combinations would be "*I don' t want to <u>venture out</u> when it' s pouring.*" as the verb *venture* could be

[4] The online free CLAWS POS tagger is available at http://ucrel.lancs.ac.uk/claws/.

easily replaced by verbs such as *go* without a dramatic change in meaning. The third step therefore involved a manual check in order to rule out verb + particle free combinations and verb + prepositional phrases.

To investigate the semantic behaviour of the multi-word verbs, three dictionaries were used: *Collins COBUILD Phrasal Verbs Dictionary* 2nd Edition (Sinclair *et al* 2002), *Longman Dictionary of Phrasal Verbs* (*Longman* hereafter) (Courtney 1983) and *Oxford Phrasal Verbs Dictionary for Learners of English* (Parkinson 2001). The dictionary meaning of each occurrence of a multi-word verb was recorded for future analysis.

For the first and second research questions, the results were analysed quantitatively. The overall frequencies of phrasal and prepositional verbs in the three sub-corpora were calculated to explore the answers to the first research question. For the second question, the average number of meanings of phrasal and prepositional verbs in the sub-corpora was calculated. As for the third question, a qualitative approach was taken. Stylistic behaviour of the high-frequency phrasal and prepositional verbs in the three sub-corpora were compared. The next section presents the results and findings.

5 Results and findings

Results from the quantitative and qualitative analyses reveal a complex picture of multi-word-verb acquisition by Chinese EFL learners. First, the overall frequencies of phrasal and prepositional verbs show that phrasal verbs seem to be more problematic for the Chinese EFL learners than prepositional verbs. Little increase in the use of phrasal verbs was found in their third-year writing. Moreover, an unexpected drop in phrasal-verb use was observed in their second-year writing. Their use of prepositional verbs, in contrast, grew steadily.

The stylistic analysis of the multi-word verbs shows a similar pattern, i.e. the learners have greater difficulty in acquiring phrasal verbs than prepositional verbs. The learners did not show much progress in their stylistic knowledge of phrasal verbs until the third year. However, their stylistic knowledge of prepositional verbs developed considerably.

Results from the semantic analysis, however, reveal a different picture. While the average number of meanings between phrasal verbs and prepositional verbs does not show a considerable difference, the types of meanings used in these two sub-types of multi-word verbs are very different. Nearly 80 per cent of the meanings of the phrasal verbs are figurative, while figurative use of prepositional verbs account for only slightly over 50 per cent of all prepositional verbs. The fact that nearly half of the meanings of prepositional verbs are literal rather than figurative may partly explain why the learners showed a greater progress in acquiring prepositional verbs than phrasal verbs. Literal meanings are semantically more transparent, hence may be easier for learners to acquire.

6 Conclusion

The findings indicate that the acquisition of multi-word verbs by EFL/ESL learners is a complex process. The learners may show more progress in acquiring certain aspects of multi-word verb knowledge but little progress in others. It also shows that multi-word verb acquisition is not a simple matter of knowing the meaning and the form. It involves many other types of knowledge. This study shows that the complex semantic behaviour of multi-word verbs seems to cause great difficulty for learners of English. Figurative meanings of multi-word verbs, for instance, may slow the learners' learning pace. These findings may be useful for teachers, teaching-material writers as well as lexicographers in preparing teaching or learning materials about multi-word verbs.

References

Liao, Y. and Fukuya, J. Y. 2004. Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, 54 (2): 193–226.

Chen, M. 2013. "Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students". *Internation Journal of Corpus Linguistics*, 18(3): 418-442.

Dagut, M. and Laufer, B. 1985. Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition*, 7: 73–79.

Gilquin, G. 2011. Corpus linguistics to bridge the gap between World Englishes and Learner Englishes. *12th International Symposium on Social Communication, Santiago de Cuba, January, 2011*. Available at: http://hdl.handle.net/2078.1/112509 (accessed August 2012).

Hulstijn, J. H. and Marchena, E. 1989. Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition*, 11: 241–255.

Laufer, B. and Eliasson, S. 1993. What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition*, 13, 35-48.

Liao, Y. and Fukuya, J. Y. 2004. Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, 54 (2): 193–226.

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

Schmitt, N. and Redwood, S. 2011. Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds.) *A Taste for Corpora. A tribute to Professor Sylviane Granger* (pp. 173-208). Amsterdam: John Benjamins.

Confronting EAP textbooks with corpus evidence: the case of academic listening

Katrien Deroey
University of Luxembourg
katrien.deroey@Ugent.be

1 Introduction

In this paper I will discuss the extent to which lecture listening textbooks reflect authentic lecture language. I will also demonstrate Sketch Engine, which allows you to easily retrieve target language from (academic) corpora, and FileMaker Pro, a database programme which I find extremely useful in processing concordances.

The degree to which EAP materials correspond to the demands of real lectures is arguably an important factor in their ultimate usefulness. As Thompson (2003, p. 6) notes, '[f]or EAP practitioners, a key issue is how to provide as accurate as possible a model of lecture organisation and help their learners to develop the skills to interpret organising signals'. To assess how representative organisational cues in EAP books are, I compare importance marking cues with those attested in the British Academic Spoken English corpus[5].

2 Corpus analysis

Importance markers identified through an initial close reading of 40 BASE lectures were retrieved from all 160 BASE lectures using Sketch Engine and supplemented with further markers attested in their cotext and the BASE word list. Additional markers from previous lecture research were also searched (Deroey and Taverniers 2012).

The investigation revealed a large variety of importance markers, the most common of which differ from those which usually appear in EAP materials. The markers were classified according to their orientation to either the participants or the content ('interactive orientation', Table 1) and their position relative to the highlighted point (Deroey 2013). Most are either content- or listener-oriented, and signal important points prospectively. The predominant markers by far were those of the type *the point is* and *remember*. These are potentially

[5] The recordings and transcriptions in this study come from the British Academic Spoken English (BASE) corpus, which was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. The corpus is available from the Oxford Text Archive http://ota.ox.ac.uk/headers/2525.xml.

multifunctional and less explicit than their far less frequently used prototypical counterparts containing adjectives (e.g. *the important point is*) or a listener pronoun (*you should note that*). It can be argued that students should therefore be trained in interpreting these prevalent, multifunctional cues alongside being exposed to markers reflecting the variety that exists in real lectures.

Interactive orientation	N	%
Content *the point is sound waves don't really interact*	363	46.4
Listener *remember South Korea is still classified as a NIC*	304	38.9
Speaker *i want to emphasize this*	79	10.1
Joint *now let us note what Descartes is doing*	36	4.6

Table1: Interactive orientation of importance markers: examples and frequencies (N=782)

3 Corpus evidence versus EAP textbooks

The EAP books I examined vary widely in their inclusion of importance markers and range of examples. Most include few and fairly prototypical importance markers (Lebauer 2010; Lynch 2004; Phillips 1999; Salehzadeh 2006; Sarosy and Sherak 2006), the origins of which are unclear. Three integrate research findings on lecture listening and/or include corpus data: Salehzadeh (2006), Kelly, Revell, and Nesi (2000) and Lynch (2004).

Salehzadeh (2006) uses some lectures from MICASE. ‘Emphasis’ cues are said to generally occur before a point, which is borne out by my corpus data. However, examples are very few and mainly prototypical (e.g. *the important thing here is…*, *what you don’t want to forget…*) and it is unclear whether these are corpus-derived.

Kelly, Revell, and Nesi (2000) relates listening skills to lecture excerpts from BASE. The chapter on distinguishing between more and less important information includes examples such as *The key point is…What’s crucial… is…*; *A point worth noting is…*; and *That’s… the main point here*. Examples are from the corpus but all contain adjectives and do not represent the predominant markers from this study.

The lectures in Lynch (2004) seem to have been organised for the course. Interestingly, his categorisation of importance markers (p. 39) closely resembles the one based on corpus data in Deroey (2013). Lecturers stress points by ‘speaking about the subject matter itself’ (e.g. *a basic point; the central problem is that…*); ‘speaking to the audience’ (*it’s important to bear in mind that…; remember that…, you shouldn’t lose sight of the fact that…*); or by ‘speaking about themselves’ (*I want to stress*). Lynch’s list of importance markers is the largest and most varied. Nevertheless, it is mostly restricted to fairly prototypical examples and it is not clear what the list is based on.

4 Conclusion

In short, I feel that much remains to be done to ensure that corpus evidence informs lecture listening materials so that students are better prepared for the demands of their course lectures. In the case of importance markers textbooks should contain examples of a wider variety of importance markers, and practise the interpretation of prevalent, potentially multifunctional markers.

References

Deroey, K. L. B. and Taverniers, M. 2012. “‘Just remember this’: Lexicogrammatical relevance markers in lectures”. *English for Specific Purposes* 31 (4): 221-233.

Deroey, K. L. B. published online 2013. “Marking importance in lectures: Interactive and textual orientation”. *Applied Linguistics*. doi: 10.1093/applin/amt029

Kelly, T., Revell, R., and Nesi, H. 2000. *Listening to lectures*. Warwick: University of Warwick.

Lebauer, R. 2010. *Learn to Listen, Listen to Learn, Level 2: Academic Listening and Note-Taking* (3rd ed.). New York: Pearson Longman.

Lynch, T. 2004. *Study listening: A course in listening to lectures and note taking*. Cambridge: Cambridge University Press.

Phillips, T. 1999. *Skills in English listening: Level 3*. Reading: Garnet Education.

Salehzadeh, J. 2006. *Academic listening strategies: A guide to understanding lectures*. Ann Arbor: University of Michigan Press.

Sarosy, P. and Sherak, K. 2006. *Lecture ready 2: Strategies for academic listening, note-taking, and discussion*. Oxford: OUP.

Thompson, S. E. 2003. “Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures”. *Journal of English for Academic Purposes* 2 (1): 5-20.

Bridging informal Massive Open Online Courses and formal English for Academic Purposes programmes with language corpora

Alannah Fitzgerald
Concordia University & The Open Educational Resources Research Hub
alannahfitzgerald@gmail.com

Shaoqun Wu, Ian H. Witten
University of Waikato
{shaoqun,ihw}@cs.waikato.ac.nz

Martin Barge, William Tweddle, Saima Sherazi
Queen Mary University of London
{m.i.barge, w.tweddle, s.n.sherazi}@qmul.ac.uk

1 Introduction

Massive Open Online Courses (MOOCs) provide a compelling opportunity for domain-specific language learning. They supply a large corpus of interesting linguistic material relevant to a particular subject area, including text, supplementary images (slides), audio and video. It follows then that these domain-specific corpora can also be used in formal English for Academic Purposes (EAP) programmes as well. Such corpora can be automatically analysed, enriched, and transformed into a resource that learners can browse and query in order to extend their ability to understand the language used, and help them express themselves more fluently and eloquently in that domain.

To illustrate this idea, an existing online corpus-based language-learning tool (FLAX) is applied to Open Educational Resources (OER), including openly-licensed Coursera and edX MOOC content, and Open Access (OA) research content for the development of domain-specific language collections for uptake by informal MOOC learners and formal EAP students. Open education acts as a bridge to formal education, and is complementary, not competitive, with it. This is one of a cluster of research hypotheses currently under investigation at the OER Research Hub for the development of open language corpora in FLAX.

2 Domain-specific corpora

The use of domain-specific corpora is a growing trend in language teaching and learning (e.g. Gabrielatos, 2005). Most corpora are based on particular domains, genres, or collections of certain types of document from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs and Barth, 2003). Among other aspects of language, the domain-specific corpora we are developing in FLAX will provide an excellent context in which to study lexical bundles (Biber & Barbieri, 2007; Biber et al, 2003, 2004) and collocations, a notoriously challenging aspect of productive language use even for quite advanced learners (Bishop, 2008; Nesselhauf, 2003).

The academic language collections in FLAX use an automated scheme that extracts salient linguistic features from academic text and presents them in an augmented text interface, designed for the non-expert corpus user (Wu & Witten, In Press). Rather than relying on complex search commands to query corpora within involved concordancer interfaces (which have been designed by and for the corpus linguist) FLAX links relevant tools and resources into streamlined online interfaces for the language learner. For example, FLAX connects academic collections in the language learning system to the Wikipedia Miner tool to extract key academic concepts and their definitions from Wikipedia articles (Milne & Witten, 2013) to assist with reading and vocabulary.

3 Open educational resources

OER were selected for re-use to demonstrate how the FLAX corpus tools can linguistically enhance MOOC and domain-specific content. Two demonstration academic English language collections in FLAX are currently under development.

One collection is based on virology courses and resources developed by Professor Vincent Racaniello of Columbia University. His lectures were already popular across a range of web channels, including iTunesU and YouTube, before being imported into the Coursera MOOC platform. These lectures, along with Racaniello's weekly podcast *This Week in Virology*, his academic *Virology* blog, and OA articles relevant to his virology courses, are all published under a Creative Commons Attribution licence for easy processing as a FLAX language support collection for the virology MOOC learners.

The other domain-specific collection in FLAX is centred on the re-use of OER for academic English for law. This collection is being developed for EAP students who will be following the Law Pathway Pre-sessional and the Critical Thinking and Writing in Law In-Sessional programmes at Queen Mary University of London this 2014-15 academic year.

Lecture transcripts and videos (streamed via YouTube and Vimeo) will be featured from four different MOOCs: Copyright Law at Harvard (edX), English Common Law at the University of London (Coursera), Age of Globalization from Texas at Austin (edX), and Environmental Law and Politics at Yale (OpenYale). Podcast audio files and transcripts from the University of Oxford's Law Faculty and the Centre for Socio-Legal Studies (OpenSpires) are also being added to the collection.

The spoken language sub-corpus will be paired with a written language sub-corpus made up of OA research articles, samples of student writing from previous Law Pathway Pre-sessionals and sections of EThoS law theses held at the British Library and written by Queen Mary law students.

4 Open systems design for the uptake of educational language corpora

Open corpus-based systems and resources like the academic English collections in FLAX have unique characteristics and challenges with regards to diffusion, adoption and integration. Insights from EAP teachers involved in the FLAX collections building process will be presented with respects to how they perceived and interacted with these open educational systems, as they exist and as we are designing them in this research. In addition to this, a deeper understanding of how to design, iterate, integrate and evaluate open technological systems in support of advanced approaches to language learning and instruction within the specific context of open educational resource initiatives will be shared for discussion with TaLC conference participants.

References

Biber, D., Conrad, S., & Cortes, V. (2003). "Lexical bundles in speech and writing: an initial taxonomy." In A. Wilson et al. (Eds.), Corpus linguistics by the lune (pp. 71–92). Frankfurt/Main: Peter Lang.

Biber, D., Conrad, S., & Cortes, V. (2004). "If you look at . . .: lexical bundles in university teaching and textbooks." Applied Linguistics, 25, 371–405.

Biber, D., Barbieri F. (2007). "Lexical bundles in university spoken and written registers." English for Specific Purposes, 26, 263–286.

Bishop, H. (2004) "The effect of typographic salience on the look up and comprehension of unknown formulaic sequences." In N. Schmidt (Ed.) *Formulaic sequences: Acquisition, processing, and use* (pp. 227-244). Philadelphia, PA, USA: John Benjamins.

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13 (6), 4-16.

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*. Retrieved on Nov 17, 2013 from http://jime.open.ac.uk/2012/18

Gabrielatos, C. (2005) "Corpora and language teaching: Just a fling or wedding bells?" *Teaching English as a second or foreign language*, 8(4). Retrieved Oct 21 2013 from http://tesl-ej.org/ej32/a1.html.

Milne, D. and Witten, I.H. (2013) "An open-source toolkit for mining Wikipedia." Artificial Intelligence, (194), pp. 222-239, January.

Nesselhauf, N. (2003) "The use of collocations by advanced learners of English and some implications for teaching." *Applied Linguistics*, 24(2), 223-242.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. International Journal of Instructional Technology & Distance Learning, 2(1).

Stubbs, M., and Barth, I. (2003) "Using recurrent phrases as text-type discriminators." *Functions of Language*, 10(1), 61-104.

Wu, S. and Witten, I.H. (In Press) "Transcending concordance: Augmenting academic text for L2 writing." Submitted to Journal of English for Academic Purposes.

Using corpus-based research and academic corpora for thesis writing

Lynne Flowerdew
(formerly HKUST)
flowerdewlynne@gmail.com

The main aims of this presentation are twofold. First, I explain how the findings of corpus-based research from the existing literature can provide useful insights for ESP materials design in the realm of thesis writing. Second, building on exemplary work in the field of corpus-driven learning (e.g. Charles 2011, 2012, 2013; Eriksson 2012; Lee and Swales 2006), I illustrate how a freely-available corpus can be used for hands-on corpus-driven enquiries. These two main aims will be elaborated in relation to workshops on PhD thesis writing, specifically the Discussion section, for science and engineering students at a tertiary institution in Hong Kong.

In view of the growing internationalization of universities around the globe, it is not surprising that thesis writing for L2 writers is assuming increasing importance in academic writing programmes (see Thompson 2013 for an overview). A very useful handbook guide looking at thesis writing from a genre-based perspective is the volume by Paltridge and Starfield (2007). The chapter on writing the Discussion section provided the initial starting-point for the ESP materials design, centred on the following rhetorical functions:

- a restatement or review of most important findings
- examples from the data illustrating important results
- whether the results were expected or unexpected / commenting
- comparison with other work / previous research
- explanations for the findings, and/or speculations
- the strengths / limitations of the study
- implications of the study (generalizations from results)
- recommendations for future research and/or practical applications

Corpus-based research on academic writing was also consulted for preparing the input material, which was greatly informed by Hyland's (1996, 2005) work on hedges, boosters and attitude markers for the functions of 'commenting on the results', 'offering explanations for the findings' and 'stating implications of the study'. Other corpus-based research found useful for informing the materials was the following: use of inclusive and exclusive pronouns (Harwood 2005; Kuo 1999), 'attended' and 'unattended' *this* (Wulff et al. 2012), and nouns with a metadiscourse function (Flowerdew 2003) (see Flowerdew 2012 for further references).

In the first session of the workshop, after some preliminary pen-and-paper exercises using authentic Discussion sections from theses to familiarise students with the main rhetorical functions found in this section, students were introduced to a freely available online corpus, the *Corpus of Research Articles* (CRA), comprising 5,609,407 words from high-impact journals across 39 disciplines. As the corpus can be searched by discipline and section, i.e. I-M-R-D, it lends itself very well to searches for lexico-grammatical items occurring only in Discussion sections (see Lin and Evans 2012 for a detailed description of the corpus). While research articles are a different genre to theses in terms of their purposes and audience, nevertheless, it can be argued that there are significant areas of overlap in lexico-grammar and rhetorical functions, and, as Geoffrey Leech has pointed out there are no perfect tailor-made corpora.

The CRA has its own phraseological search engine (Greaves 2009). The sub-corpus of Discussion sections comprises 2.3 million words, which was found to be adequate and not too overwhelming in terms of size, another key factor to consider in corpus consultation. Tasks were devised to familiarize students with the types of searches they could carry out and to illustrate the usefulness of corpus consultation. For several of the tasks 'probes' were provided as a lead-in to discovery of potentially useful lexico-grammar. Students were first asked, though, to supply appropriate phraseologies themselves. For example, for the function of 'showing comparison with previous work', students were asked to complete the following phrase: 'This finding is…'. They were then asked to compare their suggestions with concordance output, which yielded examples such as the following:

This finding is consistent with Meyer, Becker, and Vandehberghen...
This finding is consistent with observations in the theoretical ...
This finding is in agreement with those of previous studies.
This finding is similar to prior findings that federal relief...
This finding is supported by research conducted by…
This finding has been supported by …
This finding has not been previously established.

The hands-on tasks thus mediated between top-down and bottom-up approaches (see Charles 2007). Follow-up class discussion generated student queries related to prepositions and the difference in the use of tenses between present and present perfect. While not all the corpus examples related to the function under consideration (i.e. comparison with previous work), students were able to match other phrases, e.g. ‘This finding is not surprising given that previous behavioural…’ with other functions, in this case, commenting on the data and giving an explanation. Students also noted that non-standard English was found in some of the concordance output (see Flowerdew in press for a discussion on English as a lingua franca in ESP corpus-based pedagogy). In the second session of the workshop, students used the corpus to self-check phrases in their own theses and were also introduced to more sophisticated searches. The CRA site provides instructions on self-compilation of corpora so students were encouraged to do this.

References

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

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

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

Charles, M. 2013. “The power of personal corpora: students’ discoveries using a do-it-yourself resource”. Paper presented at the 7th Corpus Linguistics Conference, CL2013. Lancaster University, 23-26 July 2013.

Eriksson, A. 2012. “Pedagogical perspectives on bundles: teaching bundles to doctoral students of biochemistry. In J. Thomas and A. Boulton (eds.) *Input, Process and Products. Developments in Teaching and Language Corpora*. Masaryk University Press: Brno, Czech Republic, 195-211.

Flowerdew, J. 2003. “Signalling nouns in discourse”. *English for Specific Purposes* 22: 329-346.

Flowerdew, L. 2012. “Grammar and the research article”. In C. Chapelle (ed.) *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Flowerdew, L. in press, 2014. “Adjusting pedagogically to an ELF world: an ESP perspective”. In Y. Bayyurt and S. Ackan (eds.) *Current Perspectives on Pedagogy for ELF*. Amsterdam: De Gruyter Mouton.

Greaves, C. 2009. *ConcGram 1.0: a phraseological search engine.* Amsterdam: John Benjamins.

Harwood, N. 2005. “‘*Nowhere has anyone attempted to…. In this article I aim to do just that’.* A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines”. *Journal of Pragmatics* 37: 1207-1231.

Hyland, K. 1996. Writing without conviction? Hedging in scientific research articles. *Applied Linguistics* 17: 433-453.

Hyland, K. 2005. Stance and engagement: a model of interaction in academic discourse. *Journal of Pragmatics* 7: 173-192.

Kuo, C-H. 1999. “The use of personal pronouns: role relationships in scientific journal articles”. *English for Specific Purposes* 18: 121-138.

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

Lin, L. and Evans, S. 2012. “Structural patterns in empirical research articles: a cross-disciplinary study”. *English for Specific Purposes* 31 (3): 150-160.

Paltridge, B. and Starfield, S. 2007. *Thesis and Dissertation Writing in a Second Language: A handbook for supervisors*. London: Routledge.

Thompson, P. 2013. “Thesis and dissertation writing”. In B. Paltridge and S. Starfield (eds.) *The Handbook of English for Specific Purposes*. Oxford: Wiley-Blackwell, 283-299.

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

Advanced learner speech: The use of '*I (don't) think*' by L2 learners

Dana Gablasova
Lancaster University
d.gablasova@lancaster.ac.uk

1 Introduction

Compared to advanced learner writing, spoken language production of advanced L2 learners has so far received much less attention in research (De Cock, 2010). To address this gap, this study aims to contribute to our understanding of advanced learner speech. In particular, it examines the use of discourse markers (DMs) in spoken interaction. DMs fulfil several important functions in spoken interaction and are crucial for effective and smooth oral communication (Müller, 2005). It is for this reasons that they are of interest in language pedagogy. '*I think*' is a DM that has been studied previously (e.g. Fung and Carter, 2007). This study aims to contribute to the description and understanding of the use of '*I think*' by L2 users by examining the contextual characteristics that may contribute to a specific pattern of use of *'I think'* by L2 learners.

2 Method

The corpus used in this study consists of 10 transcribed examination sessions with advanced speakers of English taken from the GESE exam developed and administered by the Trinity College London (Trinity College London, 2013). Table 1 provides a brief description of the corpus. All examiners were native speakers of English and all test-takers were non-native speakers of English.

	Learners	Examiners	Whole corpus
Types	2,226	1,546	2,813
Tokens	22,652	13,845	36,497

Table1: Data used

3 Results and discussion

The use of *'I think*' in the current data confirmed the overuse by non-native speakers (learners) as compared to the native speakers (examiners). Whereas learners used *'I think*' on average 196 times per 10k, the examiners used it considerably less (30 per 10k). The difference was statistically significant at p<.001 (LL-score: 68.15).

In order to further explore the use of '*I think*', the speech surrounding all instances of *'I (don't) think*' used by learners was examined in greater detail. To illustrate the findings, two examples of learner-examiner interaction are used. Dispersion plotting was used to visualise the interaction between each examiner and learner. Figure 1 shows these two examples of learner-examiner interaction. Red colour indicates all instances when the learner used *I (don't) think*. In the first case (File 1), the learner used *I (don't) think* 42 times, while the examiner used the expression twice. In the second example (File 2), the learner used *I (don't) think* 40 times, while the examiner did not use it at all. As can be seen from these dispersion plots, both learners use '*I (don't) think*' very often in their speech. However, it could be argued that whereas the second learner shows a typical overuse of *I think* by the learner, the production of the first learner is more complex.

In order to see the difference, the instances where the examiner used the expression *you think* (e.g. *do you think, so you think, you think that* and *don't you think*) during the conversation need to be explored (black lines in Figure 1). Whereas in the case of the second learner (File 2), there are only 6 instances of this use, in the first example (File 1), the examiner employed this expression 13 times. It can be seen from the dispersion plot (File 1) that several uses of *I think* by the learner can be considered as a response to the examiner's prompt (e.g. *do you think*) as the examiner's use of *you think* is repeatedly closely followed by the learner's *I think.* Also, the student uses *I think* several times in response to the prompt, thus increasing the number of occurrences of *I think* in the learner data. This can be seen in the example below taken from File 1.

Example of the pairing of *you think* and *I think*

Examiner: erm so why is football so popular all over the world why do people really- why are people so addicted to football **do you think**
Student: **I think** because you feel erm erm like <.>

4 Conclusion

These data suggest that the overuse of some items (e.g. *I think*) observed in learners' speech could be in some cases explained by the interaction of the two speakers. In this particular study which investigated test-taker/examiner interaction, the examiner's speech could be priming the use of a particular DM. These findings stress the complexity of the use of DMs in speech, which is co-constructed by the two interlocutors.

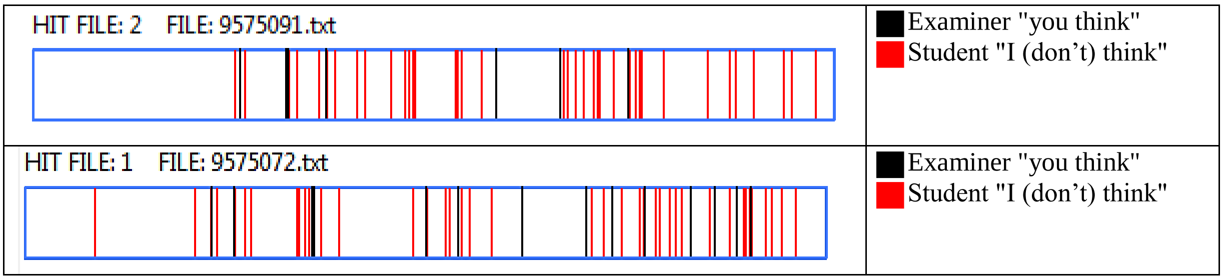


Figure 1 ‘I think’ and ‘you think’ in the learner-examiner interaction

References

De Cock, S. 2010. Spoken learner corpora and EFL teaching. In Campoy, M.R., Belles-Fortuno, B. and Gea-Vallor, M.L. (Eds.) Corpus-based approached to English language teaching. London: Continuum.

Fung, L and Carter, R. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, 28*(3)*, 410-439.

Müller, S. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.

Trinity College London 2013. *Exam Information: Graded Examinations in Spoken English (GESE)*. London: Trinity College London.

Towards a cross-linguistic analysis of perception of tone in academic textbooks

Melody Geddert
Kwantlen Polytechnic University
melody.geddert@kpu.ca

1 Introduction

Language learning at best is an arduous process. Coupled with trying to achieve success in academic areas, the task at times might seem insurmountable. Well aware that students might need some relief from the constant onslaught of information, textbook writers at times insert some degree of light-heartedness to keep the reader engaged. The concern this study addresses is the problem of language students' not recognizing tone that might be interpreted as something other than serious in academic reading materials in the English language, and whether or not linguistic group could offer any insight. Effective inter-cultural communication in a global context has now become imperative as the number of students studying in English internationally has become significant.

2 The problem

As many academic instructors can attest, the ability to identify tone in textbook reading often goes undetected by many English Second Language students, thereby reducing the students' understanding that not all of their educational experience is dull and dry.

Language instructors are often perplexed by students' inability to identify humour as a tone in some reading assignments. Many academic writers do inject a little light-heartedness here and there in order to make their writing a little less dull and hopefully to foster an interest in whatever information it is they are trying to transmit. When students are asked to read a passage or essay and then to identify any parts that seemed humorous- or less than serious- the task is not always possible for everyone. Even when the vocabulary and syntax are relatively simple, the ability to recognize the correct tone is still elusive. Perhaps the problem sometimes rests in such culturally contrasting senses of humour that it is impossible for some students to perceive the language as humorous or entertaining. Sometimes, because of pre-existing cultural schema, humour in a textbook would be completely unexpected. Although a student may still understand the information, not recognizing the author's effort makes the reading experience just a little bit less enjoyable and the reader perhaps a little less engaged than might be possible.

Research abounds on many aspects of cross-linguistic differences in humour, particularly joke telling (Attardo, 1994). Ample research also exists on how language learning can be facilitated by the incorporation of humour (Bell, 2009). However, there seems to be little information specific to cross-linguistic differences regarding tone recognition in academic materials.

3 The study

This paper presents an overview of an empirical study done at a Canadian university examining responses from approximately 400 first year university students from 14 different linguistic groups as related to perceived differences of humorous tone in academic textbooks from a range of subject areas. The questionnaire used for the study is a hand built corpus of actual passages taken from first year academic textbooks, first piloted on faculty to assure the humour value. The study then analyses the results for specific areas of difference while applying theories of formulaic language to account for some of the problematic items. This study provides some empirical evidence that when learners are not yet highly familiar with the usual contextualized phrases of a language, it is difficult to sense when register variations for the purpose of humour or some other engaging-type language have occurred. This paper proposes the need for much larger collections of humour derived from textbooks. A sizeable natural databank could help teachers give students the skills to better appreciate textbook authors' intentions. This research is in further development to a paper published in *Language and Humour in the Media,* Cambridge Scholars, 2012, which applied sociolinguistic schema theory to a prior smaller sampling that addressed both literary and non-literary texts.

References

Attardo, S. 1994. *Linguistic Theories of Humor.* New York: Mouton de Gruyter.

Bell, N. 2009. "Learning About and Through Humor in the Second Language Classroom". *Language Teaching Research* 13 (3): 243.

Investigating learning context in spoken learner corpora: Making use of learner profiles

Sandra Götz
Justus Liebig
University Giessen
Sandra.Goetz
@anglistik.uni-giessen.de

Joybrato Mukherjee
Justus Liebig
University Giessen
Mukherjee
@uni-giessen.de

It has been noted by various scholars that second-language acquisition (SLA) theory needs to take into account learning context as a determining factor in the acquisition/learning process (cf. Norris and Ortega 2001). However, SLA theoreticians face a number of challenges in systematising the infinity of individual learning processes and the multitude of specific learning contexts that affect and involve learners. These challenges are caused by the fact that learning context, defined by Ellis (1994: 197) as "the different settings in which L2 learning can take place", is by no means a variable that applied-linguistic theory and description can easily come to grips with, as these settings are shaped by many different factors (ranging from the learner's L1 and the language learning situation in his country of origin to specific learning materials in the classroom context). For one, any specific learning context includes a variety of context-determining factors, secondly – and more importantly –, one always needs to find methodological shortcuts, as it were, in order to reduce the sheer complexity of the variable context (and its effects on the learner) by abstracting away a limited number of context types and, hence, by setting up certain typologies of learning contexts for theory and description.

It is clear that a complete and detailed log of the entirety of a language learner's exposure to the second/foreign language at hand can only be produced, if at all, in a longitudinal study for very few individual learners. Thus, if we want to abstract from individual learners to the general picture of the output of a greater number of typical learners, we need to take into account comparable data from many learners that have been exposed to a defined set of typical and representative learning contexts. This is one of the central rationales in compiling learner corpora as repositories of learners' 'natural language use' (cf. Ellis 1994; Granger 2002). There are great differences in proficiency gains depending on the learning context and other sociobiographic variables. However, empirical research into interrelations between learning contexts and language learners' performance so far has mainly focused on a small number of learners and/or a very restricted set of context-related variables. This is due to the multitude of factors that are relevant when analysing learning contexts, and the lack of standardized (learner) corpora with context-related meta-information on the learners and their language-learning experiences.

Thus, in the present paper, we would like to demonstrate how research into learning context can now benefit immensely from the advent of relevant large-scale (learner) corpora. For spoken Learner English, the most important learner corpus is the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; cf. Gilquin et al. 2010), which includes data of learners with 11 different L1s, and also includes detailed learner profiles. These profiles provide information about each individual learner in the corpus and his/her language-learning experience, i.e.

- sociobiographic data: gender, age, nationality, first language, current studies, current year of studies;
- languages the learners have been exposed to: mother's/father's first languages, dominant language spoken at home, other languages spoken at home;
- target-language teaching setting: medium of instruction – primary (school), medium of instruction – secondary (school), medium of instruction – primary (university), medium of instruction – secondary (university);
- number of years of instruction at school and at university;
- time spent abroad in up to three English-speaking countries (name of country + age when time spent abroad + months of stay);
- other language(s) (first or foreign) spoken by the learner.

The present paper thus puts into perspective the potential that learner corpora offer for the analysis of the effect of learning contexts on learners' output and performance in the foreign language. We will start off by discussing some of the core issues that need to be addressed when describing and systematising learning contexts from a learner-corpus perspective and will proceed to an in-depth discussion of some case studies we conducted on the German component of LINDSEI that present different facets of making use of learner profiles in the analysis of learning context variables. The results of our case studies reveal interesting findings: While some learning context variables show clearly visible correlations with the learners' performances, e.g. there is a high positive correlation with the learners' fluency performance and the time they have stayed abroad, other variables prove to have no predictive power of the learners' performances at all, e.g. the number of years learned English has no correlation at all with the learners' degree of accuracy.

Finally, we will offer some conclusions and sketch out some avenues for future research.

References

Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

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

Granger, S. 2002. “A bird’s eye view of learner corpus research”. In S. Granger, J. Hung and S. Petch-Tyson (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. 3-33.

Norris, J.M. and Ortega, L. 2001. “Does type of instruction make a difference? Substantive findings from a meta-analytic review”, *Language Learning* 51 (s1): 157-213.

Making XML painless in the teaching and training of beginning corpus linguists

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

1 Introduction

In a recent position paper (Hardie 2014) I argue for a modest approach to the application of XML in corpus markup and annotation – in contrast to the heavyweight XML standards, such as TEI and CES, hitherto most commonly recommended as standard practice in corpus construction.

In this paper, I refine my earlier suggestions for practice in this area into a more concrete programme for the teaching/training of students and researchers in corpus linguistics. After a brief summary of the overarching argument for the “Modest XML” approach, I outline a series of stratified levels of awareness and practice that could be incorporated into pedagogy within a typical undergraduate or postgraduate course on corpus linguistics. This organisation into levels – where every level is valuable in itself as preparation for typical corpus-methodological practice, even if a student stops at that level and continues no further – offers a way-in to XML competence suited to any size of module or course in corpus methods. Each level includes explicit mention of tools relevant to that level of awareness.

2 Modest XML: a summary

XML is among the most widely used systems of textual markup in the world, and certainly is the most common choice for markup of corpora, and endorsed as such by many of the standard sources on good practice in corpus construction (see Wynne 2005). However, the two most prominent standards for XML corpus markup, the Text Encoding Initiative (TEI; Burnard and Bauman 2013) and the Corpus Encoding Standard (CES; Ide 1996) are both extremely complex and heavyweight systems. As such, they are best suited for large corpus construction projects designed to produce widely reusable resources such as the British National Corpus, which is encoded as TEI.

However, due to today’ easy availability of very large amounts of electronic text – primarily but not solely via the medium of the World Wide Web – such projects no longer represent the typical corpus construction undertaking, as they did in the 1990s. Today, it is common for individual researchers to create corpora for their own research purposes, with no expectation that the resource will ever be used by anyone else. For such undertakings a much more modest level of XML good practice is preferable. This would consists of two elements. First, knowledge of the underlying rules of XML elements, attributes and entities; the encoding of open, close and empty tags; and overall document structure, most notably the principle of perfect nesting within a single document-level root element. Second, suggestions regarding *de facto* standardised elements and attributes. The proposed list of elements includes: for written texts – p, s, head, pb, q, gap, and reg; for spoken texts – u, pause, voc, event, and stage; and for either type of text – unclear, header, body, text, w, c, and anon. The attributes that may be considered *de facto* standard include the general attributes id, n, desc and dur, as well as element-specific attributes such as level (on head), orig (on reg) , and who (on u).

I argue that the above basic knowledge regarding good practice is both (a) sufficient for most researchers' needs and (b) readily teachable to linguists at all levels from undergraduate students up to full academics. More complex aspects of XML requiring substantial technical and/or programming expertise – such as Document Type Definitions/Declarations, XML Schema, XML Stylesheets, and XPath – are explicitly and deliberately excluded.

The *Modest XML for Corpora* suggestions (Hardie 2014) attempt to encapsulate this level of good practice; these suggestions are recast below as a three-level teaching programme.

3 Teachable levels of knowledge and practice

The suggested approach to teaching good XML practice to beginning corpus linguists consists of the following levels.

First level:

- *Knowledge*: what XML is; how to spot it; common pitfalls when plaintext is used with XML software
- *Practice*: Use of entities for &, < and >
- *Software*: Configuring programs such as WordSmith/AntConc to ignore XML

Second level:

- *Knowledge*: Rules of XML elements and document structure
- *Practice*: Use of <text>, <p>, <s>, <u>, <gap> and a few other *de facto* standard elements
- *Software*: Using a web browser to check document validity; using a text editor with XML syntax highlighting.

Third level:

- *Knowledge*: Rules of attribute-value pairs
- *Practice*: Use of standard attributes such as id, n, and who. Good practice for using attributes for analytic markup, especially in spoken data and for pragmatic markup
- *Software*: Using regular expression capable software to search for elements and attribute-values.

References

Burnard, L. and Bauman, S. (eds.). 2013. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.5.0. Last updated on 26th July 2013*. TEI Consortium: available online at http://www.tei-c.org/Guidelines/P5/

Hardie, A. 2014. "Modest XML for Corpora: not a standard but a suggestion". *ICAME Journal* 38: 73-103.

Ide, N. 1996. *Corpus Encoding Standard. Version 1.5.* Expert Advisory Group on Language Engineering Standards (EAGLES): available online at http://www.cs.vassar.edu/CES/

Using Wmatrix to explore individual variation in L2 writing

Chunyu Hu
Guangdong University of Foreign Studies
gwhcy@gdufs.edu.cn

Learner corpus research is an area attracting more and more scholarly interest over the past two decades. The popularity of corpus method lies in the fact that learner corpora can make it possible to investigate some aspects of learner language which have previously been difficult or even impossible to explore (Granger 1998a). By combining insights from SLA theory and EFL teaching practice with a corpus linguistic methodology, researchers are able to describe interlanguage features and suggest implications for language teaching with greater confidence than has hitherto been possible.

Nevertheless, the design of current learner corpora is problematic in a number of ways. One of the problems is that most learner corpora, if not all, consist of spoken or written materials produced by groups of learner, which enables researchers to get statistically significant results regarding "overuse" or "underuse" (Leech 1998), at the expense of ignoring individual variation in their language performance. Considering that individuals "may do and act linguistically in ways which are not reflected by group data" (Anshen 1975, cited in Newmeyer 2005: 230), generalizations about learning from corpus evidence are *sometimes* elusive and not always likely to hold regardless of individual differences. It is thus important to investigate individual data to compensate for the potential limitations of group data.

This paper sets out to explore individual variation in L2 writing by using a web-based tool Wmatrix, an automatic tagging software that is able to assign part-of-speech (POS) and semantic field (domain) tags, and to permit the extraction of key words, key POS and key domains by applying the keyness calculation to tag frequency lists. The merits of Wmatrix lies in the mere fact that it allows macroscopic analysis (the study of the characteristics of whole texts) to inform the microscopic level (focussing on the use of a particular linguistic feature) and thereby suggesting those linguistic features which should be investigated further (Rayson 2008).

The data used in this study are essays written by two English-major juniors in China, each of whom was required to write sixteen argumentative essays, each essay being of approximately 300 words finished within 40 minutes. The sixteen essays written by learner A is named as Par7.txt which contains 7,052 words, and the sixteen essays by learner B as Par10.txt consisting of 5,640 words, given that there were sixteen participants and the two students in this study were labeled as Part7 and Par10 respectively (see Hu 2011). The native speaker referent corpus used in this study is LOCNESS-US, the sub-corpus of LOCNESS (the Louvain Corpus of Native English Essays) which is a collection of essays written by undergraduate students in the United States (Granger 1998b). Considering that the high rate of errors in learner corpora affects the accuracy rate of POS and semantic tagging, the spelling errors were corrected.

The results show that although both learners shared similarities when comparing with LOCNESS-US (e.g. significantly underusing NN2 such as *arguments* and *claims*, the conjunction *that*, *etc*.), there exists a significant degree of individual variation in the composing behaviors of the two learners despite their similar learning background. At 99% confidence (or $p < 0.01$), the cut-off of 6.63 would indicate that there are 68 words and multiword expressions, 23 POS and 29 semantic domains significantly overused or underused between Par7.txt and Par10.txt.

One of the daunting problems is in deciding on whether the apparent individual variation in wiring reflects personality style or language proficiency. For instance, Par7 preferred to use past tense modals (such as COULD, MIGHT, and WOULD) whereas Par10 used COULD and WOULD only once and did not use MIGHT at all in her 16 essays. Is it because Par10 was less competent in using the English modals or because Par10 was more direct in argumentation? Other types of data (such as *Interview* or/and *Questionnaire*) are needed to address this question.

References

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

Granger, S. 1998b. 'The computer learner corpus: a versatile new source of data for SLA research.' In Granger, S. (ed.), 3-18.

Hu, C. 2011. *Constructing the Interlanguage Modal System: L2 Acquisition of Modality by Chinese EFL Learners*. Beijing: Science Publication.

Leech, G. 1998. Preface. In Granger, S. (ed.), *Learner English on Computer* (xiv-xx).

Newmeyer, F. 2005. A replay to the critics of 'Grammar is grammar and usage is usage'. *Language, 81 (1),* 229-235.

Rayson, P. 2008. 'From Key Words to Key Semantic Domains'. *International Journal of Corpus Linguistics* 13(4):519–549.

Assessing lexical sophistication measures of multi-topic Japanese EFL writing with controlled text length and genre

Ishii Takumi
University of Tsukuba
s1330049@u.tsukuba.ac.jp

1 Introduction

Although vocabulary has attracted highly growing attention in the field of studies on second language acquisition over the past few decades (Nation, 2013), productive vocabulary (i.e., how learners use their vocabulary in writing and speaking) has been relatively under-researched in contrast to receptive vocabulary (Daller et al. 2007; Nation and Webb 2011; Schmitt 2010). Many vocabulary researchers, therefore, emphasize the need for investigating learners' vocabulary use in production (e.g., Nation 2007; Read 2000; Schmitt 2010).

One of the orthodox standpoints of measuring productive vocabulary use is lexical richness, which may be mainly classified into lexical diversity (LD) and lexical sophistication (LS). Whereas there has been a thorough assessment of the reliability and validity of LD measures for over 70 years, those of LS still need assessing in detail.

2 Review of LS measures

One major approach to analyzing LS is to classify learners' output by making reference to the word-frequency level. The idea behind this approach is that more lexically proficient learners can produce more infrequent words in their output because word frequency functions as the main objective criterion for word difficulty (Laufer and Nation, 1995; Meara and Bell, 2001) and as one of the most influential parts in vocabulary acquisition (Nation and Beglar, 2007).

The traditional, basic LS measure is the Lexical Frequency Profile (LFP; Laufer and Nation 1995) which shows the percentage of words produced at different word-frequency levels. It is, however, difficult to use the LFP in comparative studies because the LFP provides four levels of word-frequency indices. Laufer (1995), thus, developed a simplified version of the LFP. It is called Beyond 2,000 and calculated as follows:

$$Beyond\ 2{,}000 = \frac{Number\ of\ word\ families\ beyond\ 2{,}000\ basic\ word\ families}{Total\ number\ of\ word\ families}$$

Next, Daller et al. (2003) integrated LS measures with LD measures, and proposed Advanced TTR ($\mathrm{A_{TTR}}$) and Advanced Guiraud ($\mathrm{A_G}$). Each of them is defined as follows:

$$A_{TTR} = \frac{\text{Number of advanced types(i.e., typesnot included in the basic word list)}}{\text{Tokens}}$$

$$A_G = \frac{\text{Number of advanced types}}{\sqrt{\text{Tokens}}}$$

Third, Meara and Bell (2001) invented the P_Lex. It provides a single index of the likelihood of infrequent word occurrence by computing the distribution of infrequent words in production. Fourth, Kojima (2010) created S. With text coverage proportion across frequency ranks, it estimates the overall frequency level of learners' output and learners' productive vocabulary size.

Another novel approach is to examine the psycholinguistic attribute indices of learners' output. The premise of this approach is that more lexically proficient learners can produce more unfamiliar, abstract, low-imagery, and non-associative words in their output because psycholinguistic word attributes also affect the difficulty and learnability of L2 vocabulary (de Groot and van Hell 2005; Ellis and Beaton 1993; Salsbury et al. 2011; Yokokawa 2006). Recently, an increasing number of researchers (e.g., Ishii 2014; Kusanagi 2013; Salsbury et al. 2011) have adopted this new approach and selected the following psycholinguistic attributes as LS measures from the Medical Research Council Psycholinguistic Database (Wilson 1988): (a) familiarity (FAM), how familiar a word is; (b) concreteness (CON), how concrete or abstract a word is; (c) imageability (IMG), how strongly a word arouses a mental image; and (d) meaningfulness (MNG), how many other words a word is associated with.

Admittedly, all of these measures have their own improvements, and thus their own advantages. They, however, seem unsatisfactory because their reliability and construct validity lack detailed assessment. Some previous studies did not report the reliability and construct validity of their measures, and others conducted such assessment, but without controlling text lengths, topics, or genres. Further, how similar or different these LS measures in the two different approaches remains unclear.

3 The present study

The purpose of the present study is to assess nine LS measures (i.e., Beyond 2,000, P_Lex, $\mathrm{A_{TTR}}$, $\mathrm{A_G}$, S, FAM, CON, IMG, and MNG) by (a) investigating their reliability and construct validity and (b) examining their relationship in multi-topic Japanese EFL writing with controlled text length and genre. For this purpose, essays written by 72 Japanese EFL learners with different proficiency levels were collected from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa 2013), one of the largest learner corpora consisting of East Asian EFL/ESL compositions. In the ICNALE, each learner produced two topics of essays.[6] The ICNALE strictly and carefully controls key writing conditions, including essay length[7] and genre.[8] Also, corresponding essays written by 72 English native speakers were selected.

4 Results and discussion

The results showed the following: (a) only four (i.e., A_G, S, CON, and IMG) of all the nine LS measures were moderately correlated across two topics, (b) only three (i.e., A_G, CON, and IMG) of the four measures could significantly distinguish between Japanese EFL learners and English native speakers in both topics of essays, and (c) some measures in the two approaches were moderately to highly correlated, but the correlations were not consistent across two topics. As for the examination of the reliability and construct validity of the LS measures, it is worthwhile noting that these results contradict previous studies. Not only do the results demonstrate the need for reexamining the reliability and construct validity of LS measures, but they also suggest that LS measures may be strongly affected by essay topics.

To further assess LS measures, future research requires testing their reliability and construct validity in longer and shorter texts on two or more different topics in two or more different genres. Moreover, learners' topic familiarity should be taken into consideration.

References

Daller, H., Milton, J. and Treffers-Daller, J. (eds.) 2007. *Modelling and assessing vocabulary knowledge*. Cambridge University Press.

Daller, H., van Hout, R. and Treffers-Daller, J. 2003. "Lexical richness in the spontaneous speech of bilinguals". *Applied Linguistics* 24 (2): 197–222.

de Groot, A.M.B. and van Hell, J.G. 2005. "The learning of foreign language vocabulary". In J.F. Kroll and A.M.B. de Groot (eds.) *Handbook of bilingualism: psycholinguistic approaches* (pp. 9–29). Oxford University Press.

Ellis, N.C. and Beaton, A. 1993. "Psycholinguistic determinants of foreign language vocabulary learning". *Language Learning* 43 (4): 559–617.

Ishii, T. 2014. "A qualitative approach to measuring lexical developments with psycholinguistic word attributes in Japanese EFL writing". *English Corpus Studies* 21: 1–17.

Ishikawa, S. 2013. "The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English". *Learner Corpus Studies in Asia and the World* 1: 91–118.

Kojima, M. 2010. "Atarashii lexical richness sihyou no teian: gakushuusha no sansyutsu goi reberu no suitei" [Proposing S as a new measure of lexical richness: estimating frequency levels of vocabulary produced by learners]. *English Corpus Studies* 17: 1–15.

Kusanagi, K. 2013, March. *Shinrigengogaku teki tokusei ni motoduku sanshutsugoi no hyouka* [Assessment of productive vocabulary with psycholinguistic word attributes]. Paper session presented at the NICE Symposium 2013, Nagoya University, Japan.

Laufer, B. 1995. "Beyond 2,000: a measure of productive lexicon in a second language". In L. Eubank, L. Selinker and M. Sharwood-Smith (eds.) *The current state of interlanguage: studies in honor of William E. Rutherford* (pp. 265–272). Amsterdam/Philadelphia: John Benjamins.

Laufer, B. and Nation, I.S.P. 1995. "Vocabulary size and use: lexical richness in L2 written production". *Applied Linguistics* 16 (3): 307–322.

Meara, P. and Bell, H. 2001. "P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts". *Prospect* 16 (3): 5–19.

Nation, I.S.P. 2007. "Fundamental issues in modelling and assessing vocabulary knowledge". In H. Daller, J. Milton and J. Treffers-Daller (eds.) *Modelling and assessing vocabulary knowledge* (pp. 35–43). Cambridge University Press.

Nation, I.S.P. 2013. *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I.S.P. and Beglar, D. 2007. "A vocabulary size test". *The Language Teacher* 31 (7): 9–13.

Nation, I.S.P. and Webb, S. 2011. *Researching and analyzing vocabulary*. Boston: Heinle, Cengage Learning.

Read, J. 2000. *Assessing vocabulary*. Cambridge University Press.

Salsbury, T., Crossley, S.A. and McNamara, D.S. 2011. "Psycholinguistic word information in second language oral discourse". *Second Language Research* 27 (3): 343–360.

Schmitt, N. 2010. *Researching vocabulary: a vocabulary research manual*. Hampshire: Palgrave Macmillan.

Wilson, M. 1988. "MRC Psycholinguistic Database: Machine-usable Dictionary, Version 2.00". *Behavior Research Methods, Instruments, & Computers* 20 (1): 6–10.

Yokokawa, H. (ed.) 2006. *Kyoiku kenkyu no tameno dai ni gengo deitabeisu: Nihonjin eigo gakushusha no eitango shinmitsudo (moji hen)* [Familiarity for English words of Japanese EFL learners (visual version): L2 database for education and research]. Tokyo: Kuroshio.

[6] Essay topics: (a) "Is it important for college students to have a part time job?" and (b) "Should smoking be completely banned at all the restaurants in the country?"

[7] Essay length: 200–300 words

[8] Essay genre: argumentation

Authorial presence in PhD theses and research articles

Elvan Eda Işık-Taş
Middle East Technical University
Northern Cyprus Campus
edaisik@metu.edu.tr

1 Introduction

Manipulating linguistic features that mark authorial presence in writing is an important aspect of pragmatic competence that needs to be mastered by L2 writers. This study is a contrastive analysis of the use of first person pronouns in PhD theses produced by Turkish L2 writers and in published RAs of expert writers in Applied Linguistics.

Language socialization is a process through which novices or newcomers to a community acquire the "knowledge, orientations, and practices" (Garret and Baquedano-Lopez 2002) that will help them gain "membership and legitimacy" (Duff 2007) in that group. It is realized through the mastery and adoption of appropriate voices and "discourse identities" (Ivanic´ 1998) as well as linguistic conventions associated with the target community.

Numerous studies so far have investigated the relationship between writing and construction of discursive identity in L2 graduate students' writing (e.g., Cadman 1997; Hilvala and Belcher 2001; Ivanic´ and Camps 2001; Abasi et al. 2006) and in research articles (henceforth RAs) in different disciplines (e.g., Tang and John 1999; Matsuda 2001; Hyland 2001; Helms-Park and Stapleton 2003; Martinez 2005; Harwood 2005; Hu and Cao 2011; Martín and León Pérez 2014). Findings in these studies suggest that NSE and L2 writers might have different rhetorical preferences in writing, which might make the L2 writers "vulnerable to the risk of violating communicative norms " (Hyland and Milton 1997, 126) of the target language.

Hyland (2002) ordered discourse functions of self-mentions along a continuum according to their degree of power. According to his categorization, "explaining a procedure" and "expressing self-benefits" are the least powerful functions, which are followed by "stating a goal/purpose". While "elaborating an argument," reflects some authorial power, the most assertive function is "stating results and claim" "through which writers make knowledge claims, foreground their distinctive contribution or commitment to a position" (1104).

Stance can be expressed in many ways including "grammatical devices, word choice, and paralinguistic devices (e.g., loudness, pitch, and duration and non-linguistic devices such as body position and gestures)" (Biber et al. 1999, 972). The focus of this study was first person pronouns, which according to Biber et al., (1999, 977) make the attribution of stance to the writer "explicit" and "overt".

2 Corpora and methods

Two corpora were compiled in this study. The 90117-word L2 writer corpus included introduction parts of 25 PhD theses in English written between 2007 and 2012 in six Turkish universities. The 23006-word corpus of RAs, which comprised introduction parts of 25 single-authored RAs published between the same years in international academic journals nominated by expert informants, was utilized as the reference corpus.

Hunston (2007) suggests that research in stancetaking should not only "identify patterns" but should also reveal "meaning groups" (28). In this respect, co-texts surrounding individual words were also examined. Both qualitative and quantitative data analyses methods, comprising frequency counts and hand-tagged text analyses were employed. First person pronouns and determiners, I, me, my, mine, we, us, our, and ours were searched for in the corpus using AntConc 3.2.4w, a concordance software program developed by Laurence Anthony of Waseda University. All instances were analyzed and categorized within the framework of Hyland's (2002) typology of functions of self-mentions in academic texts.

3 Results

As presented in Table 1, first person pronouns were used almost five times more in RAs than in the PhD theses. In contrast to RAs, in which self-mentions were frequently employed (in 19 out of 25 texts), in PhD theses, author presence was rarely (in 5 out of 25 texts) marked. As Table 2 shows, the authors of PhD theses preferred to use first person pronouns in relatively low-risk functions, namely in "explaining a procedure" and in "stating a goal/purpose". Authors of RAs, on the other hand, clearly marked their presence in writing by using first person pronouns in diverse functions including the relatively high-risk functions of "stating results and claims" and "elaborating an argument".

	Number of texts	Number of words	First person pronouns (per 5000 words)	First Person Pronouns (per text)
PhD Theses	25	90117	1.8	0.4
RAs	25	23006	8.6	1.4

Table 1: Frequency of first person pronouns and determiners in PhD Theses and RAs

Discourse function	Frequency in PhD Theses	Frequency in RAs
Explaining a procedure	7	11
Stating results and claim	-	10
Elaborating an argument	-	6
Stating a goal/purpose	3	13
Totals	10	40

Table 2: Discourse functions of first person pronouns (Hyland, 2002) in PhD theses and RAs

4 Discussion and implications

In contrast to the authors of the RAs, who used first person pronouns to front a powerful authorial presence, authors of PhD theses in this study rarely marked their presence in writing. Findings of this study might have implications for novice writers who would like to publish their research in academic journals and more specifically, who would like to recontextualize their PhD theses as research articles. In this respect, findings might be utilized in the supervision of graduate students to help them more effectively respond to the expectations of their discourse community.

References

Abasi, A., Akbari, N. and Graves, B. 2006. “Discourse appropriation, construction of identities, and the complex issue of plagiarism: ESL students writing in graduate school”. *Journal of Second Language Writing* 15: 102-117.

Biber, D, Johannsson, S., Leech, G., Conrad and S., Finegan, E. 1999. *Longman grammar of spoken and written English*. Pearson Education Limited. Essex.

Cadman, K. 1997. “Thesis writing for international students: A question of identity”. *English for Specific Purposes* 16: 3-4.

Duff, P. A. 2007. “Problematising academic discourse socialization”. In Marriott, H., Moore, T, and Spence-Brown, R. (eds). *Learning discourses and the discourses of learning*. Melbourne: Monash University ePress. DOI:10.2104/ld070001

Garrett, P. B. and Baquedano-Lopez, P. 2002. “Language socialization: Reproduction and continuity, transformation and change”. *Annual Review of Anthropology* 31: 339–361.

Harwood, N. 2005. “We do not seem to have a theory...The theory I present here attempts to fill this gap: Inclusive and exclusive pronouns in academic writing”. *Applied Linguistics* 26 (3): 343-375.

Helms-Park, R. and Stapleton, P. 2003. *Journal of Second Language Writing* 12: 245-265.

Hirvala, A. and Belcher, D. 2001. “Coming back to voice: The multiple voices and identities of mature multilingual writers”. *Journal of Second Language Writing* 10: 83-106.

Hu, G. and Cao, F. 2011. “Hedging and boosting in abstracts of applied linguistics articles: A comparative study of English- and Chinese-medium journals”. *Journal of Pragmatics* 43: 2795-2809.

Hunston, S. 2007. “Using a corpus to investigate stance quantitatively and qualitatively”. In Englebretson (ed.) *Stancetaking in Discourse*: Amsterdam: Benjamins.

Hyland, K. 2001. “Humble servants of the discipline? Self-mention in research articles”. *English for Specific Purposes* 20: 207-226.

Hyland, K. 2002. “Authority and invisibility: authorial identity in academic writing”. *Journal of Pragmatics* 34: 1091-1112.

Hyland, K. & Milton. 1997. Qualification and Certainty in L1 and L2 students’ writing. *Journal of Second Language Writing* 6: 283-205.

Ivanic´, R. 1998. *Writing and identity: The discoursal construction of identity in academic writing*. Amsterdam: Benjamins.

Ivanic´,R. & Camps, D. 2001. “I am how I sound: Voice as self-representation in L2 writing”. *Journal of Second Language Writing* 10: 3-33

Martín, P. & León Pérez, I.K. 2014. “Convincing peers of the value of one’s research: A genre analysis of rhetorical promotion in academic texts”. *English for Specific Purposes* 34: 1–13.

Martinez, I. 2005. “Native and non-native writers’ use of first person pronouns in the different sections of biology research articles in English”. *Journal of Second Language Writing* 14: 174-190.

Matsuda, P.K. 2001. “Voice in Japanese written discourse: Implications for second language writing”. *Journal of Second Language Writing* 10: 35-53

Tang, R. and John, S. 1999. “The ‘I’ in identity: exploring writer identity in student academic writing through the first person pronoun”. *English for Specific Purposes* 18: S23–S39.

Tell me what I'm missing: Helping language learners make useful comparisons through enhancing the features of concordancing software.

Stephen Jeaco
Xi'an Jiaotong-Liverpool University
smjeaco@liv.ac.uk

Over the past twenty years or so, a variety of studies have explored the use of corpora in language teaching. One thread running through this research is that helpful activities can be built around the comparison of synonyms and alternative word forms. These comparative activities can be found in the literature on suggested teaching approaches as well as in evaluations of concordancing tools. The thread is evident in one of the earliest papers on Data-Driven Learning where Johns (1991) explained that learners often came to a concordancer wanting to compare two words. Coniam (1997) provides a rich range of activities for teachers and they are all based on learners comparing the output of two or more words. Tsui (2004) describes six fruitful concordancing activities for the classroom and half of these focus on synonymy: near synonyms; words with a very close meaning; and words which have a common translation in the target group's first language. There is confidence that corpora can reveal clear differences between synonyms (Kaltenböck and Mehlmauer-Larcher 2005).

However, although it can be rewarding, feedback from learners suggests that they can find the discovery of differences between synonymous words both difficult and time-consuming (Yeh, Liou et al. 2007). Frustration can arise from a lack of effective search skills (Sun 2003). In order to successfully retrieve relevant data and see clear patterns, there are a variety of obstacles to be overcome. Firstly, multiple queries may be required for patterns to be discernable (Gaskell and Cobb 2004). Another issue is that language learners may not be able to readily think up suitable words for comparison. Once they have begun exploring a word, they may also find it difficult to know how to extend their exploration with further queries (Gabel 2001). Learners often lack the knowledge and understanding of vocabulary which is required in order for them to conflate results by exploring lemma or usefully compare the data of one type with those for other word forms.

Given the importance placed by teachers and researchers on the power of comparisons in Data-Driven Learning, it seems strange that little support is provided in most concordancing software to facilitate this. Both *WordSmith Tools* (Scott 2010) and *AntConc* (Anthony 2004) require use of multiple windows or saved results in order to view two sets of concordance lines or collocations simultaneously. While the *Sketch Engine* (Kilgarriff, Rychly et al. 2004) includes the *Sketch-Diff* function, only the summary word sketches are available in this view, and comparing actual concordance lines would require moving backwards and forwards between pages or having multiple tabs open in the browser. The *Sketch-Diff* query box also offers no suggestions or support and requires the same POS tag to be used for both nodes.

While the default handling of lemma or types varies between concordancers, they share a high expectation that the user can select the required query mode expertly and to be able to appreciate which other forms will be merged or excluded when they view the results. In terms of multi-word units, while concordance lines can be easily converted into lists of collocates, there is usually little support for a learner wanting to know whether searching for a phrase rather than individual words could be beneficial. The differences in priming that Hoey (Hoey 2005) has shown to exist for collocations and nestings suggests that helping learners compare single words with longer structures containing those words could be a rich field for exploration. Danielsson (2007) argues that multi-word units should be seen as the norm, and independent words as special cases.

This paper presents a number of concordancing support features which could be incorporated into corpus tools to aid language learners in the discovery of differences between words and their use in phrases. The features have been incorporated into a new concordancer for language learners which is currently being evaluated. However, the focus of the presentation is on the value of such support rather than a software demonstration. The features include a split screen design for comparing results side by side and algorithms for generating suggestions for comparison. Single word suggestions are based on shared stems, synonyms from either a mono-lingual resource such as WordNet (Miller 1995) or one automatically built up from a translation dictionary. Multi-word suggestions take into account asymmetrical position sensitive collocations. Feedback from users and log data showing actual use of each of these will be presented.

The paper has implications for possible further enhancements of concordancing tools and also suggests productive pathways which teachers could encourage learners to explore in the classroom whether the features are already built into the software or not.

References

Anthony, L. (2004). "*AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit*", Tokyo, Waseda University.

Coniam, D. (1997). "A Practical Introduction to Corpora in a Teacher Training Language Awareness Programme." *Language Awareness* **6**(4): 199-207.

Danielsson, P. (2007). "What constitutes a unit of analysis in language?" *Linguistik online* **31**(2/07): 18.

Gabel, S. (2001). "Over-Indulgence and Under-Representation in Interlanguage: Reflections on the Utilization of Concordancers in Self-Directed Foreign Language Learning." *Computer Assisted Language Learning* **14**(3-4): 269-288.

Gaskell, D. and T. Cobb (2004). "Can learners use concordance feedback for writing errors?" *System* **32**(3): 301-319.

Hoey, M. (2005). *Lexical priming : a new theory of words and language.* London : Routledge.

Johns, T. (1991). "Should you be persuaded: Two samples of data-driven learning materials". In T. Johns and P. King *Classroom concordancing*. Birmingham, England: Centre for English Language Studies, University of Birmingham. **4:** 1-13.

Kaltenböck, G. and B. Mehlmauer-Larcher (2005). "Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching." *ReCALL* **17**(01): 65-84.

Kilgarriff, A., P. Rychly, et al. (2004). "*The Sketch Engine*". 2003 International Conference on Natural Language Processing and Knowledge Engineering, Beijing.

Miller, G. A. (1995). "Word Net: A Lexical Database for English." *Communications of the ACM* **38**(11): 39-41.

Scott, M. (2010). WordSmith Tools. Oxford, Oxford University Press.

Sun, Y.-C. (2003). "Learning process, strategies and web-based concordancers: a case study." **34**: 601-613.

Tsui, A. B. M. (2004). "What Teachers Have Always Wanted to Know-and How Corpora Can Help". In J. M. Sinclair *How to Use Corpora in Language Teaching*. Amsterdam, Netherlands: Benjamins**:** 39-61.

Yeh, Y., H.-C. Liou, et al. (2007). "Online Synonym Materials and Concordancing for EFL College Writing." *Computer Assisted Language Learning* **20**(2): 131-152.

Teaching corpus linguistics: What should we be doing?

John M. Kirk

jk@etinu.com

The series of TALC conferences grew out ICAME conferences to explore how Corpus Linguistics as was being developed by the early-mid 90s could or should be brought into the university classroom and taught, and to consider how it might be done and what goals and practices might inform the activity. During those days when desktop computing was only in its infancy, there was an emphasis on skilling up students to become researchers in their own right. It was not hard to recruit enthusiasts for this pioneering enterprise, which provided students with a feeling of being let in on an important scientific development of paradigmatic importance. Enthusiasm, however, had always to be curtailed by modularisation and feasible assessment. By 1994, I was able to report on the approach which I had developed and which was proving attractive to students and feasible to manage and assess. At the same time as corpora and the software for exploiting them were rapidly developing with some serious rigour, so too was the scholarly literature which had come to harness corpus-linguistic techniques. By 1996, a long-awaited textbook finally appeared (McEnery & Wilson 1996), and there have been several others appearing periodically since. With this burgeoning literature, it became clear at least to me that ways had to be found to help students deal beneficially with this largely research-based literature. Students had to learn to be critical of every aspect of research methodology and, in due course, in 2002, my course materials for teaching critical skills became published. The situation has moved on rapidly and exponentially. Desktop and especially laptop hardware as well as broadband and ubiquitous communications on the one hand and online corpus resources of undreamed-of magnitudes and extremely fast search and output speeds have utterly revolutionised teaching opportunities, enabling much work to be undertaken outside of the university computer lab. Nevertheless, the ease of one-click operations has not replaced the need for rigorous manual analysis of data when needed, or rigorous critical reading of an unprecedentedly large research literature, increasingly published in an electronic format in addition to or instead of traditional hard copy. Some of these developments have been charted as ‘generations’ in the latest textbook (McEnery & Hardie 2012), which valuably raises many of the central issues germane to good corpus linguistics teaching. As TALC returns to

Lancaster, it would thus seem timely to reflect again on what it is that university teachers are doing pedagogically when they are teaching corpus linguistics.

References

Kirk, John M. 1994. ‘Teaching and Language Corpora: The Queen’s Approach’, in A. Wilson and A. McEnery (eds.) *Teaching and Language Corpora*. (University of Lancaster Department of Modern English Language and Linguistics Technical Reports) pp. 29–51.

Kirk, John M. 2002. ‘Teaching Critical Skills in Corpus Linguistics Using the BNC’. In eds. B. Kettemann and G. Marko, *Teaching and Learning by Doing Corpus Analysis.* Amsterdam: Rodopi. pp. 155–164.

McEnery Tony & Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony & Andrew Hardie. 2012. *Corpus Linguistics*. Cambridge: Cambridge University Press.

How do L1 and proficiency level influence L2 written production?

Yuichiro Kobayashi
Japan Society for the Promotion of Science
kobayashi0721@gmail.com

Mariko Abe
Chuo University
abe.127@g.chuo-u.ac.jp

1 Introduction

The application of large computational databases of written and spoken samples produced by language learners has been developed as a way to reveal the influence of one’s first language (L1) on interlanguage development. Many previous Contrastive Interlanguage Analysis (CIA) studies have used the International Corpus of Learner English (ICLE), which is one of the most well-known written learner corpora, with samples from learners from more than 20 native language backgrounds (Granger 1996). However, because a sufficiently large-scale learner corpus with proficiency level information based on an objective rubric has not been available to the public, few learner corpus-based studies have targeted East Asian learners of English.

2 Purpose

The present study aims to explore the influence of one’s first language and English proficiency on the writing in English as a foreign language. The researchers compare the written production of East Asian learners of English (in Hong Kong, Korea, Taiwan, and Japan) to examine the use of 58 linguistic features, such as vocabulary, parts-of-speech, grammar, and discourse particles. The following research questions are investigated to accomplish this purpose:

1) How do L1 and CEFR levels effect L2 written production?
2) What linguistic features distinguish L2 learner groups from different L1 background and different proficiency levels?

3 Procedure

We used the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa 2011), which is considered to be the largest East Asian composition database. Written compositions from 2,000 EFL learners of English were analyzed in terms of the 58 linguistic features in Biber’s (1988) list. Table 1 summarizes proficiency levels (CEFR levels) and L1s of learners.

CEFR Level	Hong Kong	Korea	Taiwan	Japan	Total
A2	2	150	58	308	518
B1.1	60	122	174	358	714
B1.2	104	176	122	98	500
B2	30	116	44	34	224
C1	4	36	2	2	44
Total	200	600	400	800	2,000

Table 1: Corpus size

This study applied the linguistic feature list used by Biber (1988), but instead of employing factor analysis as in his study, we used multivariate methods, such as correspondence analysis and hierarchical cluster analysis, since they are more suitable for investigating similarities among variables (Oakes 1998; McEnery and Hardie 2012). Correspondence analysis was used as a first step to identify analysis points for a detailed investigation, and a hierarchical cluster analysis was then conducted to classify the resulting groups of correspondence analysis into larger meaningful groupings. In addition, we used the Kruskal-Wallis test to identify linguistic features significantly used in each group.

4 Results and discussion

East Asian learners with different proficiency levels were classified into four groups: (1) Hong Kong learners and C1 level learners in Japan and Korea, (2) A2, B1, and B2 level learners in Japan, (3) B1.2, B2, and C1 level learners in Taiwan and B1.2 and B2 level learners in Korea, and (4) A2, B1.1 level learners in Taiwan and Korea (Figure 1). These results suggest that there is an influence of first language on the L2 written output of learners in East Asia. However, we can further conclude that there is also an influence of proficiency levels on the L2 written output of learners in Korea and Taiwan.

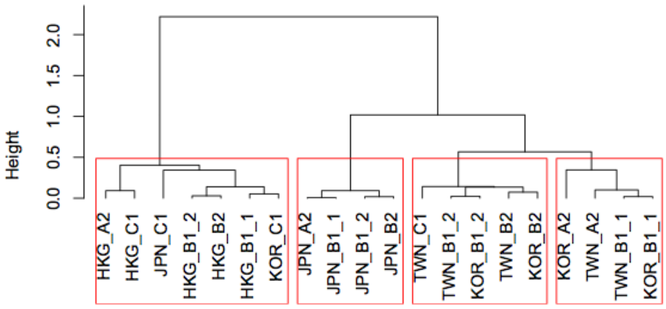


Figure 1. Clusters of learner groups

These clusters in Figure 1 are divided by the frequency of certain linguistic features. Therefore, the Kruskal-Wallis test was conducted to identify linguistic features that can distinguish the four clusters. As a result, learners in Hong Kong frequently used linguistic features characteristic of academic writing, such as nominalizations, conjuncts, and predictive modals. In contrast, learners in Japan made significant use of linguistic features characteristic of spoken language, such as first person pronouns, private verbs, present tense, and independent clause coordination. Moreover, learners in Taiwan and Korea, especially at novice levels, commonly used indefinite pronouns and emphatics.

5 Conclusion

The present study suggests an influence of first language and proficiency levels on the written production of East Asian learners of English. What is more, we were able to specify the linguistic features that can distinguish learners in four groups. A more detailed qualitative analysis may imply whether the learner language is influenced by universal phenomena or by developmental characteristics of specific L1s. However, as it is, this study contributes to our understanding of the nature and characteristics of learner language.

References

Biber, D. 1988. *Variation across speech and writing.* New York: Cambridge University Press.

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

Granger, S. 1996. From CA to CIA and back: An integrated contrastive approach to bilingual and learner computerized corpora. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.

Ishikawa, S. 2011. A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa and K. Poonpon (eds.) *Corpora and language technologies in teaching, learning and research* (pp. 3-11). Glasgow: University of Strathclyde Press.

McEnery, T. and Hardie, A. 2012. *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

Oakes, M. 1998. *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Can corpora help improve translation trainees results in specialised translation?

Natalie Kübler
CLILLAC-ARP,
University Paris Diderot, Paris Sorbonne Cité
nkubler@eila.univ-paris-diderot.fr

Mojca Pecman
CLILLAC-ARP,
University Paris Diderot, Paris Sorbonne Cité
mpecman@eila.univ-paris-diderot.fr

Alexandra Volanschi
CLILLAC-ARP,
University Paris Diderot,
Paris Sorbonne Cité
avolansk@eila.univ-paris-diderot.fr

1 Introduction

Over the last fifteen years, much has been written on the way corpora can be used to enhance translation students' production or raising tranlsation students' awareness on specific translation problems. Starting with Aston 1999 who presented how to use different kinds of corpora to help in the translation process or for translator education. Bernadini and Zanettin (2000) publish papers presented at the first Corpus Use and Learning to translate Conference. Zanettin et al. (2003) publish then a second monograph in which many papers show how to use corpora in different ways. Beeby et al. (2009) suggest in their introduction to the monograph they published two ways of looking at corpora, namely learning to use corpora to translate and learning to translate using corpora. Kübler & Aston (2010) list the different steps in the translation process in which corpora can be used : Documentation, drafting and revision.

As considerable research in translation studies has been focussing on the specific aspects of translated texts (Baker 1999; Puurtinen 2003, Olohan & Baker 2002; Olohan 2004; Mauranen 2007; Frankenberg-Garcia 2009), experimental research on how to use corpora in the translation process is more recent.

Bowker & Bennison (2003) propose a Student translation Tracking System to help students evaluate their own translation. Pearson (2003) uses a small parallel corpus in a specialised translation course to allow students to compare their translations with those of professional translators, with a view to raise students awareness on the strategies uses by professionals. In the framework of the MeLLANGE project, Castagnoli et al. (2011) approach a way to raise students' awareness of the different strategies adopted in the translation process by using in translation training a multilingual annotated learner translator corpus, which also contains professional translations.

2 Aim of the paper

This paper aims at demonstrating how corpora can help in the different processes involved in specialised translation. Research in terminology and translation (Bowker 1998, Maia (2002), Author1 2003, 2011, Sánchez-Gijón 2009) has confirmed that corpora are necessary tools for bilingual terminology and for getting familiar with the domain in the translation process. Over the years, we have compiled students' corpus analyses which show how they use corpora in the documentation, drafting and revision processes. Our aim is now to formalise this knowledge through a structured classroom experiment.

3 Methods

We are dealing with three groups of twenty students who each have to translate a 500-word extract of a very recent paper in earth science from English into French. They first have to find term candidates in their extract and compile a comparable corpus (English and French) of articles containing the term candidates they detected. All the individual corpora are then assembled in two common corpora, one in English and one in French and are saved on a server. Students will then use a customised version of the IMS Corpus Workbench (Evert & Hardie 2011) to query the corpus of ca. ten million words in each language. The first step consists in working on the bilingual terminology. All the English terms and their French equivalents are stored in the ARTES terminological and phraseological database (Author1 & Author2 2012) and are accessible to all. The second step consists in translating the extract of 500 words.

Each group of students will first have to translate the first 200 words only using the ARTE term base and no corpora. This will take place in a limited time frame. The translations will be annotated, using the MeLLANGE error typology (Castagnoli et al. 2011). In the next classroom session students will be specifically shown how to use corpora to help them in translating their text. In two classroom sessions, students will have to correct their translations using all the available corpora; they will be required to comment on how they used which corpus to improve their translations. This first batch of comments will be compiled in a 'comments corpus'.

The next three classroom s`essions will be dedicated to the translation of the remaining 300 words, using the ARTES database and the available corpora. Students will have to submit the remaining translation with a more thorough analysis of how they used corpora to help them in the translation process. The translations will be annotated.

4 Results

Statistics on error types will be made on the first annotated corpus of translations without corpora and on the second annotated corpus of translations with corpora. Previous observations have shown an improvement with the use of corpora. However, we expect statistical results to demonstrate it. The corpus of students' comments will also be analysed.

References

Aston, Guy 1999. 'Corpus use and learning to translate', *Textus*, 12, 289-313.

Baker, Mona 1993. Corpus Linguistics and Translation Studies – Implications and Applications. In *Text and Technology. In Honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (eds), 233-250. Amsterdam & Philadelphia: John Benjamins.

Beeby Allison, Rodriguez Inés Patricia and Sánchez-Gijón Pilar (eds). 2009 *Corpus Use and Translating*. Amsterdam: Benjamins

Bernardini, Silvia and Zanettin, Federico (eds) 2000. *I corpora nella didattica della traduzione/Corpus Use and Learning to Translate*, Bologna: Cooperativa Libraria Universitaria Editrice.

Bowker, Lynne & Bennison, Peter 2003. Student Translation Archive and Student Translation Tracking System. Design, Development and Application. In *Corpora in Translator Education*, F. Zanettin, S. Bernardini & D. Stewart (eds), 103-118. Manchester: St. Jerome.

Castagnoli, Sara, Ciobanu, Dragos, Kübler, Natalie, Kunz, Kerstin, Volanschi, Alexandra (2011)"Designing a Learner Translator Corpus for Training Purposes". In N. Kübler. (ed) (2011) *Corpora, Language, Teaching, and Resources : From Theory to Practice.* Bern: Peter Lang. 221-248.

Evert, Stefan and Hardie, Andrew (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Frankenberg-Garcia, Ana (2009) "Compiling and using a parallel corpus for research in translation". *International Journal of Translation*, vol. XXI-1, pp 57-71

Kübler, Natalie & Aston, Guy. 2010. "Using Corpora in Translation". in M. Mc Carthy & A. O'Keefe (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge. 505-515

Kübler, Natalie & Pecman, Mojca. "The ARTES bilingual LSP dictionary: from collocation to higher order phraseology". In *Electronic Lexicography*, S. Granger & M. Paquot (eds). Oxford: Oxford University Press.pp 187-209

Kübler, Natalie 2011. 'Working with different corpora in translation teaching'. In Ana Frankenberg-Garcia, Lynne Flowerdew, and Guy Aston (eds) *New Trends in Corpora and Language Learning*. London: Continuum. 62-80

Loock, Rudy 2013. « Using corpora to define traget-language use in translation », in Xiao, R. (ed), *Corpus-Based Contrastive and Translation Studies: Papers from the 2010 conference on Using Corpora in contrastive and Translation Studies.*

Maia, Belinda 2002. 'Corpora for Terminology Extraction – the Differing Perspectives and Objectives of Researchers, Teachers and Language Services Providers'. In Elia Yuste-Rodrigo (ed) *Proceedings of the Workshop Language Resources for Translation Work and Reaserch. LREC 2002*, Las Palmas de Gran Canaria, Spain.

Mauranen, Anna. 2007. 'Universal tendencies in translation', in M. Rogers and G. Anderman (eds) *Incorporating Corpora: The Linguist and the Translator.* Clevedon: Multilingual Matters.

Olohan, Maeve & Baker, Mona 2000. Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation?. *Across Languages and Cultures* 1(2): 141-158.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies,* London: Routledge.

Puurtinen, T. 2003. Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals?. In *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*, S. Granger, J. Lerot & S. Petch-Tyson (eds), 141-154. Amsterdam: Rodopi.

Sánchez-Gijón, P. 2009. Developing documentation skills to build do-it-yourself corpora in the specialised translation course. In *Corpus use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, A. Beeby, P. Rodríguez & P. Sánchez-Gijón, 109-128. Amsterdam & Philadelphia: John Benjamins.

Zanettin, F., Bernardini, S. and Stewart, D. (eds) (2003) *Corpora in Translator Education*, Manchester: St. Jerome.

Studying phrasal verbs in specialised genres

Maggie Leung
The Hong Kong Polytechnic University
magleung1217@gmail.com

Phrasal verbs have attracted a considerable amount of attention over the past thirty years due to their extensive use in English language and their syntactic and semantic characteristics. Previous studies have used different definitions of phrasal verbs (e.g. Bolinger 1971; Lindner 1981; Biber et al. 1999; Schneider 2004). This study adopts an inclusive approach in defining phrasal verbs by including both the combinations of a verb plus an adverbial particle, and a verb plus a prepositional particle which function as a single unit (Courtney 1983; Sinclair 1989; Halliday 2004).

Lexical words are often regarded as the element which contributes most to what a text is about. Phrasal verbs are special in the sense that they are formed by both a lexical word (the verb) and a grammatical word (a particle) which in combination contribute to the aboutness (Philips 1989) of texts. Phrasal verbs have been described as a common feature of English more frequently used in spoken or informal contexts. Much of the discussion in the literature is based on the use of phrasal verbs in general English, whereas the patterns of use of phrasal verbs in more specialised contexts are under-researched.

This study adopts a genre-based approach to examine the use of phrasal verbs in engineering English in terms of their frequencies, forms and functions, and examines the extent of genre specificity of phrasal verbs. Specifically, the study attempts to answer the following research questions: (1) Is the use of phrasal verbs the same or different across different engineering genres? (2) Do the inflectional forms of a phrasal verb have the same or different co-selections, and if so, does this impact the meanings? The study compares and analyses the most frequent phrasal verbs across 31 genre-based sub-corpora in a corpus of engineering English. In particular, it analyses the co-selections of phrasal verbs in different genre-based sub-corpora to examine the extent to which a phrasal verb may be specific to a specialised genre. The possible meanings which phrasal verbs may have in the genres are discussed in terms of the functions and characteristics of the genres.

The data used for the study are from the Hong Kong Engineering Corpus which consists of 9.2 million words of texts collected from the engineering sector in Hong Kong. This profession-specific corpus is formed by 31 genres used in the field, such as agreements, code of practice, consultation papers, reports, and speeches. The sub-corpora were annotated using Wmatrix (Rayson 2009) for part-of-speech tagging. The annotated corpora were then searched using WordSmith Tools (Scott 2004) for all the possible combinations of phrasal verbs. ConcGram (Greaves 2009) is used to perform concordancing as it shows all possible configurations of a single search in the concordances, namely contiguous instances (e.g. '*comply with*'), non-contiguous instances (e.g. '*comply* in all aspects *with*'), and positional variation (e.g. '*with* which companies must *comply*'). The most common phrasal verbs in each engineering genre are identified and compared in terms of their relative frequencies, and are analysed using Sinclair's (1996) five categories of co-selections.

Preliminary findings show that a phrasal verb may have different co-selections in different genres, and hence, different extended units of meaning. For example, *comply with* is found among the top ten most frequent phrasal verbs in both Reports and Standards, but the inspection of their co-selections suggests that the extended units of meaning of *comply with* in Reports is mainly evaluating whether requirements are met or not, whereas the co-selections in Standards are different and suggest a different meaning, i.e. obliged to act according to the authorized requirements. Moreover, it is found that the inflectional forms of a phrasal verb may have varied frequencies of occurrence and they do not necessarily share the same co-selections. It is thus argued that phrasal verbs should not be lemmatized in dictionaries or grammar books when presenting their frequency of occurrence, meanings and use.

There are pedagogical implications for teaching and learning English for specific purposes as the study provides insights into the authentic patterns of language use in the engineering industry. The findings of the study will raise learners' awareness of the genre specificity of phrasal verbs. It is argued that learners will have a better understanding of how different contexts may contribute to different units of meaning if they analyse the textual environment in which the phrasal verbs are used. The methodology and results of the study have research implications for examining the genre specificity of phrasal verbs in other registers or specialised corpora.

Acknowledgements

The work described in this paper is substantially supported by a research studentship funded by the Research Grants Council of the Hong Kong Special Administrative Region (Project No.: 4-RTY2).

References

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

Bolinger, D. 1971. *The phrasal verb in English*. Cambridge: Harvard University Press.

Courtney, R. 1983. *Longman dictionary of phrasal verbs*. England: Longman.

Greaves, C. 2009. *ConcGram 1.0: a phraseological search engine*. Amsterdam: John Benjamins.

Halliday, M. A. K. 2004. *An introduction to functional grammar*. (Third edition). London: Arnold.

Lindner, S. J. 1981. "A lexico-semantic analysis of English verb particle constructions with *out* and *up*". PhD dissertation. University of California.

Philips, M. 1989. *Lexical structure of text [Discourse analysis monographs 12]*. Birmingham: University of Birmingham.

Rayson, P. 2009. *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Schneider, E. W. 2004. "How to trace structural nativization: particle verbs in world Englishes". *World Englishes* 23(2): 227-249.

Scott, M. 2004. *WordSmith Tools version 4*. Oxford: Oxford University Press.

Sinclair, J. McH. 1996. "The search for units of meaning". *Textus* 9: 75-106.

Sinclair, J. McH. et al. 1989. *Collins COBUILD Dictionary of Phrasal Verbs*. London: Collins.

A corpus-driven study of corporate governance reports

Maggie Leung
The Hong Kong Polytechnic University
maggie.sn.leung@connect.polyu.hk

Martin Warren
The Hong Kong Polytechnic University
martin.warren@polyu.edu.hk

In recent years, corporate governance has become a major issue after a number of high profile cases in which companies were found to have a lack of or inadequate corporate governance, such as the collapse of Enron in 2001 and WorldCom in 2002, and, more recently, the money-laundering scandal involving HSBC in 2012. Corporate governance covers discipline, independence, fairness, transparency, responsibility, accountability, and social awareness (Gill 2002). Good corporate governance is considered essential to improve economic efficiency, enhance the market and investors' confidence, and maintain the stability of the financial system (Ho 2003: 55). In Hong Kong, as in many countries, it is important that companies adhere to strong corporate governance since it holds the key to sustaining the growth of the city's economy, stock market, supports investor confidence, attracts international capital, and creates liquidity (Chamber of Hong Kong Listed Companies, 2007). Listed companies are required by the Stock Exchange of Hong Kong to submit interim and annual corporate governance reports, and make these reports available on their websites. While the significance of corporate governance reports is well-recognised, there is a lack of research regarding the language and discourse organisation of this genre.

This paper offers a partial description of this relatively new genre by examining the corporate governance reports of a cross-section of major companies in Hong Kong in terms of their generic move structure, lexico-grammatical features, and phraseologies associated with the moves. It adopts the 'top-down' corpus-based approach to discourse analysis (Biber, Connor and Upton 2007) and combines both quantitative and qualitative approaches to discourse studies of language use. The study is collaborative with companies and professional associations in the financial services sector. Expert advice was sought for the design of study, data collection, data analysis, interpretation and dissemination.

The Hong Kong Corpus of Corporate Governance Reports is comprised of the corporate governance reports of companies listed in the Hang Seng Index which is the main index of the Hong Kong Stock

Exchange. The corporate governance reports were downloaded from the websites of listed companies in Hong Kong with the permission of the companies.

First, the rhetorical moves of each corporate governance report were identified to establish which were obligatory and which were optional. The move structure of the reports was then compared to the guidelines set out by the regulatory authority which include both mandatory disclosure information and recommended disclosures in order to determine the extent to which the companies simply meet or go beyond the stipulated requirements. The results of the move-structure analysis show that most of the companies simply comply with the regulations in terms of the information required to be presented in their corporate governance reports. Some of the companies included additional moves in their reports which are not required by the authorities, for example, Remuneration of directors and management, Business ethics, and Dividends. The sequencing of the moves was also established and the most frequent patterns will be described. The moves will be described along with the differences in sequencing and some of the more frequent move-specific phraseologies to determine the functions or motivation why companies include the additional information.

To carry out an analysis of the lexico-grammatical features and phraseologies used, the whole corpus and the move sub-corpora were examined using ConcGram (Greaves, 2009) to generate word lists and two-word concgram lists. The concordances of the two-word concgrams were then studied to determine which ones were meaningfully associated and which ones were simply chance co-occurrences. Some of the most frequent phraseologies in each of the moves are described and discussed in terms of their form, pattern and functions.

The findings of the study provide initial insights into the discursive practice and strategies adopted by listed companies in their corporate governance reports and the extent to which they are complying with or exceeding the requirements. The description and analysis of the patterns specific to corporate governance reports have implications for the learning and teaching of English for Specific purposes. The findings could be used to raise awareness and lead to a better understanding of this new genre. The methodology and findings might also inform other studies of specialised corpora and genre analysis in general.

Acknowledgements

The research described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5440/13H).

References

Biber, D., Connor, U and Upton, T. (eds.). 2007. *Discourse on the Move: Using Corpus Analysis to describe Discourse Structure*. Amsterdam/ Philadelphia: John Benjamins.

Chamber of Hong Kong Listed Companies. 2007. http://www.chklc.org/web/eng/index.htm, retrieved on 9 September 2009.

Gill, A. 2002. “Corporate governance in emerging markets - saints & sinners: who’s got religion?” Symposium on Corporate Governance and Disclosure: The Impact of Globalisation. The School of Accountancy, The Chinese University of Hong Kong, February 2002.

Greaves, C. 2009. *ConcGram 1.0: a phraseological search engine*. Amsterdam: John Benjamins.

Ho, S.S.M. 2003. “Corporate governance in Hong Kong: Key problems and prospects”, 2nd Edition, Copyright 2002, 2003. Centre for Accounting Disclosure & Corporate Governance. School of Accountancy, The Chinese University of Hong Kong.

Motivation and the learner corpus

Tim Marchand
J. F. Oberlin University
marchand@obirin.ac.jp

Sumie Akutsu
J. F. Oberlin University
smakutsu@obirin.ac.jp

1 Introduction

This paper explores the issue of motivation and the construction of a learner corpus from computer-mediated communication (CMC). The course in question is from an EFL context of university students in Japan where lesson materials and tasks are provided online through a news-based blog. Comments on the blog form the basis of a learner corpus, which in previous research has been analysed with reference to native speaker norms, allowing needs to be identified and addressed in subsequent materials (Marchand and Akutsu, 2013).

This paper examines the importance of motivation from three perspectives: on a global scale, to evaluate the motivational benefits of using CMC for pedagogical purposes; on a topic-by-topic level, to assess to what extent differences in the intrinsic level of interest in each lesson topic influences learner contributions to the corpus; and finally on an individual basis, to account for the importance of motivation as a learner variable in corpus construction.

2 Background

Recent reviews of the current state of learner corpus research have suggested avenues of future research to enhance our understanding of the field (Meunier, 2010). For example, there has been a call to expand the types of tasks and genres of learner data collected, some of which may better reflect the real-world forms of native-produced data often found in reference corpora (Granger, 2009). Meanwhile when compiling a learner corpus, the intrinsic and extrinsic motivation of individual students has remained an underexplored area of research (Ishikawa and Ishikawa, 2013) and one which, if unaccounted for, is likely to contribute to the notion of proficiency in corpus texts being a “fuzzy” variable (Carlsen, 2012). This paper seeks to broach these issues by examining certain motivational aspects that underlie the production of corpus texts of a new genre type, computer-mediated communication.

3 CMC and the Learner Corpus

Computer-mediated communication has been recognized as an effective way of connecting with the current population of students, since blogging and social networking are modes of communicating that many language learners use in their daily lives (Alm, 2006; Erbaggio et al., 2010). Therefore, from the outset of the research project, learner motivation was considered to be a key factor in the construction of a corpus based on CMC.

In this study, CMC takes the form of a blog where, each week, an article about a recent news item, is posted online for students to access. Students write their reactions to the story on the class blog, and these comments then form the basis of the learner corpus. Over the course of one academic year, learners are expected to have submitted at least 12 written reactions to the news articles, and after several years of running the course, the learner corpus is now approaching 200,000 tokens in size.

The size of the corpus, and the fact that the news stories themselves cover a range of topics which the learners may find intrinsically more or less interesting, offers the opportunity to examine how motivation and engagement with lesson materials affects learner performance in producing corpus texts.

4 Topic Interest and Learner Motivation

Three aspects of motivation will be addressed in this paper. Firstly, the overall motivational benefits of the CMC-based course will be discussed with reference to questionnaire data. The questionnaire inquired after how the learners’ experience of the course impacted on their own motivation and language learning.

Secondly, the question of topic interest and its effects on learner output will be explored. After submitting comments and reactions on the blog to news articles, learners were asked to rate their level of interest and knowledge on the news topics. Correlating the scores for each news topic with the corpus texts that are produced, the paper will explore whether increased engagement with lesson materials has any discernable effects on the nature of the written responses.

Finally, motivation as a learner variable will be discussed in terms of corpus construction. With reference to learner profiles obtained through questionnaire data, the paper compares and contrasts the corpus texts of identifiably motivated and unmotivated learners, arguing that learner motivation is an important construct that needs to be considered when exploring issues of proficiency levels within learner corpora.

The results suggest that not only is the use of CMC in the language classroom of sound pedagogical value for its ability to engage learner interest, but also that it presents a suitable platform for conducting research into the correlation between motivation and corpus texts. The paper concludes by affirming that learner motivation is an important factor that needs to be considered when constructing a learner corpus, such as in the deliberation of what prompts to use, if the goal of capturing genuinely authentic samples of learner language is to be realised.

References

Alm, A. (2006) CALL for autonomy, competence and relatedness: Motivating language learning environments in Web 2.0. *The JALT CALL Journal 2(3):* 29-38.

Carlsen, C. (2012) Proficiency level - a fuzzy variable in computer learner corpora. *Applied Linguistics 33(2*): 161-183.

Erbaggio, P., Gopalakrishnan, S., Hobbs, S., & Liu, H. (2010) Enhancing student engagement through online authentic materials. *International Association for Language Learning Technology 42(2).*

Granger, S. (2009) The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation. In: Aijmer, Karin (ed.) *Corpora and language teaching* (pp.13-32). Amsterdam & Philadelphia: John Benjamins.

Ishikawa, S. and Ishikawa, Y. (2013) How writers' personal attributes influence their L2 use: A study based on the ICNALE. Paper presented at the Symposium on Learner Corpora 2013, University of Padua.

Marchand, T. and Akutsu, S. (2013) New lines of research in computer-mediated communication – insights from a learner corpus of Japanese EFL students' writing. Paper presented at BAAL 2013, Edinburgh.

Meunier, Fanny (2010) Learner corpora and English language teaching: checkup time. *Anglistik: International Journal of English Studies 21(1)*: 209-220.

A joint venture: Cinderella and the Ugly Duckling of ELT acting together

Sanja Marinov
University of Split
smarinov@efst.hr

The aim of this presentation is to share an idea of combining two tools in a single exercise aimed at improving accuracy and raising awareness of both the foreign language being taught as well as the learners' mother tongue.

For this purpose a translation exercise is combined with a well-targeted concordance exercise. Thus the joint venture, where translation is seen as Cinderella and concordance exercise as the Ugly Duckling of modern language teaching.

It is interesting that different authors refer to different components of ELT as Cinderella. For Kelly (1969) it was pronunciation which was very important in the structural approach to teaching, where accuracy was upheld over fluency, but lost in importance with the introduction of communicative language teaching (Isaacs 2009).

Strangely enough, it was the idea of the "primacy of speech" that struck the first blow to eradicating translation from LT, threatening to create another Cinderella. It was, however, the orthodox Grammar Translation method that gave translation a bad name. The recognised inadequacy of the Grammar Translation method led to the establishment of the Reform Movement and Direct Method, the latter responding both to the need for an improved approach to language teaching as well as to the practical requirements of teaching multilingual classes in different language backgrounds by native-speaker teachers (Cook 2012).

Translation is not completely absent from ELT but what makes it a Cinderella is its position of an outlaw in the language teaching theory for around a hundred years (Cook 2012). This joint venture is an attempt to reintroduce it in the classroom because it "has been too long in exile .....It is time it was given a fair and informed appraisal." (Widdowson 2003:160).

"The process of translation is seen as a slow and laborious one, focused more upon accuracy than fluency." (Cook 2012:88). In this sense our Cinderella shares common characteristics with our Ugly Duckling. Concordance work is also focused on particular language issues, language system rather than on fluency and communication. The ultimate goal, however, remains the same: broadening language horizons and raising awareness of what there is in a language, making a shift in how we look at language and possibly contributing to developing more autonomous and self-sufficient learners. It is at least partly due to the ideas of the communicative approach to language teaching that a more focused language work and some sweating invested into combining language to get more favourable results is seen as strange to language teaching. Communicative approach shifted interest from the systematic learning of lexis and grammar to learning language for communication. However, the availability of new, improved materials based on a new theory of language which raised awareness of the importance of accuracy led to a revived interest in language. A change came at a time when corpus research had progressed to the extent that it could be used as a valuable source of information about language features that need to be taught (Kennedy 2009). This is where and when our Ugly Duckling was born but is finding some hard time in turning into a beautiful swan. Or rather, we have not yet been able to present it as such to a large enough audience.

Both partners in this joint venture fit into Scrivener's idea of interventionist teaching which requires teachers to be more assertive, their interventions to be more muscular, to push and nudge students to achieve more. Both tools, each in its own way, are a response and a contribution to the "demand-high teaching" where, keeping the advantages of the communicative approach, we expect more, ask for more and introduce strategies, techniques and interventions that achieve more (Scrivener 2013).

Therefore, this paper will provide an analysis of how the proposed joint venture worked for a group of fifty 1st year students of the Faculty of Economics, University of Split. The students were given a task consisting of several parts: (i) a translation of five Croatian sentences into English, each sentence containing a language element that commonly presents a problem to the native speakers of Croatian; (ii) an analysis of three short sets of concordance lines tackling three language problem areas that might be of potential use in improving the translations; (iii) using the knowledge acquired in improving their translations if they thought they could or wanted to do it. The feedback on students' attitudes was collected by a questionnaire which consisted of 18 questions/statements broken down into four sections.

References

Cook, G.2012. *Translation in Language Teaching: An Argument for Reassessment.* Oxford: Oxford University Press.

Isaacs, T. 2009. "Integrating form and meaning in L2 pronunciation instruction". *TESL Canada* Journal, 27: 1-12. Available at: http://www.teslcanadajournal.ca/index.php/tesl/article/viewFile/1034/853

Kelly, L.G. 1969. *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C.-1969.* Rowley, MA: Newbury House.

Scrivener, J. 2013. "A proposal: for active interventionist teaching". In T. Pattison (ed*) IATEFL 2012 Glasgow Conference Selections*. Kent: University of Kent

Widdowson, H. G. 2003. *Defining issues in English Language teaching*. Oxford: Oxford University Press.

A corpus-driven study of the learning of disciplinary genres

Ryan T. Miller
Carnegie Mellon University in Qatar
rtmiller@qatar.cmu.edu

Silvia Pessoa
Carnegie Mellon University in Qatar
spessoa@qatar.cmu.edu

1 Introduction

The study of academic and disciplinary writing has benefitted greatly from recent development of corpora such as the British Academic Written English corpus (BAWE; Nesi, 2011) in the UK and the Michigan Corpus of Upper-level Student Papers (MICUSP; Römer & Wulff, 2010) and Portland State University Corpus of Student Academic Writing (PSU C-SAW; Conrad & Albers, 2008) in the US. Studies using these corpora have revealed linguistic and rhetorical differences between writing in different disciplines. However, much of this research has focused on broad disciplinary boundaries, rather than the specific genres that novices in those disciplines need to learn to produce. In addition, corpora of student academic writing generally include only assignments that received high scores, masking the complete range of abilities among learners.

On the other hand, research within the framework of Systemic Functional Linguistics has yielded rich descriptions of written genres within disciplines such as history (Coffin, 1997, 2004, 2006; Veel & Coffin, 1996) and biology (Humphrey, 2013; Humphrey & Hao, 2013). For example, Coffin (2004) identified within school history writing three genre families consisting of 11 distinct genres, each with its own social purposes and linguistic features. However, genre research has largely relied on either manual analysis of a small number of texts, or on analysis of genres based solely on instructor descriptions, without including analysis of actual writing.

The study of not only academic writing, but disciplinary writing is of importance recently due to a rapid increase in participation in English-medium higher education around the world. Disciplinary writing is challenging to learners in these contexts because it involves the dual purposes of understanding a discipline's content as well as the expectations and demands of the various genres and audiences within the discipline. In addition, although many universities offer general instruction in academic writing, few offer instruction in specific disciplinary genres. Thus, there is a critical need for better understanding of discipline-specific genres and how they are learned.

2 The present study

In the present study, we use corpus linguistic methods to investigate the recurring linguistic choices that instantiate specific written genres, with the goal of identifying the linguistic and rhetorical features of these genres and assessing students' ability to include these features in their own writing. In the present study, we focus on the discipline of information systems (IS), the study and use of computers and information technology tools as instruments to generate, process, and distribute information. Parallel to the technical aspects of this discipline, IS professionals are also required to perform documentation and analysis of IS solutions. These tasks are realized via the written genres of project reports and case analyses, genres that are often assigned in IS courses. In the present study, we investigate the linguistic and rhetorical features of these genres, and how undergraduate students at an English-medium university in the Middle East learn to produce these genres.

3 Data and methods

The data for the current study are drawn from a four-year longitudinal study of academic and disciplinary literacy development at a branch campus of an American university in the Middle East. For the current study, sample project reports and case analyses were collected, as well as student-written texts from across the entire four-year information systems curriculum. We use the corpus-based text analysis tool DocuScope (Ishizaki & Kaufer, 2012) to conduct rhetorical analyses of the focal genres, and UAM CorpusTool (O'Donnell, 2008) to conduct linguistic analyses based on Systemic Functional Linguistics (Halliday & Matthiessen, 2004). These tools and methods are used in order to reveal how language is used to realize these genres within a social context.

First, analyses of professional project reports and case analyses are conducted in order to identify features of the focal genres. The analysis is carried out through statistical comparison of the sample project reports and case analyses with a larger reference corpus in order to identify the salient linguistic and rhetorical features of the genres, as well as to identify the target levels of these features. Following the identification of salient features, we conduct additional analyses of student writing in order to evaluate students' acquisition of the genre features over time.

4 Findings

Findings reveal that non-native English-speaking students are able to acquire a number of the linguistic and rhetorical features of disciplinary genres in the field of information systems. However, results also suggest that acquisition of these features could be aided through explicit instruction. Pedagogical applications of the findings for the explicit instruction of the salient features of IS genres are discussed.

References

Coffin, C. (1997). Constructing and giving value to the past: An investigation into secondary school history. In F. Christie & J. Martin (Eds.), Genre and institutions (pp. 196-230). London: Continuum.

Coffin, C. (2004). Learning to write history. *Written Communication, 21*, 3, 261-289.

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

Conrad, S. & Albers, S. (2008, October). *A new corpus of student academic writing*. Paper presented at the American Association of Corpus Linguistics, Provo, UT.

Halliday, M.A.K. & Matthiessen, C.M.I.M. (2004). *An introduction to functional grammar.* (3rd edition). London: Hodder Arnold.

Humphrey, S. (2013). Designing a reading pedagogy for undergraduate biology students. *Linguistics and the Human Sciences, 7*, 55–76.

Humphrey, S., & Hao, J. (2013). Deconstructing written genres in Undergraduate Biology. *Linguistics and the Human Sciences*, 7, 29–53.

Ishizaki, S., & Kaufer, D.S. (2012). Computer-Aided Rhetorical Analysis. In P. McCarthy & C. Boonithum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 276–296). Hershey, PA: Information Science Reference.

Kaufer, D.S., Geisler, C., Ishizaki, S., & Vlachos, P. (2005). Computer-support for genre analysis and discovery. In Y. Cai (Ed.), *Ambient intelligence for scientific discovery: Foundations, theories, and systems* (pp. 129–151). New York: Springer.

Nesi, H. (2011). BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 213-228). London: Continuum.

O’Donnell, M. (2008, April). *The UAM CorpusTool: Software for corpus annotation and exploration.* Paper presented at the XXVI Congreso de AESLA, Almeria, Spain.

Römer, U. & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research,* 2, 99-127.

Veel, R., & Coffin, C. (1996). Learning to think like an historian: The language of secondary school history. In R. Hasan and G. Williams (Eds.), *Literacy in society* (pp. 191–231). London: Longman.

Self-repetitions in learners' spoken language -- a corpus-based study

Marek Molenda
University of Lodz
marek.j.molenda@gmail.com

Piotr Pęzik
University of Lodz
piotr.pezik@gmail.com

In our presentation we would like to focus on self-repetitions in spoken English, as used by Polish EFL students. While there already exist many definitions of the phenomenon in question, most of them are based on the characteristics of native speakers' oral performance in a given language (e.g. Henry, 2002; Kjellmer, 2008). On the other hand, the way in which learners use self-repetitions might provide an insight into the nature of oral fluency and its constituents, as well as students' progress towards formulaicity (McCarthy, 2005). Thus, the aim of the presentation is to compare and contrast certain aspects of native speakers' usage of self-repetitions (as descried by Blankmenship & Kay, 1964; Tannen, 1989; Candéa, 2000, Henry, 2002; Kjellmer, 2008) with the data retrieved from the spoken component of the PELCRA Learner English Corpus[9] (Pęzik, 2012).

Following Henry's classification (2002) we decided to separate the instances of *performance repetitions*[10] and *competence repetitions*. While the latter category refers primarily to repeating content in order to achieve grammatical accuracy (*we **had had** a dog*), the former type is connected with performance effects, e.g. (*acoustic **was was** terrible for me*). In the course of the research it was decided to focus primarily on performance repetitions, which -- though criticized because of their "disorderly appearance" (Kjellmer, 2008, p.38), or even labeled as *dysfluencies* (e.g. Shriberg 1994) -- seem to constitute an important part of spoken language and reflect the true qualities of oral production, that is "the momentary, the individual, the performative, the disorderly" (Hill, 1986).

In terms of structure description, it was decided to utilize Candéa's classification (2000), where each repetition is divided into two major parts, namely "le répétable" (*the repeatable*) and "le répété" (*the repeated*). These two elements can be starting points for two different types of analysis. Firstly, by focusing on *the repeatable*, one may attempt to discover which elements of spoken language are repeated most frequently, and in which context(s) such a phenomenon is most likely to occur. In the case of EFL learners, this type of research could provide certain answers concerning differences between native and non-native speakers. Some possible research directions were suggested by Henry (2002) who was interested in the type of words repeated (*function* vs *content* words) as well as their placement (at the boundaries between syntactic structures or within them).

Moreover, the instances of repetitions might be classified according to their functions. Although the "variety of functions served by repetition is impressive" (Kjellmer, 2008, p.45) and "It would be hubris (and hopeless) to attempt to illustrate every form and function of repetition" (Tannen, 1989, p.85), it is still true that there are some basic criteria that might be applied in order to understand the purpose and/or the reason for which this device is used in speaking. Our classification, based on the works of Denke (2005) and Henry (2002) includes the following labels:

- Hesitational repetitions
- Rhetorical repetitions
- Repairs

On the other hand, focusing on *the repeated* provides information concerning the way in which a given repetition is realized. In this case, the factors described by Henry (2002), such as *the length of the repeated phrase*, *the number of repetitions*, or *the succession of the repeated elements* (immediate vs interrupted) can oftentimes help decide whether or not the use of certain phrases is natural in the context given. For instance, the frequency of occurrence of *so* repeated three times:

Okay so
so
so you are also into politics...

is significantly higher in PLEC corpus than in spoken components of the BNC and COCA.

The main finding of our research is the fact that the majority of repetitions used by the students are function words and hesitation markers. However, both kinds of repetitions might provide an insight into certain lexical issues, such as compositional efforts to use phrases which are idiomatic for native speakers:

in
in
in these pictures,

boundaries between formulaic phrases:

and then get back
aah get back home,

[9] PLEC is available at: pelcra.pl/plec

[10] Henry uses the term *repetition* to refer to *self-repetitions* only, as opposed to Kjellmer, who introduced a division between *self-* and *allo- repetitions*. For the sake of brevity, we adopted Henry's approach in this abstract.

or struggles to find correct lexical items:

play against players
from
from
aah from outs
from other
from other countries right ?

Finally, we decided to group the repetitions according to learners' level of linguistic proficiency, and compare their usage. Such an approach makes it possible for us to focus on the use of these devices across different groups of learners. We hope that the results contribute to the ongoing discussion on the use of repetitions in learners' spoken language.

References

Blankenship J., Kay C. 1964. "Hesitation phenomena in English speech : a study in distribution". *Word*. 20: 360-372.

Candéa M. 2000 *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Étude sur un corpus de récits en classe de français*. Unpublished PhD thesis, Université Paris III (Sorbonne Nouvelle).

Denke, A. 2005. *Nativelike performance. Pragmatic markers, repair and repetition in native and non-native English speech.* Stockholm: Department of English, Stockholm University

Hill, J. 1986. "The refiguration of the anthropology of language". *Cultural Anthropology*. 1: 89-102.

Henry, S. 2002 "Etude des répétitions en français parlé spontané pour les technologies de la parole". Paper presented at *Récital*, Nancy: 24-27 June.

Kjellmer, G. 2008 "Self-repetition in spoken English discourse" In T. Nevalainen, I. Taavitsainen, P. Pahta and M. Korhonen (eds.) *The dynamics of linguistic variation: Corpus evidence on English past and present.* Amsterdam: John Benjamins

McCarthy, M. 2005 "Fluency and confluence." In M. McCarthy (ed.), *Explorations in corpus linguistics*. New York: Cambridge University Press.

Pęzik, P. 2012 "Towards the PELCRA Learner English Corpus." In P. Pęzik(Ed.) *Corpus data across languages and disciplines, Lodz studies in language.* Vol. 28. Frankfurt am Main: Peter Lang.

Shriberg, E.E. 1994. *Preliminaries to a theory of speech disfluencies* Unpublished PhD thesis, Department of Psychology, University of California, Berkeley.

Tannen D. 1987. "Repetition in conversation as spontaneous formulaicity". *Text*. 7 (3): 215-243.

Productivity and difficulty of function words

Naoki Nakamata
Kyoto University of Education
nakamata@kyokyo-u.ac.jp

1 Introduction

Each function word shows its own productivity; some words combine with many kinds of verbs and some words combine only with a very limited number of words.

The aim of this paper is to propose a new way of quantifying the productivity of 100 functional words taught in a Basic Japanese Course. Furthermore, it is demonstrated that this productivity has a strong relationship with learning difficulty: both "high-productive" words and "low-productive" words are easy, but "mid-productive" words are difficult.

2 Preceding studies of productivity

The term "productivity" has been discussed in studies of suffixes such as English *-ity* or *-ness* (Taylor 2003). Further, some formulae to quantify the productivity have been proposed. For example, Baayen (1991) proposes a formula, given below, where n_1 means the number of *hapax legomena*, or words that appear only one time in a corpus collocated with the target item, and N means all tokens of the target item.

$$p = \frac{n_1}{\mathrm{N}}$$

However, this formula cannot be applied in the context of the aim of this paper, because some function words show a large number of tokens. To confirm this, see Table 1.

	N	n_1	p
te-iru	985,113	9,547	0.010
te-aru	15,088	662	0.044

Table 1: Application of Baayen's formula

Japanese has two function words representing statives, *te-iru* and *te-aru*. *Te-iru* can follow both intransitive and transitive verbs, while *te-aru* follows only transitive verbs. So logically, the productivity of *te-aru* should be less than that of *te-iru*. But when Baayen's formula is applied, the index of *te-iru* becomes very small because of the large number of tokens. Thus, a new way of quantifying productivity is necessary.

3 Candidates for productive index

To measure productivity beyond large differences between items, six candidate measures are prepared.

(a) Guiraud Index
(b) Standardized Type–Token Ratio
(c) Frequency Ratio of Top 10 Verbs
(d) Revised Perplexity
(e) Gini Coefficient
(f) Entropy

4 Evaluation and decision between indices

Which candidate can measure productivity best? To decide, 100 functional words following verbs were surveyed by searching system *Chuunagon*,[11] which can search the Balanced Corpus of Current Written Japanese with morpheme information. All functional words were calculated for all six indexes. More than 40 million sentences are gathered in total.

After that, the candidates were evaluated from several perspectives. Results are below, and can be summarized as follows. (a) Guiraud Index is considered the best approach, for several reasons: (1) It can distinguish between *te-iru* and *te-aru* with a large difference; (2) it is little effected by token frequency—the aim is to clarify the degree of productivity itself, so it is important to exclude the effect of frequency as much as possible; (3) the results fall within a suitable range (0–30); they also show a nearly normal distribution; and (4) it shows a U-shaped difficulty effect most significantly discussed in section 6 below.

On this basis, a new productivity index is proposed. The formula is quite simple:

$$p = \frac{\text{type}}{\sqrt{\text{token}}}$$

(In this paper, type means the number of kinds of verb that collocate with the target item.)

5 Property of productivity

The results given above are consistent with the preceding insights regarding the descriptive grammar of Japanese. High-productive words contain voice markers and tense markers; since they involve no constraint on verb meanings, they can collocate with any verbs. Low-productive words, in contrast, contain a lot of markers for asking, proposal, and prohibition, as well as honorifics. These all are used more in oral communication than in written words.

6 Productivity and difficulty of acquisition

Productivity has a strong relation to difficulty of acquisition. All 100 functional words considered here were surveyed in the KY corpus, the most famous corpus of spoken learner's Japanese. The corpus contains three mother languages (Chinese, English, Korean) and four levels (Novice, Intermediate, Advanced, Superior)—90 leaners in total.

High-productive words appear in novice learners' utterances and are widely used at the intermediate level. Middle-productive words don't appear until the advanced level. Some low-productive words don't appear at the intermediate level, but these words never appear at the superior level either, and are hardly used by native speakers. And some low-productive words do appear at the intermediate level.

This means that middle-productive words are the most difficult to acquire. Figure 1 is a scattergram of productivity and appearance at the intermediate level. The horizontal axis shows the productivity index and the vertical axis, the ratio of intermediate leaners' usage to superior learners' usage (Superior level = 1). Notice that zero-usage words gather in the middle of the productive area (13–19).

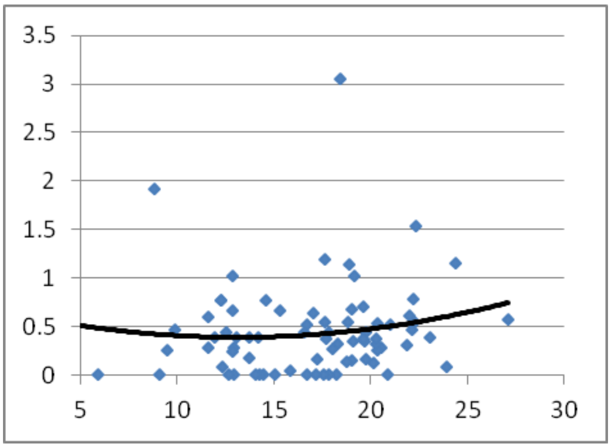


Figure 1: Productivity and usage at the intermediate level

7 Conclusion

In this paper, a new formula to calculate word productivity is presented, and by calculating productivity for 100 functionwords in Japanese, it is demonstrated that productivity has a strong relation to difficulty of acquisition. Table 2 gives a summary.

[11] chunagon.ninjal.ac.jp/

High-productive	Middle-productive	Low-productive
Easy to learn (as grammar)	Difficult to learn (as grammar)	Easy to learn (as lexicon)

Table 2: Productivity and Difficulty

This spectrum matches Langacker (1983)'s claim of continuity of grammatical items and lexical items.

References

Baayen, H. 1991. "Quantum aspects of morphological productivity". In Geert Booij (ed.) *Yearbook of morphology*. Dordrecht/Boston/London: Kluwer.

Langacker, R.W. 2008. *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

Taylor, J.R. 2003. *Cognitive Grammar*. New York: Oxford University Press.

Corpora

Balanced Corpus of Current Written Japanese, National Institute of Japanese Language.

KY Corpus, Kamada, O. and Yamauchi, H.

Problems with learning location prepositions in Czech and English: Prepositions *v* versus *na* contrasted with *in* versus *at*

Renata Novotná
Institute of the Czech National Corpus,
Charles University, Prague
renata.novotna@ff.cuni.cz

1 Introduction

This contribution is based on the data of the parallel corpus Intercorp including Czech and other 32 languages (about 867 mil. words). The Czech-English parallel corpus which is part of the Intercorp corpus has about 55 mil. words.

2 Frequency of the prepositions

The prepositions chosen for this study are very frequent in both languages and are often source of mistakes, especially by foreign users. Prepositions *v* and *na* are two most frequent prepositions in Czech. According to Frequency Dictionary of Czech (Čermák and Křen eds 2004) preposition *v* is the most frequent and preposition *na* the second frequent preposition. Prepositions *in* and *at* in English are also very frequent – according to frequency dictionary Word Frequencies in Written and Spoken English Based on the British National Corpus (2001) the preposition *in* is on the fifth place and preposition *at* on the twentieth place.

3 Functions of prepositions

The location (or position) meaning of the prepositions studied is the part of the adverbial function of prepositions. The prepositions can have three functions according to their position in the sentence (Čermák 1996):

a) adverbal – relation between the verb and the noun (*vzpomínat na prázdniny* – to remember the holidays)

b) adnominal – relation between two nominal expressions (*prázdniny u moře* – holidays by the sea)

c) adverbial – relation to the whole proposition (e.g. *v létě* – in the summer).

This contribution will be concentrated on the adverbial function of the Czech and English prepositions studied.

4 Semantic groups of prepositions

According to Čermák (1996) Czech one-word prepositions can be divided into seven semantic groups:

1	Identification	5,6 %
2	Classification	4,8 %
3	Qualification	4,0 %
4	Determination	26,4 %
5	Causality	8,0 %
6	Localisation	37,2 %
7	Temporalization	14,0 %

Table 1: Semantic groups of Prepositions

As the table 1 shows the localisation group has the highest percentage.

5 Meaning of prepositions studied

The Czech prepositions *v* and *na* can take both accusative and loccative cases. In the location sense the preposition *v* is used only in loccative case (where?) while the preposition *na* in loccative case (where?) and accusative case (which direction?). As far as the English prepositions *in* and *at* are concerned, according to English-Czech Explanatory Dictionary (1998) the first meaning of these prepositions is location: *in* – 1. something that is **in** something else is enclosed by it or surrounded by it, 2. if something is **in** a place, it is there; *at* – 1. you use **at** to say where something happens or is situated. In both cases the Czech equivalents are *v* and *na*. This study will be therefore concentrated on the first meaning of the preposition ***in***, e. g. *in the box – **v** krabici* and the differences of the second meaning, such as *in the room – **v** pokoji, **in** the garden – **na** zahradě,* then on the location meaning of preposition ***at***, such as *at the hospital – **v** nemocnici* vs *at the university – **na** univerzitě*.

6 Examples of excercises

In conclusion, at least two examples of working with the corpus Intercorp will be given. The best method is the data-driven learning. The first task for the students is to search for the collocations *na zahradě* (75 occurences) and *v zahradě* (115 occurences) and study the difference: *na zahradě* – referring to gardening, e. g. *proč si chirurg bere na práci na zahradě rukavice - why a surgeon takes gloves to work in the garden* x *v zahradě* – referring to plants and trees: počítal *v zahradě stonky cibule - counted the onion plants in the garden*. The English equivalent in both cases is *in the garden*. If we search *at the garden* we will find out that the preposision *at* is the valency exponent of previous verb, such as *to look at the garden*. The task of the students is to study if there is the same rule for *na poli – v poli* (in the fields), cf. *celý den pracují na poli – they worked the whole day in the fields* x *stromy roztroušené v poli - some trees here and there in the fields*. The collocation *v poli* is also used in military sense: *vojáci v poli – soldiers in the field*. The other example concerning the problems with learning location prepositions the collocations *ve škole* and *na škole* vs *at (the) school* and *in (the) school* will be studied. The students are given concordances of *ve škole* (243) and *na škole* (47). Their task is to discover the difference between Czech *ve škole* (i. e. studying there) and *na škole* (i. e. teaching there). The English equivalents for *na škole* are *in the school* and *at the school* for ve škole are *at school* and *in school*.

References

Čermák, F. 1996. Systém, funkce, forma a sémantika českých předložek. *Slovo a Slovesnost* 57, 30-46.

Čermák, F. and Křen, M. (eds) 2004. *Frekvenční slovník češtiny*. Praha: NLN.

Čermák, F. and Holub, J. 2005. *Syntagmatika a paradigmatika českého slova I. Valence a kolokabilita*. Praha:Nakladatelství Karolinum.

Čermák, F. and Rosen, A. 2012. The Case of InterCorp, a multilingual parallel corpus. *IJCL* 17:3, 411-427.

Český národní korpus - InterCorp. Ústav Českého národního korpusu FF UK, Praha. Available online at WWW: <http://www.korpus.cz>.

Klégr, A. and Malá, M. and Šaldová, P. 2012. *Anglické ekvivalenty nejfrekventovanějších českých předložek*. Praha: Nakladatelství Karolinum.

Leech, G and Rayson, P. and Wilson, A. 2001. *Word Frequencies in Written and Spoken English Based on the British National Corpus*. London: Longman..

Anglicko-český výkladový slovník 1998. Praha: NLN (Chief editor of the English part J. Sinclair, Czech supervision F. Čermák).

Cobuild Collins 1991. *English Guides 1:Prepositions, Helping learners with real English*. Glasgow: Caledonian International Book Manuf.

Multi-word formulaic phrasal construction bundles: phraseology, terminology, corpora, and EAP pedagogy

David Oakey
Iowa State University
djoakey@iastate.edu

This paper presents a comparison of recently published lists of phraseological items intended for use in English for Academic Purposes (EAP) pedagogy. The importance of phraseological items in first and second language acquisition and use is today well recognized, and EAP practitioners are aware of the need for their students to acquire these linguistic forms and their meanings, and identify the appropriate registers in which to use them.

However, phraseology is an area with a wide scope and long history. The plethora of phraseology terminology has been troublesome since well before the application of computers to applied linguistics. The phraseology terminology problem is that the same term is used by different people for different things; or that the same term is used by one person for too wide a range of things; or that different terms are used by one person or by different people for the same thing. All terms come with their own epistemological "baggage", carrying along with them echoes of the contexts in which they have been coined, used and/or misused by different scholars over the years.

Research into identifying pedagogically useful phraseological items has been informed by a diverse range of theoretical perspectives. Lexicographers refer to particular word combinations as 'frozen metaphors', 'frozen phrases' or 'fossilized forms', implying that meanings and forms can change over time or become static. Cognitive linguists employ a building metaphor to hint at how the mind acquires and processes such combinations, as in 'preassembled speech', 'pre-formulated units', or 'ready-made expressions'. Sociolinguistic perspectives, which highlight the role of word combinations in language use, duly focus on the repeated, routine nature of the social situations in which they occur, as in 'formulaic speech' and 'conventionalized forms' (Wray 2002: 9). Common lexical, syntactic, semantic, pragmatic, and methodological principles for selecting such linguistic units for EAP are not yet established.

Recently the picture has been made clearer for theorists by large scale studies of phraseology which have produced lists of phraseological items backed by corpus evidence. The picture for those teaching EAP, on the other hand, has been clouded by these studies since they have either used additional terms or appropriated existing ones: '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). The shared pedagogic premise for these lists, in keeping with the tradition of lists of single words (West 1953; Coxhead 2000; Gardner and Davis 2013) is that since these items are found to occur in academic registers they should accordingly be taught to EAP learners.

Confronted with so many lists and so many different names for the items on them, 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. While some items appear on more than one list, the fact that different names for the items are used - an exacerbation of the traditional phraseology terminology problem - makes it unclear whether these items are similar enough to be considered as the same linguistic feature.

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 combines lexical, syntactic, semantic, pragmatic, and methodological criteria from traditional "Eastern European" lexicography for identifying 'restricted collocations' (Aisenstadt 1981; Howarth 1996), from "Empirical Firthian" lexicology for identifiying 'extended lexical units' (Stubbs 2001), and from "Usage-Based" cognitive linguistics for defining 'phraseologisms' (Gries 2008). For example, the semantic criterion for a restricted collocation is that it must be partially semantically transparent, in which one element of the combination has a literal, unidiomatic meaning; the same criterion for an extended lexical item is that it has an observed semantic preference, such as a particular lexical set, semantically related word-form or lemma; while for a phraseologism the definition involve an item's semantic non-compositionality or non-predictability. Particularly important for EAP learners, whose linguistic choices must be appropriate to the register in which they will use English, is the pragmatic criterion: for a restricted collocation this means it must be institutionalized (somehow distinctive and memorized); for an extended lexical unit the pragmatic criterion involves the item's discourse function and its distribution in text types; the pragmatic criteron for identifying phraseologisms is less explicit.

The results of this comparison reveal common ground shared by the items in these lists, and the paper explores 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.

References

Ackermann, K. and Chen, Y-H. 2011. "The academic collocation list". *Pearson Education*. www.pearsonpte.com/RESEARCH/Pages/CollocationList.aspx

Aisenstadt, E. 1981. "Restricted collocations in English lexicology and lexicography". *ITL (Instituut voor Toegepaste Linguistiek) Review of Applied Linguistics*, 53: 53-61.

Biber, D., Conrad, S., and Cortes, V. 2004. "If you look at ...: Lexical Bundles in University Teaching and Textbooks". *Applied Linguistics*, 25(3): 371-405.

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

Coxhead, A. 2000. "A new academic word list". *TESOL Quarterly*, 34 (2): 213-238.

Durrant, P. 2009. "Investigating the viability of a collocation list for students of English for academic purposes". *English for Specific Purposes*, 28: 157-169.

Gardner, D., and Davies, M. 2013. "A new Academic Vocabulary List". Applied Linguistics, advanced access 1-24.

Gries, S. T. 2008. "Phraseology and linguistic theory". In S. Granger & F. Meunier (Eds.), *Phraseology: an interdisciplinary perspective.* (pp. 3-25). Amsterdam: John Benjamins.

Howarth, P. A. 1996. *Phraseology in English academic writing*. Tübingen: Max Niemeyer.

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

Liu, D. 2012. "The most frequently-used multi-word constructions in academic written English: a multi-corpus study". *English for Specific Purposes* 31 (1): 25- 35.

Martinez, R., & Schmitt, N. 2012. "A phrasal expressions list". *Applied Linguistics*, 33 (3): 299-320.

Simpson-Vlach, R., & Ellis, N. C. 2010. "An academic formulas list: new methods in phraseology research". *Applied Linguistics*, 31 (4): 487-512.

Stubbs, M. 2001. *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

West, M. 1953. *A General Service List of English Words*. New York: Longmans, Green and Co.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Hypal: A user-friendly tool for automatic parallel text alignment and error tagging.

Adam Obrusnik
Masaryk University
adam.obrusnik@gmail.com

1 Introduction

Parallel corpora are of great importance in the field of applied linguistics, especially in translation and language teaching. For a parallel corpus to be reliable, it is necessary to align the body of texts as accurately as possible. Apart from parallel corpora, error-tagged corpora are sometimes used in language teaching as well.

This contribution presents a user-friendly tool called *Hypal* which has the capabilities of automatic text alignment, manual text alignment (if necessary) and manual error-tagging.

This contribution includes a brief description of an automatic alignment algorithm developed by the author in his thesis (Obrusník, 2013) as well as a demonstration of the web-based graphical user interface and the possibilities of using *Hypal* together with other tools. Projects currently using *Hypal* will also be briefly mentioned.

2 Automatic text alignment

The algorithms that have been used for parallel text alignment can be generally divided into two categories; statistical models (e.g. Och and Ney 2003) and anchor-based models (Hofland and Johannson 1998).

The automatic alignment procedure used in *Hypal* incorporates features from both approaches. It is in principle similar to the algorithm described by Varga et al. (2005) but since it has been developed independently, it differs in several features, mainly the metric used for statistical scoring and the matrix formalism it incorporates.

The matrix formalism that has been worked up recently to better formulate the ideas presented in the author’s thesis is relatively flexible and, most importantly, it allows a very intuitive incorporation of constraints and restrictions.

As an example, consider M sentences in language A and N sentences in language B. Then an M × N matrix can be constructed with each element expressing the similarity of a particular A-B sentence pair. To calculate the score of a particular alignment, the matrix is multiplied by a projection matrix, if applicable, (corresponding to sentence merging) and an orthogonal transformation matrix (corresponding to interchanging of sentences). Furthermore, from applying certain reasonable constraints, e.g. merging 3 sentences at maximum, it follows that the M × N matrix does not have to be populated entirely and populating only several diagonals is sufficient.

3 Parallel alignment interface

The parallel alignment interface allows the user to save parallel texts to a database and align them automatically or semi-automatically.

The alignment procedure consists of several steps. Firstly, in the case of longer parallel texts, it is advisable to verify the paragraph-level alignment proposed by the program. The program then performs POS tagging of both the language versions (if a tagger and disambiguator are available) and the automatic alignment.

The user can subsequently view the automatic alignment and make changes to it. Aligned texts can be exported either as Sketch Engine-compatible *.vert files or as *.tmx files for translation memories.

As mentioned above, it is strongly preferable to POS-tag the texts prior to the alignment. *Hypal* is currently capable of working with all languages supported by TreeTagger (Schmidt 1994), e.g. English, French, German, Italian. In addition, a POS-tagger for Czech and Slovak by Spoustová et al. has been implemented.

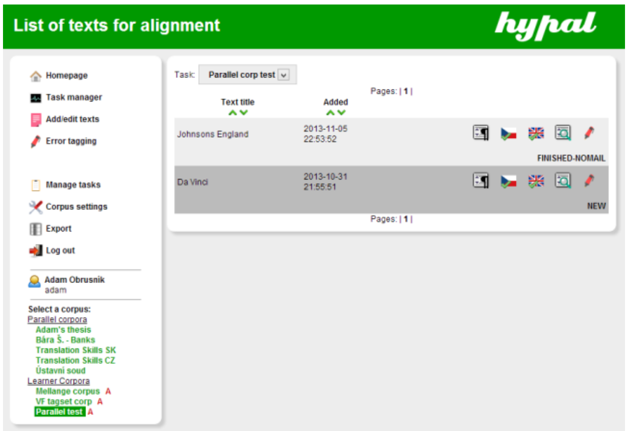
4 Error tagging

The second feature of *Hypal*, which can be used independently of the parallel corpus interface, is user-friendly error-tagging of texts. For this purpose, *Hypal* also includes a student interface, through which students can submit their assignments and view the error-tagged versions once the teacher has reviewed the texts.

The error tagging interface then allows teachers to tag the errors in the submissions, either using their own error tag set or a pre-defined one. The consistency of the error tag set used is ensured by a system of user rights and privileges. The error tagged texts can also be exported, for instance to a Sketch Engine-compatible format.

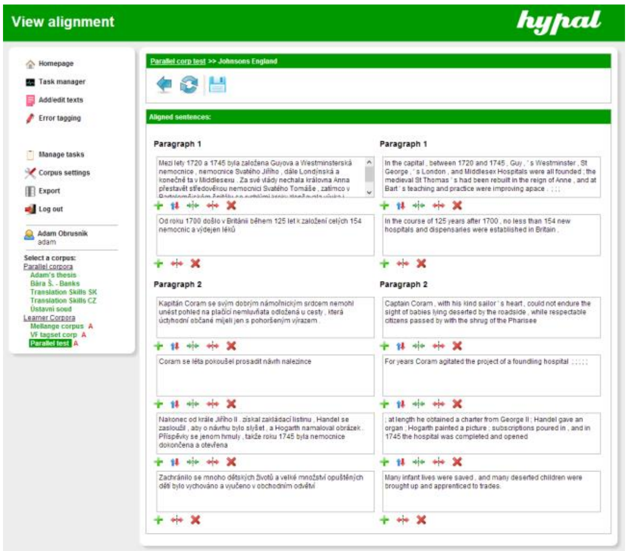
In future, it should also be possible for the students and teachers to view the error statistics obtained from larger bodies of texts. This would allow the students to focus on their own problematic areas and the teachers could easily see what types of mistakes are most frequent among their students.

5 Screenshots



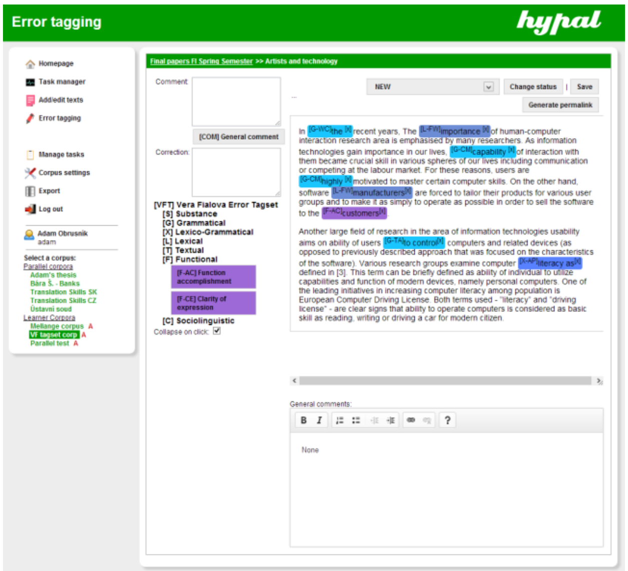
Text title	Added	
Johnsons England	2013-11-05 22:53:52	FINISHED-NOMAIL
Da Vinci	2013-10-31 21:55:51	NEW

Figure 1: Text overview – parallel corpus



Paragraph 1	Paragraph 1
Mezi lety 1720 a 1745 byla založena Guyova a Westminsterská nemocnice , nemocnice Svatého Jiřího , dále Londýnská a konečně ta v Middlesexu . Za své vlády nechala královna Anna přestavět středověkou nemocnici Svatého Tomáše , zatímco v	In the capital , between 1720 and 1745 , Guy , ' s Westminster , St George , ' s London , and Middlesex Hospitals were all founded ; the medieval St Thomas ' s had been rebuilt in the reign of Anne , and at Bart ' s teaching and practice were improving apace . ; ; ;
Od roku 1700 došlo v Británii během 125 let k založení celých 154 nemocnic a výdejen léků	In the course of 125 years after 1700 , no less than 154 new hospitals and dispensaries were established in Britain .
Paragraph 2	**Paragraph 2**
Kapitán Coram se svým dobrým námořnickým srdcem nemohl unést pohled na plačící nemluvňata odložená u cesty , která úctyhodní občané míjeli jen s pohoršeným výrazem .	Captain Coram , with his kind sailor ' s heart , could not endure the sight of babies lying deserted by the roadside , while respectable citizens passed by with the shrug of the Pharisee
Coram se léta pokoušel prosadit návrh nalezince	For years Coram agitated the project of a foundling hospital ; ; ; ; ;
Nakonec od krále Jiřího II . získal zakládací listinu , Handel se zasloužil , aby o návrhu bylo slyšet , a Hogarth namaloval obrázek . Příspěvky se jenom hrnuly , takže roku 1745 byla nemocnice dokončena a otevřena	; at length he obtained a charter from George II ; Handel gave an organ ; Hogarth painted a picture ; subscriptions poured in , and in 1745 the hospital was completed and opened
Zachránilo se mnoho dětských životů a velké množství opuštěných dětí bylo vychováno a vyučeno v obchodním odvětví	Many infant lives were saved , and many deserted children were brought up and apprenticed to trades.

Figure 2: Manual editing of the automatic alignment



[VFT] Vera Fialova Error Tagset
- [S] Substance
- [G] Grammatical
- [X] Lexico-Grammatical
- [L] Lexical
- [T] Textual
- [F] Functional
  - [F-AC] Function accomplishment
  - [F-CE] Clarity of expression
- [C] Sociolinguistic

Collapse on click: ☑

In [G-WC]the [x] recent years. The [L-FW]importance [x] of human-computer interaction research area is emphasised by many researchers. As information technologies gain importance in our lives, [G-CM]capability [x] of interaction with them became crucial skill in various spheres of our lives including communication or competing at the labour market. For these reasons, users are [G-CM]highly [x] motivated to master certain computer skills. On the other hand, software [L-FW]manufacturers[x] are forced to tailor their products for various user groups and to make it as simply to operate as possible in order to sell the software to the [F-AC]customers[x].

Another large field of research in the area of information technologies usability aims on ability of users [G-TA]to control[x] computers and related devices (as opposed to previously described approach that was focused on the characteristics of the software). Various research groups examine computer [X-AP]literacy as[x] defined in [3]. This term can be briefly defined as ability of individual to utilize capabilities and function of modern devices, namely personal computers. One of the leading initiatives in increasing computer literacy among population is European Computer Driving License. Both terms used - "literacy" and "driving license" - are clear signs that ability to operate computers is considered as basic skill as reading, writing or driving a car for modern citizen.

Figure 3: Error tagging interface (monolingual)

6 Conclusion

This contribution introduces *Hypal*, a graphical interface primarily designated for automatic and semi-automatic alignment of parallel texts. The web-based software also has relatively advanced error-tagging functionality.

References

Hofland, K., Johansson, S. (1998). *The Translation Corpus Aligner: A program for automatic alignment of parallel texts*. In Corpora in Cross-linguistic Research: Theory and Method, and Case Studies, 87-100.

Obrusnik, A. (2012). *A hybrid approach to parallel text alignment.* Masaryk University.

Och, F.J., Ney, H. (2003). *A systematic comparison of various statistical alignment models*. Comput. Linguist., 29 (1), 19-51.

Schmidt, H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of International Conference on New Methods in Language Processing.

Spoustová, J., Hajič, J., Raab, J., Spousta, M. (2008). *COMPOST - Common POS Tagger*. Retrieved October 30, 2012 from http://ufal.mff.cuni.cz/compost.

Varga, D., Németh, L., Halácsy, P., Kornai, M., Trón, V., Nagy, V. (2005). *Parallel corpora for medium density languages* In Proceedings of the RANLP 2005, pages 590-596.

Corpus pedagogy: Analysing EFL business student attitudes towards corpora

John O'Donoghue
Technical University of Applied Sciences Wildau
john.odonoghue@th-wildau.de

Chae Kwan Jung
Korea Institute for Curriculum and Evaluation
ckjung@kice.re.kr

1 Introduction

While corpora have assumed a significant role in shaping both dictionaries and materials design in recent decades, the part which corpora may play in the classroom has seemed subject to controversial debate. Many studies have reported enthusiastically on the use students make of corpus (Thurston and Candlin 1998; Bernardini 2000; Kennedy and Miceli 2001; Sripicharn 2004; Gavioli 2005; Chambers and O'Sullivan 2006; Lee and Swales 2006; Charles 2011). Many of these studies have chosen linguistics-oriented and/or post-graduate students as participants, the subjects in this study, however, were undergraduates studying business (see Boulton 2010 for similar participants). This paper reports on the attitudes of these EFL students towards the use of corpus. The question, therefore, was whether a group comprising only undergraduate students with little linguistics orientation would benefit from corpus work as positively as the subjects in the studies mentioned above.

2 Methods

In June 2012, forty students at a German technical university were introduced to a large reference corpus (British National Corpus, or BNC, accessed using bncweb.lancs.ac.uk) and shown basic techniques for exploring the resource. The subjects in this study were all undergraduates in a Business Administration programme. Their English proficiency levels ranged from B1 to C1 according to the Common European Framework of Reference for Languages (CEFR). A questionnaire survey was conducted with students to ascertain their experience of corpus. In order to clarify some of the issues arising from the results of the questionnaire, the decision was made to take a qualitative approach using a semi-structured interview. Due to time constraints and the in-depth nature of the interviews, it was judged that six students would provide reasonable diversity of experience to yield relevant feedback on corpus learning. By using the NVivo

10, the interview transcripts were scrutinised to identify the main issues experienced by the interviewees.

3 Results and analysis

Table 1 represents the answers by students to key questions.

Category	Useful	Not useful	Mean*	S.D.
Purpose	87.2%	12.8%	4.56	1.12
Usage of vocabulary	79.5%	20.5%	4.20	1.16
Usage of phrases	79.5%	17.9%	4.60	1.08
Usage of grammar	61.5%	33.3%	3.83	1.19
Reading skill	28.2%	64.1%	3.08	1.06
Writing skill	46.2%	48.7%	3.54	1.30
Future writing	69.2%	28.2%	3.81	1.29
Useful resource	59.0%	41.0%	3.79	1.30
Helpful writing	74.4%	20.5%	4.34	1.34
Helpful reading	15.4%	79.5%	2.32	1.18
Improved understanding	59.0%	38.5%	3.57	1.18
Cut-off sentences	41.0%	35.9%	3.63	1.19

Table 1: Using the British National Corpus (O'Donoghue and Jung 2013:58)

Two of the statements that the students reacted to most positively were that they understood the purpose of using the BNC (87.2%) and that they found the BNC helpful for learning the usage of phrases (79.5%). This response would correlate with the phraseological approach to language that corpus promotes, "the way in which meaning is sometimes associated with a whole phrase rather than a single word" (Sinclair 2003:10). Moreover, the students valued the BNC in terms of learning the meaning of vocabulary (79.5%). These findings confirm those of Yoon and Hirvela (2004), whose students responded that corpus use was most helpful for learning the usage of vocabulary and phrases.

There was a drop of nearly 20% when it comes to analysing how useful corpora are for learning grammar. The students may associate grammar with tenses in a more structured format and therefore might have ignored the grammatical element in favour of a lexical concentration. Despite the fact that corpus can, of course, be used to aid reading comprehension, as Gavioli (2005) observed, it was valued essentially as a writing tool: 74.4% cited the latter against only 15.4% for the former.

Many of the comments made by students in the interviews corroborate the inductive approach that corpus work encourages. S1 said that she enjoyed the inductive method because "you have to figure it out yourself" and believed this process enhanced memory storage of such items in contrast to the rule-based approach experienced in school. S1 also mentioned that it helped her learn "chunks". The second student used the BNC for her term essay, which provided her with ideas fitting the general context. The BNC was useful in showing her how words work, especially in the financial domain. S3 also used the BNC for her essay "to know in which context I can use the word". The fourth student compared the activity to mind-mapping and found that being able to discover so many different aspects of a lexical item was an "adventure".

Unsurprisingly, the students criticised certain aspects of corpus work. S2 found navigating her way around the BNCweb rather challenging, confusing, and time-consuming (for example, having to enter brackets for certain queries). The fifth student found navigating around the BNC "not that easy to use and handle". He picked upon the term *query* as an example of a word in the BNC interface that many non-native speakers would not necessarily understand. The final student returned to the issue of how time-consuming and "overwhelming" corpus research may appear.

4 Conclusion

In conclusion, some students enjoyed using corpus as they had to work things out for themselves, using the inductive approach. Moreover, they appreciated that concordance lines provided more context than dictionaries. Others experienced problems with the BNC interface, experiencing the process of investigating words as time-consuming, finding some vocabulary difficult, and even feeling overwhelmed by the amount of data. For these reasons many preferred to use dictionaries in their writing. It seems, however, that when time is not a pressing factor, corpus is considered a positive alternative. Corpus may ultimately be viewed as an adventure that leads in many different directions, and this can seem attractive to some students and disorientating to others. The task of the teacher may lie consequently in focusing on those directions which students find immediately relevant and rewarding and minimizing sources of distraction and disorientation.

References

Bernardini, S. 2000. “Systematizing serendipity: Proposals for concordancing large corpora with language learners”. In L. Burnard, and T. McEnery (eds.) *Rethinking pedagogy from a corpus perspective* (pp225-234). Bern: Peter Lang.

Boulton, A. “Data-Driven Learning: Taking the Computer out of the Equation” *Language Learning* 60 (3): 534-572

Chambers, A. and O’Sullivan, Í. 2006. “Learners’ writing skills in French: Corpus consultation and learner evaluation”. *Journal of Second Language Writing* 15 (2006) 49-68.

Charles, M. 2011. “Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes”. In A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds.) *New trends in corpora and language learning*. London: Continuum.

Gavioli, L. 2005. *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Kennedy, C. and Miceli, T. 2001. “An evaluation of intermediate students’ approaches to corpus investigation”. *Language Learning & Technology* 5 (3): 77-90.

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

O’Donoghue, J. and Jung, C.K. 2013. “Corpus pedagogy: Analyzing corpus use in the classroom and EFL business student attitudes towards corpora”. *English Language Teaching* 25(3): 51-74.

Sinclair, J. 2003. *Reading concordances*. Edinburgh: Pearson Longman.

Sripicharn, P. 2004. “Examining native speakers’ and learners’ investigation of the same concordance data and its implications for classroom concordancing with ELF learners”. In G. Aston, S. Bernardini and D. Stewart (eds.) *Corpora and language learners*. Amsterdam: John Benjamins.

Thurstun, J. and Candlin, C. 1998. “Concordancing and the teaching of the vocabulary of academic English”. *English for Specific Purposes* 17(3): 267–280.

Yoon, H. and Hirvela, A. 2004. “ESL student attitudes toward corpus use in L2 writing”. *Journal of Second Language Writing* 13(4): 257–283.

An English collocations e-workbook: design and applications

Adriane Orenha Ottaiano
Universidade Estadual Paulista
“Júlio de Mesquita Filho” (UNESP)
adriane@ibilce.unesp.br

1 Introduction

McIntosh, Francis and Poole (2009: v) point that “language that is collocationally rich is also more precise” and, according to the same researchers “a student who chooses the best collocation will express himself much more clearly and be able to convey not just a general meaning, but something quite precise”. Other studies have also revealed the relevance of collocations in the current sphere of second language learning and teaching (Thomas, forthcoming; Orenha-Ottaiano 2012; Meunier and Granger 2008; Nesselhauf, 2005; Sinclair 2004; Conzett 2000; Lenko-Szymanska 1997; etc.). The claim underlying this paper is that specific teaching material on collocations should be designed, in order to allow teachers to work with the referred phraseologisms in the classroom more effectively and help learners use them more accurately and productively, taking into account the difficulties they have to master native like phraseological units.

Furthermore, and more importantly, this study argues that the selection of these collocations should be geared to targeting learners of a particular L1 background and thus teaching material should be designed with a careful selection of collocations focusing on specific difficulties learners of a particular L1 have (Mackin 1978). Bearing that in mind, this investigation proposes to address collocational aspects extracted from a parallel corpus called *Translation Learner Corpus* made up of C1 and C2 level university students’ translations from Portuguese into English. The original texts that comprise the corpus are newspaper articles taken from well-known Brazilian newspapers and magazines. The typology of the texts is related to current world news such as *Financial crises in Europe*; *Unemployment*; *Elections in the US*; *Bullying*; *Marijuana Legalization* etc.

2 Methodology

WordSmith Tools (Scott 2008) was used to extract the data and help raise the most frequent collocational patterns used by the translation learners in comparison to the original texts, the influence of the mother tongue on their choices, among other aspects. *The Corpus of Contemporary*

American English (Davies 1990-2012) was also employed to check frequency and recurrence of collocational patterns extracted. When the collocations proposed were not acceptable in the native speakers' language, other collocations were discussed as translation options, and then included in some exercises that comprise the e-workbook. Based on the collected data and the analysis of the results, some corpus-based collocational activities have been specifically designed to L2 learners of English whose L1 is Portuguese, taking into account the difficulties the Brazilian university learners had regarding the use of collocations.

As the collocations E-workbook was intended to be designed for Brazilian Portuguese speakers, the exercises were tested and selected during a 180-hour course entitled "Corpus linguistics and Phraseology applied to the pedagogical practice of English teachers from Public Schools", under our supervision. During this course, public school teachers had the chance to learn the theoretical and methodological concepts of Corpus Linguistics and Phraseology.

This experience may be regarded as a great opportunity for them, bearing in mind that research on the referred area has not reached the intended target audience as much as we have expected to. Moreover, besides gaining knowledge of the theoretical issues, teachers were also given the collocational activities built for the proposed E-workbook. Teachers were encouraged to do and evaluate them, so that it could be chosen the ones which are more suitable for their learning and for the teaching of their own students. Students from a BA in English Language and a BA in Translation have also been exploring the exercises and giving us feedback.

3 Final Remarks

It is expected that this study may contribute to a more effective change in the current paradigm, in what concerns the most traditional concepts of ESL teaching and learning. Under a Corpus Linguistics perspective and having fostered awareness of the importance of Phraseology and collocations to ESL teaching and learning, the benefits from this research may reflect on the target audience's environment as the teachers involved will also influence their co-workers as well as their own students, helping them learn the referred lexical patterns more effectively, besides shedding light on new ways of teaching and learning phraseology, especially collocations.

Additionally, students can count on a new electronic material specially designed for Brazilian Portuguese learners of English and, through the explicit learning of collocations and corpus-based strategies, they may be able to increase their proficiency in English and hence achieve native-like naturalness.

Acknowledgements

I gratefully acknowledge the financial support provided by FAPERP (Fundação de Apoio à Pesquisa e Extensão de São José do Rio Preto).

References

Davies, M. 1990-2012. *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available: <http://corpus.byu. edu/coca/.>. Acessed: April 20th, 2013.

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

Mackin, R. 1978. On collocations: words shall be known by the company they keep. In: Strevens, Peter (ed.) *In Honor of A. S. Hornby*. Oxford: Oxford University Press. 149-165.

Mcintosh, C.; Francis, B. and Poole, R. (eds.) 2009, *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford: Oxford University Press.

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

Orenha-Ottaiano, A. 2012. English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Acta Scientiarum*. 34(1): 241-251.

Scott, M. 2008. *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.

Sinclair, J. McH. 2004. *How to use corpora in Language Teaching*. Amsterdam: John Benjamins

Thomas, J. (forthcoming). Stealing a march on collocation. *TALC 10 Proceedings*.

The use of the items *of* and *to*: what they tell us about multi-word-units

Michael Pace-Sigge
University of Eastern Finland
michael.pace-sigge@uef.fi

This paper looks at the usage patterns found for the highly frequent items *of* and *to* in corpora of different sets of semi-prepared speech. The paper disputes claims that these two items can merely be referred to as prepositions or that they carry little meaning in themselves. Instead, looking at historical roots (cf. Borström: 1965, Hook: 1975, O'Dowd: 1998) and earlier corpus linguistic research (Sinclair et al.: [1970] 2004), this paper aims to demonstrate that both of and to should be understood as essential parts of multi-word-units (MWUs). By way of exemplification, the focus is on the transcripts of BBC Reith Lectures, London School of Economics (LSE) public lectures, and public lectures on key works of art at National Museums Liverpool (NML). The material used here provides a general example which shows both features of conversational English and written English.

This research disputes claims that these two items can merely be referred to as prepositions and/or particles which carry little meaning in themselves. Instead, the investigation will demonstrate that both *of* and *to* have clear semantic and pragmatic functions. Several grammar books appear to reflect this through highlighting that *of and to* are far less easy to classify than is generally assumed. Second, these particular words fulfil specific roles within larger lexical items, which reflect both the roots and the communicative functions they fulfil. The unconscious specific usage pattern for the particular sub-set of English language use described here, mirrors findings by Biber (2000) and Hoffmann (2005) and is seen as support for the lexical priming theory (Hoey: 2005). Being highly frequent items, *of* and *to* provide a key argument as to why language should not be seen as single-word units which have autonomous categories and functions: English, characteristically, is dominated by what John Sinclair terms lexical items: multi-word-units. Another argument to see *of and to* as an integral link part of important formulaic clusters found in the English language, and particularly of importance for corpus linguists, is the fact that traditional classifications seem to struggle to describe or even contain example of what these items are and when they are used. Drawing on material first presented by Sinclair (1991), Rice (1999), Biber (2000) and Hoffmann (2005) it will be shown that *of and to* are far less easy to classify than is generally assumed. This paper will show that *of* occurs in a far more stable and fixed mode than *to* (this mirrors earlier findings by Pace-Sigge, 2009). The two words are constituent parts of longer lexical items: this appears to reflect both the roots and the communicative functions they fulfil. It is crucial too, to be aware of the fact (as Sinclair: 1990, and Stubbs: 1996, have pointed out) that certain word forms predominantly appear in one single construction. This needs to be taken into account in our understanding of language. When we look at the distribution of the items, we recognise that these are not only frequent items in themselves they are also part of the most frequent clusters found in any given corpus, as Pace-Sigge (2013: 191) has highlighted with reference to the large CanCode corpus (see O'Keefe et al.: 2007). *Of* appears the more frequent the longer the cluster, while *to* usage decreases in frequency the longer the cluster. The findings presented here will show that, while highly frequent clusters may appear with varying percentages of use in different sub-corpora, a clear pattern of usage and application appears amongst the most frequently occurring clusters. These uses are more formulaic amongst freely spoken texts than amongst prepared spoken texts. Nevertheless, the nesting found for the material investigated presents a fairly uniform picture.

This paper aims to highlight that teaching English as a foreign language, in particular to students whose L1 have no free-standing prepositions or particles, can be improved by focusing on teaching the use of the highly frequent items *of* and *to* and their relevance as building blocks in the English language. One should look, according to Hoey (2005:184), to surround the learner with evidence: "priming is the result of a speaker encountering evidence and generalising from it". One such piece of evidence is not only to highlight the fact that of and to are highly frequent, but to demonstrate in what collocational and colligational forms of nesting these items are mostly found to appear.

References

Biber, Douglas. 2000. Investigating Language Use through Corpus-Based Analyses of Association Patterns. In: Barlow, Michael and Suzanne Kemmer, eds.: *Usage Based Models of Language.* Stanford, California: CSLI Publications, 287-314.

Brorström, Sverker. 1965. *Studies on the use of the preposition* of *in the 15th century correspondence.* Stockholm: Almqvist & Wiksell.

Hoey, Michael. 2005. *Lexical Priming*. A new theory of words and language. London: Routledge.

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions*. London: Routledge.

Hook, J.N. 1975. *History of the English Language*. New

York: The Ronald Press Company.

O'Dowd, Elizabeth M. 1998. *Prepositions and Particles in English. A Discourse-functional Account*. New York/Oxford: Oxford University Press.

O'Keefe, Michael Mc Carthy and Ronald Carter. 2007. From Corpus to Classroom. Cambridge: CUP.

Pace–Sigge, Michael. 2009. Why TO is a weird word. Available online at www.academia.edu/4107981/Why_TO_is_a_weird_word (last accessed 10/11/2013)

Pace–Sigge, Michael. 2013. *Lexical Priming in Spoken English Usage*. Houndmills, Basingstoke: Palgrave Macmillan.

Rice, Sally. 1999. Patterns of Acquisition in the Emerging Mental Lexicon: The Case of to and for in English. In: *Brain and Language* 68, 268–276

Sinclair, John M., Susan Jones and Robert Daley. [1970] 2004. *English Collocation Studies: The OSTI Report.* London. Continuum.

Sinclair, John M. 1991. *Corpus Concordance Collocation.* Oxford: OUP.

Stubbs, Michael. 1996. *Text and corpus analysis. Computer-assisted analysis of language and culture*. Oxford: Basil Blackwell.

Inaccurate pronunciation in students' interpreting performance: Evidence from a learner corpus

Jun Pan
Hang Seng Management College, Hong Kong
janicepan@hsmc.edu.hk

Jackie Xiu Yan
City University of Hong Kong
ctjackie@cityu.edu.hk

1 Introduction

Although language competence has been regarded as a prerequisite for the learning of interpreting, language deficiencies usually pose as a problem for many students in interpreting classrooms, especially for those at the tertiary level training programmes (see Shaw et al. 2004; Pan and Yan 2012). One of the language problems displayed in students' interpreting performance is inaccurate pronunciation (see Pan and Yan 2012), which may be caused by students' language deficiency, or influenced by the sociolinguistic backgrounds of their language learning. More often, pronunciation problems may result from students' cognitive stress in performing the difficult task of interpreting itself. Therefore, a thorough investigation of interpreting students' problems of inaccurate pronunciation will provide important insights into not only the enhancement of students' interpreting performance but also the improvement of their language competence for the learning of interpreting.

2 Research background

Inaccurate pronunciation has been used as one of the factors in many evaluation schemes of students' interpreting performance (e.g., Lindquist 2005). In many cases, the study of pronunciation problems in interpreting is about phonation, a disfluency indicator, rather than a language problem (e.g., Pio 2003; Yang, 2005; Cai 2007). Despite the understanding that the study of inaccurate pronunciation may offer regarding the psychological and cognitive aspects of interpretation learning, there is a paucity of research in this area.

With the growing recognition of its advantages, the corpus linguistic approach has increasingly been applied in various fields such as language assessment (e.g. Alderson 1996), the study of learner language (e.g. Granger 2002) and the study of features of interpreted language (e.g. Shlesinger 1998). According to Leech (1992), the approach of computer corpus linguistics has four distinguishable focuses:

- Focus on linguistic performance, rather than competence;
- Focus on linguistic description, rather than linguistic universals;
- Focus on quantitative, as well as qualitative models of language;
- Focus on a more empiricist, rather than rationalist view of scientific inquiry. (p. 107)

The development of corpus-based interpreting studies (Shlesinger 1998) has provided existing paradigms for the analysis of interpreting performance. However, investigation of the construction as well as the practical application of interpreting learner corpus is quite scarce.

3 The study

In light of these research gaps, this study aimed to investigate the problem of inaccurate pronunciation in students' consecutive interpreting performance with the application of corpus analysis methods. In this study, a small corpus of university students' interpreting test outputs (i.e., Chinese-English consecutive interpreting) was constructed, which included audio files as well as their transcriptions in computer readable formats. The total transcription work involved audio recordings of a total of 92,400 seconds (i.e., 1,540 minutes). Given the specific purpose of the present study, both linguistic and extralinguistic annotations were added to the corpus data. Problems of inaccurate pronunciation were annotated based on a scheme specifically developed for the present study, the purpose of which was the research of pronunciation problems of interpreting students and the enhancement of their overall interpreting performance, rather than the development of phonological pedagogy for second / foreign language learners. The scheme included the annotation of problematic vowels and consonants at the segmental level and stress problems at the word level, with a subsystem adapted from H. Yang and Wei (2005). The annotated corpus was then analyzed by the software Wordsmith 5.0.

4 Results and implications

The study revealed the distribution of different pronunciation problems of interpreting learners, with segmental problems the absolute majority (over 95%). In contrast, word-level stress errors only accounted for less than 5% of pronunciation problems. Within the segmental problems, consonant errors, especially the substitution of consonants (almost 40%) weighed most as compared to other subtypes of segmental problems.

A further investigation into the mispronounced words indicated that students' sociolinguistic backgrounds such as their dialects played an interesting role in their pronunciation problems in consecutive interpreting. Moreover, the study also provided evidence that students tended to pronounce words inaccurately when they encounter difficult or unfamiliar concepts in interpreting.

Findings of the study suggested that interpretation training should take into consideration students' sociolinguistic backgrounds as well as their cognitive needs. Students usually brought into interpreting classrooms their idiosyncratic features, the synergy of which contributes to some shared path in the learning of interpreting, which is identifiable through the construction of such a learner corpus.

Given the scant number of interpreting learner corpus, the present study provides valuable methodological insights into relevant corpora construction. It also reveals the significance of developing learner corpus in translation and interpreting studies. In addition, findings of this study indicate the important areas in both curriculum development and pedagogical enhancement in interpreter training at different levels.

References

Alderson, C. 1996. "Do corpora have a role in language assessment?". In J. Thomas and M. Short (eds.) *Using corpora for language research: Studies in honour of Geoffrey Leech*. London/New York: Longman.

Cai, X. 2007. *Kouyi pinggu [Interpretation and evaluation]*. Beijing: Zhongguo Duiwai Fanyi Chuban Gongsi [China Translation and Publishing Company].

Granger, S. 2002. "A bird's-eye view of learner corpus research". In S. Granger, J. Hung and S. Petch-Tyson (eds.) *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Leech, G. 1992. "Corpora and theories of linguistic performance". In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium*. Berlin/New York: Mouton de Gruyter.

Lindquist, P. P. 2005. "Technologies, discourse analysis, and the spoken word: The MRC approach: An empirical approach to interpreter performance evaluation and pedagogy". *Meta* 50 (4): 1492-1421.

Pan, J. and Yan, J. X. 2012. "Learner variables and problems perceived by students: An investigation of a college interpreting program in China". *Perspectives: Studies in Translatology* 20 (2): 199-218.

Pio, S. 2003. "The relation between ST delivery rate and quality in simultaneous interpretation". *The Interpreters' Newsletter* 12: 69-100.

Shaw, S., Grbić, N. and Franklin, K. 2004. “Applying language skills to interpretation: Student perspectives from signed and spoken language programs”. *Interpreting* 6 (1): 69-100.

Shlesinger, M. 1998. “Corpus-based interpreting studies as an offshoot of corpus-based translation studies”. *Meta* 43 (4): 486-493.

Yang, C. S. 2005. *Kouyi jiaoxue yanjiu: Lilun yu shijian [Interpretation Teaching and Research: Theory and Practice]*. Beijing: Zhongguo Duiwai Fanyi Chuban Gongsi [China Translation and Publishing Company].

Yang, H. and Wei, N. 2005. *Construction and data analysis of a Chinese learner spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.

The development of non-literal language in L2: Evidence from a text corpus by French learners of English

Justine Paris
Sorbonne-Nouvelle
Vincennes-St-Denis
Justine.paris@univ-paris3.fr

Significant research has shown that metaphor pertains to our way of thinking (Lakoff and Johnson, 1980; Lakoff and Turner, 1989) and to cognition (Gibbs, 1995 and 2006). Some researchers think that the comprehension and production of metaphor –and more generally, of non-literal language- is highly dependent on relational and pragmatic knowledge (Sperber & Wilson, 1986/1995; Carston, 2000; Gentner et al, 2001).

In the light of this theoretical framework, a certain number of academics started to look into non-literal language in relation to second language acquisition and teaching (Danesi, 1992 and 1995; Andreou et coll., 2009). For Danesi (1992a, 1992b et 1995), L2 learners’ use of non-literal language – its absence, in fact – explains why learners never reach a language level comparable to that of a native speaker. For him, mastering figurative language in a foreign language is the sign of a high language proficiency level. L2 learners’ level can be evaluated on their capacity to metaphorize in the L2, as well as on their capacity to adopt the natives’ worldviews / conceptual perception. Other researches then proposed teaching strategies (Irujo, 1986 ; Deignan et al 1997; Cooper, 1998; Lennon, 1998; Boers, 2000 et 2009; Komur et Cimen, 2009) but most of them focus on figurative language comprehension (Cooper, 1999; Kosciuk, 2003; Siegal and Surian, 2004). Still, little is known about L2 learners’ production of non-literal language.

In order to get a broad picture of L2 learners’ non-literal performance, two groups of thirty students majoring in French literature and Communication Studies in France were asked to write essays for the purpose of a required English course. The students were separated in terms of their proficiency level in English: participants in group 1 showed a B1 level (with reference to the Common European Framework of Reference for Languages – CEFR) while participants in group 2 showed a C1 level. This enabled us to gather data on significantly different proficiency levels and to observe their impact on the learner’s capacity to produce non-literal language in their L2. In order to avoid any priming effects, the essay topics did not encourage the students to be metaphorical in their writing (e.g. *could you go and live in a far away country?*). Six types of non-literal sequences building could be identified in the learners' essays – namely overextensions, L1 transfers, personifications, comparisons, idioms and metaphors. Each of these forms and their functions in learner discourse will be presented before outlining the general trends found in the corpus (learners' preferences in relation to their proficiency level).

Despite the differences in proficiency levels, non-literal language was globally rare in the essays, and a large majority of this language reflected the L1 of the learners. Looking at these results, we asked the participants to rewrite their essays in their mother tongue, French, for the purpose of cross-linguistic observations. Since we know that language learners tend to cling to literalness in their L2 – up to giving an "unnatural over-literalness" aspect to their discourse (Danesi, 1992) – it appeared relevant to observe the way our participants made use of non-literalness in their mother tongue, to get an element of comparison. These new essays revealed that the majority of L1 transfers observed in the English essays resulted from prefabricated forms in French, i.e. relatively frozen sequences (French idioms, collocations, and formulae). These sequences happened to be the most common non-literal ones found in the essays in French. Therefore, it is not surprising that the learners attempted to use them when using the L2.

Overall, the results showed that metaphorizing in one's L2 is possible, contrary to a pessimistic view in the literature. Non-literal sequences may be much rarer in L2 than in L1 but it is yet present rather early in the learning process – even if they may reflect the L1's conventions. In addition, it seems like we do not treat non-literalness the same way depending on the type of language that we are using (one's native vs. a foreign language). While the essays in L1 revealed a high number of conventional figurative forms (idioms) and fewer novel ones; the essays in L2 revealed an experimental approach to non-literalness via transitory figurative forms (overgeneralization, L1 transfers) and a preference for functional non-literal forms (discursive idioms). Finally, the various corpus analyses showed that the development of non-literal use in L2 is gradual, and that it progressively evolves towards a system resembling the L1's.

References

Andreou, G. & I. Galantomos (2009). Conceptual Competence as a Component of Second Language Fluency. *Journal of Psycholinguistic Research, 38*(6): 587-591.

Boers, F., Piriz, A. M. P. et al. (2009). Does Pictorial Elucidation Foster Recollection of Idioms? *Language Teaching Research, 13*(4) : 367-382.

Boers, F. (2000). Metaphor Awareness and Vocabulary Retention. *Applied Linguistics 21*(4) : 553-571.

Carston, R. (2000). The Relationship Between Generative Grammar and (Relevance Theoretic) Pragmatics *Language & Communication*, *20*(1) : 87-103.

Cooper, T. C. (1998). Teaching Idioms. *Foreign Language Annals, 31*(2), 255-266.

Cooper, T. C. (1999). Processing of Idioms by L2 Learners of English. *Tesol Quarterly 33*(2): 233-262.

Danesi, M. (1992a). Metaphor and Classroom Second Language Learning. In J. Beer, C. Ganelin & A. J. Tamburri (Eds.), *Rla : Romance Languages Annual 1991, Vol 3* (Vol. 3, pp. 189-194).

Danesi, M. (1992b). Metaphorical Competence in Second Language Acquisition and Second Language Teaching: The Neglected Dimension. In J. E. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics (« Language, communication, and social meaning »*, pp. 489-500.

Danesi, M. (1995). Learning and Teaching Languages: the Role of "Conceptual Fluency". *International Journal of Applied Linguistics*, *5*(1): 3-20.

Deignan, A., Gabrys, D., et Solska, A. (1997). Teaching English Metaphors Using Cross-Linguistic Awareness-Raising Activities. *ELT Journal, 51*(4), 352-360.

Gentner, D. (2001). Metaphor is Like Analogy. In Gentner,D., Holyoak,K.J., Kokinov, B.N. (Eds), *The Analogical Mind: Perspectives from cognitive Science.* Chapter 6.

Gibbs, R. W. (1995). Idiomaticity and human cognition. *Idioms: Structural and psychological perspectives*: 97–116.

Gibbs Jr, R. W. & Tendahl, M. (2006). « Cognitive effort and effects in metaphor comprehension: Relevance theory and psycholinguistics ». *Mind & Language* 21 (3): 379–403.

Irujo, S. (1986b). A Piece of Cake: Learning and Teaching Idioms. *English Language Teaching Journal, 40*(3): 236-242.

Komur, S., & Cimen, S. S. (2009). Using conceptual metaphors in teaching idioms in a foreign language context. *Sosyal Bilimler Enstitüsü Dergisi (İLKE) papers*, 205-222.

Kosciuk, M. (2003). An Investigation of Metaphor Comprehension by Two Second Language Learners. In B. Bartlett, F. Bryer & D. Roebuck (Eds.), *Reimagining Practice: Researching Change, Vol 2* (pp. 116-130).

Lakoff, G., & Johnson, M. (1980). *Metaphor We Live By*. Chicago, University of Chicago Press.

Lakoff, G., & Turner, M. (1989). *More than Cool Reason: A Field Guide to Poetic Metaphor.* University

Of Chicago Press.

Lennon, P. (1998). Approaches to the Teaching of Idiomatic Language. *Iral-International Review of Applied Linguistics in Language Teaching, 36*(1), 11-30.

Siegal, M., et Surian, L. (2004). Conceptual Development and Conversational Understanding. *Trends in Cognitive Sciences, 8*(12), 534-538.

Sperber, D. &Wilson, D. (1986/1995). *Relevance, Communication and Cognition*. Oxford: Blackwell.

Vocabulary profiling of Canadian High School Diploma exam expository writing

Geoffrey G. Pinchbeck
University of Calgary, Canada
ggpinchb@ucalgary.ca

1 Introduction and background

This paper presentation will examine the relationship between written vocabulary use and academic success in mainstream, university-bound Canadian high-school students. Canadian large urban centres are undergoing a rapid demographic shift (Citizenship and Immigration Canada, 2008), one result of which has been a call for academic language to be given a more prominent role in mainstream public educational planning across the curricula in Canada (Biemiller, 2012) and this parallels similar trends in the U.S. (Nagy and Townsend, 2012; Ranney, 2012; Snow, 2010) and elsewhere. This call for research was inspired initially from studies on children of immigrants who do not use a majority language at home, and who have been identified as academically at risk in secondary and post-secondary settings (Abada et al., 2008; Roberge, Siegal, & Harklau, 2010). English Language Learners (ELLs) in an English-dominant K-12 school system rapidly acquire competence in spoken English syntax and phonetics, and therefore are often quickly moved from English as a Second Language classes into mainstream classes, despite a significant but less apparent gap in academic lexical competence as compared to more English-proficient students (Cameron, 2002). Although the relationship between language and academic success is well known (e.g. Hart and Risley; Verhoeven et al., 2011) and commonly identified in research on ELLs (e.g. Roessingh and Douglas, 2012), the lack of an operationalized model of adolescent academic language development impedes attempts to assess and strategically promote vocabulary development of all learners (i.e. monolingual L1, bilingual L2, and minority-language ELLs).

2 Rationale

There is a rich research literature in adult second language acquisition and early childhood language development; however, empirical research on vocabulary development in secondary school (i.e. grades 7-12) is lacking. One obvious difference between mainstream K-12 and ESL/EFL settings is that learners within the mainstream school system are not placed by their language proficiency, but according to their age, and this may have led to a wide range in academic language competencies co-existing within the same grade level. Conventional instruments available for language measurement are expensive, not diagnostic, not based on any developmental model, and are not systematically administered (Pearson et al., 2007). The construction and analysis of a representative corpus of high-school writing would permit a description of adolescent vocabulary development and might allow vocabulary thresholds required for given levels of academic success to be estimated.

3 Methods

A random sample 1600-student database of high-school exam essay texts (>1,500,000-word corpus) and associated transcript data were obtained from the government. Lexical frequency profiles (Anthony, 2009; Heatley and Nation, 1994; Laufer and Nation, 1995) were generated by aligning essay vocabulary with word lists derived from reference corpora of 1) adult British (Cobb, n.d. b), 2) adult American English (Davies, 2010), and, importantly, 3) a K-12 American textbook and reader corpus (Zeno et al. 1995). Using a regression approach, vocabulary profiles were then compared to the following associated data: 1) official government exam essay scores (holistic trait-based rubric), 2) writing error data (detailed analytical coded rubric) (EFWR, 1993; 2003), and 3) student high-school transcripts. Additionally, individual word families used by academically successful students were compared to those in writing of average students by Text-Lex Compare (an identical technique to Cobb, n.d. a). Finally, lexical frequency profiles for sub-corpora of words that were unique to average students, or to successful students, or shared by both groups were then obtained to determine the range in word frequencies that exist in the writing across this population.

4 Results

First, a comparison of transcripts and essay scores of the random sample population with the whole population of grade 12 academic-track students strongly indicated that the sample is representative. Second, using regression analytical approaches, we have identified two lexical indexes that independently explain large and significant variance of both essay quality and general academic success. Third, a lexical subtraction/comparison of written word usage by average or successful students also indicated that the bulk of the differences that exist within this population of high-school L1 and bilingual learners falls within a domain of mid-frequency vocabulary (Schmitt and Schmitt, 2012).

5 Significance

This analysis should inform a strategic K-12 academic language pedagogy by 1) providing transparency to the construct of academic language that is required for academic success, 2) enabling students, teachers, and program designers to more accurately match learners with vocabulary-appropriate reading materials, 3) allowing at-risk students to be identified more quickly, and 4) making data from a corpus of written academic discourse of Canadian, university-bound grade 12 students available for further research.

References

Abada, T., Hou, F., & Ram, B. (2008). *Group differences in educational attainment among the children of immigrants.* (11F0019M. No. 308). Retrieved from http://www.statcan.gc.ca/pub/11f0019m/11f0019m2008308-eng.pdf.

Anthony, L. (2009). AntWordProfiler. [*Computer software*]. *http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html*.

Biemiller, A. (2012). Words for English-Language Learners. *TESL Canada Journal, 29*(Special Issue 6), 198-203.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*(2), 145-173. doi: 10.1191/1362168802lr103oa

Citizenship and Immigration Canada. (2008). *Annual report to parliament on immigration, 2008*. Ottawa: Retrieved from www.cic.gc.ca/english/pdf/pub/immigration2008_e.pdf.

Cobb,T. (n.d. a). Text Lex Compare [Online computer software. Available: http://www.lextutor.ca/text_lex_compare/

Cobb,T. (n.d. b). Web VP/BNC-25 Vocabprofiler [Online computer software] (an adaptation of Heatley and Nation’s, 1994 Range). Available: http://www.lextutor.ca/vp/bnc

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34, 213-238.

Davies, Mark. (2010). The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing, 25*(4), 447–465. doi: 10.1093/llc/fqq018

EFWR. (1993). Detailed Marking Code: The Effective Writing Programme, University of Calgary.

EFWR. (2003). The Assessors’ Guide for the Effective Writing Test: The Effective Writing Programme, University of Calgary.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H. Brookes Publishing Co.

Heatley, A., & Nation, P. (1994). Range [Computer software]. Wellington, NZ: Victoria. University of Wellington. Available: http://www.victoria.ac.nz/lals/about/staff/paulnation

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, *16*, 307-322. doi: 10.1093/applin/16.3.307

McCarthy, Philip, & Jarvis, Scott. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav Res Methods, 42*(2), 381-392. doi: 10.3758/brm.42.2.381

Nagy, William E., & Townsend, Dianna. (2012). Words as Tools: Learning Academic Vocabulary as Language Acquisition. *Reading Research Quarterly, 47*(1), 91-108. doi: 10.1002/RRQ.011

Pearson, P. David, Elfrieda, H. Hiebert, & Michael, L. Kamil. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282-296.

Ranney, Susan. (2012). Defining and Teaching Academic Language- Developments in K-12 ESL. *Language and Linguistics Compass*, *6*(9), 560-574. doi: 10.1002/lnc3.354

Roberge, M., Siegal, M., & Harklau, L. (Eds.). (2010). *Generation 1.5 in College Composition: Teaching Academic Writing to U.S.-Educated Learners of ESL.* New York: Routledge.

Schmitt, Norbert, & Schmitt, Diane. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, FirstView*, 1-20. doi:10.1017/S0261444812000018

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary Growth and Reading Development across the Elementary School Years. *Scientific Studies of Reading, 15*(1), 8-25. doi: 10.1080/10888438.2011.536125

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1999). *The educator's word frequency guide.* New York: Touchstone Applied Science Associates.

I won't do that again! - Using Task-Specific and Individual Learner Corpora to Enhance the Noticing, Awareness and Editing Skills of Advanced ESP Students

Catherine Riley
School of International Studies
Università degli Studi di Trento
catherine.riley@lett.unitn.it

The students on an English medium postgraduate course in International Studies are (excessively?? Purcell 1998) concerned with improving formal accuracy in both their spoken and written production. Taking full account of student perceptions of their own language needs is essential if motivation is to be kept high (Dörnyei). Therefore even though the English Language course is primarily content focused and task based, some language awareness and focus on form exercises and activities (e.g. Long 1991, Ellis 2003) are provided, not least to meet the students' own perceived needs (cfr Jodaie *et al.* 2011). Moreover, a balance between a meaning driven approach (content) and focus on form is conducive to successful learning (Lyster 2007).

Written tasks vary in length and type, from press releases to book reviews, academic papers to institutional reports, formal letters to informal emails. In the past, alongside general feedback, both positive and negative (Hyland and Hyland 2006) a self-correction approach, where error types are indicated and students re-edit their own work, followed by a second correction has proved to be of some effect, though some errors, both idiosyncratic and transversal, persist. Ever in search of making corrective feedback more effective and also enhancing the effectiveness of language focused activities/lessons, an approach developing greater awareness was sought. To this end, a self-built learner corpus of current and previous years' assignments and papers was used to generate different types of language exercises, which the students professed to find useful. However, the effect of these form-focused exercises was difficult to quantify. Indeed, some common errors continued to persist. It was therefore decided that a more personal approach was needed. Rather than use the growing course corpus, micro Task-specific corpora were created after each assignment. It was thus easy to identify common mistakes in similar contexts and text types and generate focused language exercises. Moreover, it provided an opportunity to try to understand WHY the students, of different L1 backgrounds, persisted in making these mistakes.

Using the concordance in class to identify and illustrate the common errors and incorrect (and correct) usage in their own work and not merely to create exercises has a stronger awareness raising effect on many of the students. The concordances are all anonymous, but students are able to see how many are making the same mistake and also to identify instances of the correct use of the form from their classmates' work. The task-specific corpus is also used to create exercises on specific items. Students are particularly aware of their own sentences (whether correct or not), even though they are completely anonymous.

A similar approach is used with spoken language. Students perform various oral tasks of varying lengths and types, scripted, semi-scripted and unscripted. These are recorded, usually only the audio but at least twice a semester also video recorded. Students then check scripts against delivery or write transcripts, which they then correct and comment (self evaluation). Both class and individual feedback is given using the task/class/individual corpora of transcripts and focused exercises are created. Identifying good use of language as well as errors is encouraged. Students are invited to practice selected extracts of their revised speeches in pairs and in plenum to both personalise and practice new/unfamiliar/problematic forms, including pronunciation and prosody, and to eradicate fossilised erroneous forms.

One positive outcome is the degree of negotiation of meaning and metalinguistic discussions also regarding usage in the students' different L1s. Moreover, students develop an awareness of certain features of lexis in use such as connotation, collocation and colligation (cfr Carter 2006). This awareness has translated into improved noticing skills (Harmer 2003), and also more careful editing (Myhill and Jones 2006). In individual feedback sessions, students have also reported using the online tools used in class, such as collocation dictionaries, Google as a corpus, the BYU corpora (Davies), while drafting work. Students have also been encouraged to create a personal corpus to monitor their own progress on recurring errors.

It is difficult to understand whether the use of task-specific and individual corpora is more effective than other form-focused activities. However, qualitative feedback suggests the approach is greatly appreciated. Moreover, their enhanced metalinguistic knowledge and noticing skills are transferrable to content courses, in particular in the analysis of legal texts and cases (Riley 2013). Students have also mentioned being on the look out for their own typical/idiosyncratic mistakes. Indeed, more than one student, referring to a mistake either highlighted in class or included in an exercise created with the concordancer has declared 'I won't do that again!'

References

Carter, R., 2007. "Spoken Grammar, Written Grammars: From Corpus to Classroom", TESOL Colloquium, Paris December 2007

Davies, M. (various dates) http://corpus.byu.edu/

Dörnyei, Z. and E. Ushioda. 2010. *Teaching and Researching: Motivation (Applied Linguistics in Action)* London: Longman.

Ellis, R. 2003. *Task-based Language Learning and Teaching*. Oxford: Oxford University Press

Harmer, J. 2003. "Do your students notice anything?" *Modern English Teacher* 12/4: 5-14

Hyland, F. and Hyland, K. 2006. "Feedback on Second Language Students' Writing", *Language Teaching* 39/2: 83-101.

Jodaie, M, *et al.* 2011. "A Comparative Study of EFL Teachers' and Intermediate High School Students' Perceptions of Written Corrective Feedback on Grammatical Errors' *English Language Teaching* 4/4: 36-48

Long, M. 1991. "Focus on form: A design feature in language teaching methodology". In de Bot, K., Ginsberg, R., & Kramsch, C. (eds.) *Foreign Language Research in Cross-cultural Perspectives*. Amsterdam: John Benjamins. 39-52.

Lyster, R. 2007. *Learning and Teaching Languages through Content: A counter-balanced approach*. Amsterdam: John Benjamins.

Myhill, D and Jones, S. 2006. "More than Just Error Correction" *Language Teaching* 39:83-101

Purcell, K., 1998. "Making Sense of Meaning: ESL and the Writing Center". *The Writing Lab Newsletter* 22:1-5

Riley, C.E. 2013. "A Long Hard Climb – Getting from the Bottom to the Top of the CLIL Incline". *Recherche et pratiques pédagogiques en langues de spécialité*, 32/3: 30-56.

Evaluation of the frequency and types of written errors in French as a foreign language: a corpus-based analysis

Ariane Ruyffelaert
Department of Linguistics, Ghent University;
Department of French Philology,
University of Granada, Spain
ariane.ruyffelaert@ugent.be

1 Introduction

In foreign language learning in academic contexts, it is a challenge to achieve a near-native level. This is especially important in the case of future foreign language teachers, because they are responsible to translate their knowledge to the next generation. In this sense, a continuous formation is required to obtain an academic level. However, the frequency and types of written errors made by future teachers of French as a foreign language (FFL) is still unknown. For this reason, the aim of this study was to analyze the frequency and types of written errors in essays made by Spanish postgraduate students in FFL teaching.

2 Methodology

The corpus consisted of written essays conducted with all the students (n=11) of a Postgraduate program in FFL teaching from the University of Granada (Spain). The students had all acquired their master's degree in French philology or translation. They were asked to write an opinion essay (with a maximum of 400 words) in French about the importance of FFL in a multilingual Europe. This general topic was selected for the compositions so that the content was the least constrained by thematic limitations. Learners could not make use of any additional help source (dictionary, grammar or textbook, nor were they allowed to ask the teacher, researcher or their classmates for help). The writing assignment was conducted in their own classroom under their teacher and the researcher's supervision. The students were assigned one hour to complete the task.

The handwritten essays were collected and transcribed into word files using Microsoft Office Word 2007. All essays were assessed in two different phases. In the first phase of the analysis, the essays were uploaded to an online correction tool, called *CorpuScript (*http://www.schrijven.ugent.be/opdrachten/*).* This tool was developed at Ghent University (Belgium) to address the need to grouping errors in categories defined in advance. Thus, effects related to the subjectivity of correction are more or less neutralized (Hadermann and Demeulenaere 2013). For this study, 14 types of written errors were evaluated. They can be categorized in five groups: (1) spelling mistakes; (2) grammar errors; (3) vocabulary errors; (4) discourse errors and (5) content errors. In addition, the following subtypes of errors: omission, misselection, inclusion and order were considered in the groups of errors 2, 3 and 4 (James 1998). In the second phase of the analysis, the written compositions were scrutinized for all types of errors as described above. The errors were identified, classified, and counted up with the use of *CorpuScript*.

3 Results

The text length of the analyzed essays ranged from 233 to 619 words (mean = 391.54). Analysis of the data revealed that all the essays presented different kind of errors. A total of 280 errors were observed, with an average of 25.45 errors per text. Categorizing these errors, we found that grammar errors were the most common with a total of 95 (33.93 %) followed by 89 (31.79 %) vocabulary and 88 (31.43 %) spelling errors. Finally, discourse (1.79 %) and content errors (1.07 %) were the most infrequent. The in-depth analysis by group of errors showed the following results: of the 95 grammar errors, 48 (50.53 %) were omission type errors followed by 28 (29.47 %) misselection, 15 (15.79 %) overinclusion and 4 (4.21 %) order errors. Within the 89 vocabulary errors, we observed 80 (89.89 %) misselection, 4 (4.49 %) order, 3 (3.37 %) overinclusion and 2 (2.25 %) omission errors.

4 Discussion and conclusion

This study revealed a hierarchy of frequency of error types. Grammar errors were the most frequent. Lexical errors and misspellings follow grammar errors in decreasing order of frequency. Discourse and content errors were very infrequent. These results allowed us to confirm that these students dominate discourse organization. However, despite the fact that the postgraduate students have acquired their master's degree, they still commit surprisingly a lot of grammar errors and experience problems finding the adequate lexicon. Practice of discourse organization plays an essential role in the development of writing skills. Nevertheless, vocabulary training and writing is also necessary to develop and enlarge vocabulary (retaining newly learned words, activating receptive vocabulary into a productive one) (Muncie 2002; Lee 2003 and Llach 2007).

In conclusion, the presence of these errors could have a negative impact in the professional performance of these advanced learners or future teachers. They must practice their writing skills with the aim of remedying their errors and enhancing the quality of their writing. This way they will be better teachers for their future learners.

Acknowledgments

This study was financed by the Department of French Linguistics, Ghent University, Belgium. The authors are grateful to the Professors P. Hadermann (UGent, Belgium) and J. Suso (UGR, Spain).

References

Hadermann, P. and Demeulenaere, A. 2013. “Perfectionnement de la compétence écrite en langue étrangère. Littératie et environnement d’apprentissage informatisé.” LE LANGAGE ET L'HOMME. 48(1). p.163-174.

James, C. 1998. Errors in language learning and use: exploring error analysis. London: Longman.

Lee, S. H. 2003. “ESL learners’ vocabulary use in writing and the effects of explicit vocabulary instruction.” System31, 537–561.

Llach, M. P. A. 2007. “Lexical errors as writing quality predictors.” Studia Linguistica, 61: 1–19.

Muncie, J. 2002. “Process writing and vocabulary development: comparing Lexical Frequency Profiles across drafts.” System30, 225–235.

Teaching a foreign language through learner corpus

Marilei Amadeu Sabino
UNESP – São José do Rio Preto, Brazil
amadeusm@ibilce.unesp.br

Learner corpus research (LCR) stands at a crossroads among some disciplines as corpus linguistics, second language acquisition, foreign language teaching, and the results of the investigations conducted in this area may bring benefits to several research fields, namely, lexicography, contrastive linguistics, teaching methodology, cognitive linguistics, second language acquisition, foreign language teaching, language testing, natural language processing and translation.

Collocations are one of the several types of phraseologisms and although a lot has already been done in terms of phraseological research, it still remains a lot to be done in terms of extracting, describing, defining, teaching and learning these structures.

Granger et al. (2002, p. 7) argue that computer learner corpora are “[…] electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose”. A very significant advantage of learner corpora is the fact that the researcher can have a record of the learners’ production which may enable him to report what learners actually produce in terms of phraseological patterns.

Altenberg and Eeg-Olofsson (1990), Sinclair (1991), Fontenelle (1994), Granger (1998), Orenha-Ottaiano (2004; 2012), Meunier and Granger (2008) claim that the learning of collocations and other prefabricated chunks is crucial to learners who aim to produce fluent speech and they assert that the use of corpora in the foreign language classrooms promotes the teaching of these chunks. Thus, based on the well-known importance of providing students with the ability to use these prefabricated structures well, we built a parallel learner corpus made up of students’ translations from Portuguese into Italian language. Therefore, this paper aims at showing some results of an investigation carried out in a Brazilian public university with students that attend a translation course.

The subjects of this research are university students from the 3rd year of a B. A. in Translation Course, whose level of Italian varies from intermediate to upper-intermediate. The original texts that comprise the corpus are newspaper articles taken from very popular Brazilian newspapers and magazines. The typology of the texts is related to current world news and the topics selected were

"One year after Tsunami in Japan"; "Financial crises in Greece and in Europe"; "Unemployment"; "Elections in the US"; "Bullying"; "Abortion", etc. These texts originally written in Portuguese were translated into Italian by a group of 10 students. With the help of *WordSmith Tools* (Scott 2004), it was possible to extract the data and analyse students' collocations.

The methodology of this investigation, corpus design and compilation are based on a similar research carried out by Orenha-Ottaiano (2012) in the same university, with the same translation students, the same original Portuguese texts, but translated into English.

Our aim is to compare, in a second stage, the collocations used by the Brazilian learners of Italian to the ones employed by the Brazilian learners of English, in order to check if:

a) Brazilian learners of English and Italian as foreign languages have the same difficulties in producing collocations;

b) they produce similar collocational errors; and

c) there is some kind of influence of the mother tongue on their choices.

Some of the problems found in the translation from Portuguese to Italian are related to the following collocations: "cessar fogo", "travar combates", "máxima autoridade rebelde", "governo transitório", "medidas de prevenção", "chegar ao poder", "zona do euro", "cobrir os empréstimos", "pacote de cortes", "rombo fiscal", to name a few.

For example, as learners are usually influenced by their mother tongue (Portuguese), they translated the collocation "entrevista coletiva" into "conferenza collettiva", when they should have used "conferenza stampa". And by ignoring the frequently used collocation "derrubou a resistência" in Italian, they translated it into "ha rovesciato la resistenza", "ha annullato la resistenza", "ha fatto cadere la resistenza", instead of into "ha piegato la resistenza".

The investigation allowed us to observe the students' collocational choices and patterns; the influence of the mother tongue on these choices; the most frequent collocational errors produced; and the most/least used type of collocations employed by them.

As a result of their production, we recognize the importance of teaching and encouraging students to explore the potential benefits of using corpora in translation. We also argue that when the teaching of collocations is in a more explicit (or intentional) way, it brings more benefits to learners than in the cases teachers hope it happens automatically, i. e., in an implicit (or incidental) way. As previously mentioned, the results of this research will be compared to Orenha-Ottaiano's findings and further discussed in a paper.

References

Altenberg, B.; Eeg-Olofsson, M. 1990. "Phraseology in Spoken English: presentation of a Project". In: AARTS, J.; MEIJS, W. (Ed). Theory and practice in Corpus Linguistics. Amsterdam: Randpi, p. 1-26.

Fontenelle, T. 1994. "Towards the construction of a collocational database for translation students". Meta 39 (1), p. 47-56.

Granger, S. 1998. *Learner English on computer*. London/ New York: Longman.

Granger S.; Hung, J.; Petch-Tyson, S. (Ed.) 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: John Benjamins.

Meunier, F.; Granger, S. 2008. "Phraseology in foreign language learning and teaching. Where to and from?" In: MEUNIER, F.; GRANGER, S. (Ed.). Phraseology in foreign language learning and teaching. Amsterdam: John Benjamins, p. 247-252.

Orenha-Ottaiano, A. 2004. *A compilação de um glossário bilíngüe de colocações, na área de jornalismo de negócios, baseado em corpus comparável*. Master's thesis, Universidade de São Paulo, São Paulo.

Orenha-Ottaiano, A. 2012. "English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy". *Acta Scientiarum*, 34 (1), p. 241-251.

Sinclair, J. 1991. *Corpus, concordance and collocation*. Oxford: Oxford University Press.

Thomas, J. E. (forthcoming). "Stealing a march on collocation". *TALC 10 Procceedings*.

Understanding L2: the case of polysemous words in fixed expressions

Henry Tyne
Université de Perpignan Via Domitia
henry.tyne@univ-perp.fr

There is much evidence in the literature of various types of transfer or influence in L2 (from L1, last language acquired…). Putting aside the obvious influences that are often labelled 'errors', there is also influence in the form of knowledge or norms shaped by general language experience (Slobin 1997) usually via L1. This has been observed in the case of communicative competence, for example (Shaw 1992).

In the field of lexis, there have been studies looking at the emergence of meaning-form associations in L2. The classic "eye for eye" experiment (Kellerman 1986) on the role of L1 knowledge on learners' judgments of polysemous words showed how similarity and frequency intuitions in L1 can lead to the prediction of transferability to the L2. While these findings tell us little about the actual use of given words in the target language (see Laufer 2000 on the question of avoidance in L2), they do give us a means of investigating the role played by 'knowledge' and informed judgments. The question of the stimulus used for investigating this type of phenomenon is rarely discussed in detail and it is often assumed that words or expressions are somehow either 'there' (i.e. acquired) or not and that L1 (or other) influence is somehow constant.

This paper studies presents findings from a study of learners' (c. 40) judgements of certain nonliteral expressions containing polysemous words denoting body parts in L2 French (*tête*, *nez*, *œil*…; head, nose, eye…). In particular, it looks at how L2 forms are apprehended by learners according to the stimuli that accompany judgment questions. These are: 'naked' samples with simple definitions and samples within concordances. Learners are asked to rank given expressions presented in pairs according to acceptability in L2; they are also asked to rate their degree of understanding of each expression and whether there is an 'equivalent' in L1.

Results show that the nature of the stimulus can impact upon the way learners deal with forms to arrive at informed judgments. The results also show that L1 knowledge (i.e. perceived similarity with L2 expressions) and the amount of time spent in the L2-speaking environment can affect the degree of accurateness and certainty in the learners' judgements in relation to the type of stimulus.

Corpus use for reference purposes in language learning/teaching is fairly well documented and various studies have compared corpus use with dictionary use, for example. In this paper we consider corpus use for encouraging informed judgments: how does the way the language is presented to the learner equate with the learner's ability to 'understand' or accept certain L2 expressions, some of which may never have been encountered before? This paper suggests there may be a trade-off between 'acceptability' and 'understanding' in some instances and a case is made for greater use of enhanced input through data-driven learning.

References

Kellerman, E. 1986. An eye for an eye: crosslinguistic constraints on the development of L2 lexicon. In E. Kellerman & M. Sharwood Smith (eds.), *Crosslinguistic Influence in Second Language Acquisition*. Oxford: Pergamon, p. 35-48.

Laufer, B. 2000. Avoidance of idioms in a second language: the effect of L1-L2 degree of similarity. *Studia Linguistica* 54(2): 186-196.

Shaw, P. 1992. Variation and universality in communicative competence: Coseriu's Model. *TESOL Quarterly* 26(1): 9-25.

Slobin, D. 1997. The universal, the typological and the particular in acquisition. In D. Sloblin (ed.), *The Cross-Linguistic Study of Language Acquisition. Vol. 5. Expanding the Contexts*. Mahwah, New Jersey: Lawrence Erlbaum, p. 1-39.

Teaching Czech Verbs on Elementary Level: Comparing Textbook Corpus and National Corpus Data

Pavlína Vališová
Masaryk University
evalisova@gmail.com

This paper presents a project of compilation of a small specialised corpus, which consists of textbooks of Czech language as a foreign language (CFL), and discusses how this type of pedagogical corpus can be used in language teaching. Comparing textbook language with authentic language in national corpora it is possible to determine, if we really teach students authentic Czech and how we can improve the teaching materials.

The planned corpus consists of 17 contemporary textbooks of Czech as a foreign language and it is divided in three parts according to the Common Framework of References of Languages (CEFR) – A1, A2 and B1. This paper uses the A1 subcorpus, which has been already finished, and includes 11 textbooks: 7 complete elementary textbooks and 4 parts of textbooks which are intended for more levels (approx. 100 000 words). The Sketch Engine tool is used to build and explore the corpus.

The present analysis focuses on the verb forms in elementary level textbooks and their context. The method is corpus-driven (Huston, 2010): first, the verbs presented in 11 textbooks of Czech as a foreign language for A1 level are examined and then compared with the Czech National Corpora and the description of A1 level according to the CEFR (Hádková, 2005)[12].

The A1 level according to the Common European Framework for Languages (CEFR) is considered as the lowest level, which does not cover almost any grammar. Nevertheless, Czech, as a highly inflected language, has to include more grammatical features than analytical languages in its level description according to the CEFR. The students on A1 level can: "Use simple phrases and sentences to describe where I live and people I know." (European Levels – Self Assessment Grid) and without the basic grammar minimum, it would be made impossible for the student to understand and speak (Cvejnová, 2010). However, elementary textbooks tend to simplify the language as much as possible, e.g. try to avoid perfective verbs despite the fact that some of them belong to the most frequent vocabulary, e.g. *přijít, zapomenout, zůstat*, as the data from the national corpus shows.

The grammar minimum in A1 level includes: present, past and future tense of imperfective verbs and modal verbs. Thus, our research questions focus on which verbs we should choose to be presented on elementary level: 1) Should we present perfective verbs on this level already? 2) Which modal verbs should we choose?

The aspect category lies on the borderland between grammatical category and lexical meaning of the verbs (Cvrček, 2010: 245). The most of Czech verbs exist in pairs which differ in expressing finished action or emphasizing the result of the action (e.g. *udělat, koupit*) and imperfective verbs which express unfinished action or process (e.g. *dělat, kupovat*), but the category is far more complex – the verbs can be created by prefixes or suffixes which can change the meaning of the verb as well. That is why the students on A1 usually learn only imperfective verbs and the aspect category itself is a part of the syllabus later (levels A2 and B1). Comparing the textbook input with authentic data shows interesting results that frequent perfective verbs are presented merely in one particular form to memorize (infinitive, imperative or past tense) in A1 level textbooks. Therefore, national corpus could be an excellent assistant in choosing the most frequent form and the suitable context. The same applies for the modal verbs as well because the choice and use of modal verbs vary in A1 textbooks.

The CFL textbooks do not take into account the most frequent collocations because they usually choose the context related to the topic of the lesson. The textbooks often contain invented texts that have been constructed for didactical purposes around a particular topic or grammatical feature and do not include authentic examples. It is argued that corpus evidence should be taken into account in order to achieve a higher degree of naturalness in textbook language (Bernardini 2004; Römer 2005). The data from CFL textbook subcorpus prove that even elementary level can benefit from the national corpus, e.g., in the choice of verbs and their most common context. This type of specialised corpora thus helps to improve the teaching materials to present contemporary and authentic language.

References

Bernardini, S. 2004. "Corpora in the classroom. An overview and some reflections on future developments". In J. Sinclair. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Cvrček, V. et al. 2010. *Mluvnice současné češtiny.* Praha: Karolinum.

Czech National Corpus – SYN2010. 2010. Praha : Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.

[12] The A1 level in Czech is important since it is a level of the Examination of the Czech Language for Permanent Residence in the CR: <http://check-your-czech.com>.

European Levels – Self Assessment Grid. Accessible at: <http://europass.cedefop.europa.eu/en/resources/european-language-levels-cefr>.

Gouveneur, C. and Meunier, F. 2009. “New types of corpora for new educational challenges. Collecting, annotating and exploiting a corpus of textbook material”. In K. Aijmer (ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Hádková, M., Línek, J. and Vlasáková, K. 2005. *Čeština jako cizí jazyk. Úroveň A1*. Olomouc: Univerzita Palackého v Olomouci.

Hrdlička, M. 2010. *Gramatika a výuka češtiny jako cizího jazyka*. Praha. Karolinum.

Huston, S. 2010. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Römer, U. 2004. “A corpus-driven approach to modal auxiliaries and their didactics”. In J. Sinclair (ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 2004, s. 185–199.

Römer, U. 2005. “Looking at looking: Functions and contexts of progressives in spoken English and 'school' English”. In A. Renouf and A. Kehoe (eds.). *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.

The Common European Framework in its political and educational context. Accessible at: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Hands-on and hands-off data-driven learning of verb-preposition collocations in L2 German

Nina Vyatkina
University of Kansas
vyatkina@ku.edu

1 Introduction

This study reports on the learning outcomes from a data-driven learning (DDL) intervention for teaching verb-preposition collocations to college-level American learners of German. Following Boulton (2012), the study compares the effects of paper-based and computer-based DDL activities and explores correlations between learning outcomes, learner proficiency, and DDL receptivity.

2 Target structure

Collocations are an important aspect of depth of vocabulary knowledge. Concordance exercises have been shown to be more beneficial for teaching collocations than traditional activities (Cobb 1997, Daskalovska 2013, Liou et al. 2012). Boulton (2012) has shown that computer-based (hands-on) and paper-based (hands-off) activities were equally effective for teaching certain English verb constructions.

The present study continues this line of research focusing on L2 German and verb-preposition collocations. These collocations cause considerable lexico-grammatical difficulties for learners because there is no direct equivalence between the German and English prepositions and because either verb or preposition can assign inflectional markers to the subsequent noun phrase.

3 Participants and institutional setting

The DDL intervention was administered in a German course taught by the researcher at a large public North American university. 10 L1 English participants with an average German proficiency of B1.2 (CEFR) took the course for their major, minor, or an elective. Course activities included extensive reading, writing, discussion, grammatical analysis, and regular assignments based on the DWDS corpus [13] that prepared participants for the experiment.

4 Research questions

The research questions draw and expand on

[13] www.dwds.de

Boulton's (2012):

1. Do learning outcomes improve following DDL intervention and are the gains retained after a month?
2. Is there a difference in learning outcomes on paper and on computer?
3. Is there a difference between gap-filling and sentence-writing outcomes?
4. Do outcomes, preferences, and proficiency relate to each other?
5. How do learners' reactions to DDL change over a 16-week course?
6. Do learners prefer DDL on paper or on computer?

5 Procedures

Receptivity was measured with a pre/post-course questionnaire in which participants rated their expectations and satisfaction regarding DDL.

Proficiency was measured with a standardized diagnostic test[14].

All learners participated in both the hands-on and hands-off condition (5 days apart). For learning outcomes testing, 10 verb-preposition collocations for each condition[15] were selected that frequently appear in common teaching materials, DWDS, and in the novel that students read in the course. The tests included a gap-filling and a sentence-writing part. Learner responses were scored for accurate use of prepositions and noun phrase inflections. A delayed post-test was administered 4 weeks later.

During teaching interventions, learners analyzed concordances, wrote out verb-preposition-case collocations, and compared their results with a partner. In the hands-on condition, learners independently found concordances in DWDS, whereas in the hands-off condition, concordances were supplied in worksheets.

6 Results

RQ1. Learner outcomes significantly improved following DDL and, although the scores decreased a month later, the outcomes remained significantly higher than before DDL (figure 1).

RQ2. There was no difference between the conditions (figure 1). Participants had a somewhat better pretest knowledge of the items in the hands-on condition but they showed learning and attrition rates equivalent to the hands-off condition.

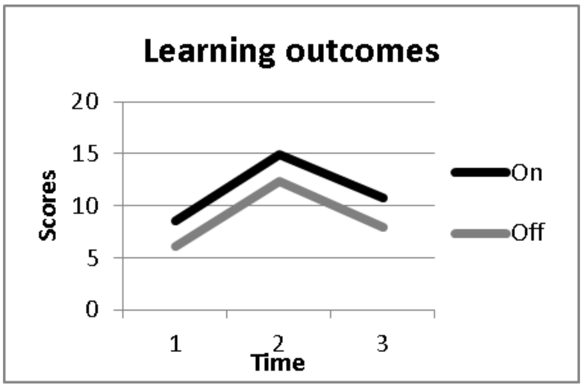


Figure 1. Learning outcomes at 3 time points for the hands-on and hands-off condition

RQ3. DDL was effective for both gap-filling and sentence-writing, although more so for the former.

RQ4. Learning outcomes correlated with learner proficiency but not receptivity. Also, learner preferences correlated with their proficiency moderately but insignificantly.

RQ5. Student post-course satisfaction with DDL activities did not significantly correlate with their pre-course expectations, although the receptivity of most (6) students increased.

RQ6. The correlation between the post-course hands-on and hands-off participants' receptivity was significant, and so was the correlation between their overall post-course receptivity and their liking of each activity type.

7 Summary and discussion

The results of this study show that both hands-on and hands-off activities are effective for L2 learning and are well received by high-intermediate learners, thus confirming Boulton's (2012) findings and extrapolating them to L2 German. Moreover, this study shows that learners retain some gains for up to a month. Unlike Boulton's, this study found that outcomes correlated with proficiency, which can be explained by the high lexico-grammatical complexity of the target structure. Finally, receptivity of most participants in this study increased, which can be explained by their high motivation. These results, although very encouraging, need to be interpreted with caution as the participant number was low, so more replication and extension studies are needed.

References

Boulton, A. 2012. "Hands-on / hands-off: Alternative approaches to data-driven learning". In J. Thomas and A. Boulton (eds.) *Input, process and product: Developments in teaching and language corpora* (pp. 152-168). Brno: Masaryk University Press.

Cobb, T. 1997. "Is there any measurable learning from hands-on concordancing?" *System* 25 (3): 301-315.

[14] www.ondaf.de

[15] Items were matched by corpus frequency across conditions.

Daskalovska, N. 2013. “Corpus-based versus traditional learning of collocations”. *Computer Assisted Language Learning*. DOI: 10.1080/09588221.2013.803982: 1-15.

Liou, H.-C., Yang, P.-C. and Chang, J.S. 2012. “Language supports for journal abstract writing across disciplines”. *Journal of Computer Assisted Learning* 28 (4): 322–335.

The construction and application of an error annotated learner translation corpus in translation classes

Jackie Xiu Yan
City University of Hong Kong
ctjackie @cityu.edu.hk

Honghua Wang
City University of Hong Kong
honwang-c @my.cityu.edu.hk

The seminal paper of Baker (1993) argued for a marriage between Corpus linguistics and Translation Studies (TS), the theoretical and applied branches of TS in particular. Baker (1993: 248) suggested that corpus-based investigations could be made to explore universal features of translation, translational norms among many other issues and emphasized “an urgent need to explore the potential for using large computerised corpora in translation studies”. Thereafter the corpus-based approach has been considered “as a viable and fruitful perspective within which translation and translating can be studied in a novel and systematic way” (Laviosa 1998: 474).

Past decade has seen many researchers applying corpus-based methods in TS including translation style (Saldanha 2011), translation units (Kenny 2011) and the fluency of translators (Uzar and Waliński 2001), to name just a few. And increased interest has been paid to corpus-based translator training since the New Millennium (Laviosa 2010). For instance, Bowker (2001) proposed a methodology for a corpus-based approach to translation evaluation.

Translation evaluation has been regarded as a “controversial issue” in TS (Colina 2010: 43). The multifaceted nature of this issue has been of concern to lots of researchers, teachers and trainers. The significance of translation evaluation, an indispensable part of translation teaching, cannot be overemphasized. However, giving feedback to translation students is a challenging task. Traditionally, translation teachers evaluated student translations based on their own experience. This practice has received much criticism due to its subjectivity and incomprehensiveness. Students are often kept in the dark as to their error patterns in general and their own problems in particular. In view of this, some scholars have introduced the corpus approach to evaluate translation and interpreting students’ performance (Bowker 2001; Uzar and Waliński 2001). The study of Leung and Yip (n.d.) in particular looked into interpreting students’ performance through the construction of a bilingual corpus in a tertiary institution in Hong Kong. With the corpus-based approach, teachers can provide students with more constructive, concrete and objective feedbacks (Bowker 2001). However, the findings from Leung and Yip (n.d.) on interpreting cannot be directly applied to written translation training classes. Therefore, despite these initial investigations, more empirical studies are called for in this regard.

Under this backdrop, the current study was designed with an aim to investigate Hong Kong translation students' performance through a corpus-based approach. An Error Annotated Learner Translation Corpus (EALTC) was built to analyze the performance of translation students at a tertiary institution in Hong Kong. Students' translation works from relevant Chinese-to-English/ English-to-Chinese translation courses were collected, which included tutorial exercises, assignments and final examination. The hard copies of student translation data were scanned and the electronic versions proofread. A textual analysis of all the source texts was conducted to identify possible problems students may encounter in translation. These problems may include, rendition errors (misunderstanding of source text, undertranslation, overtranslation and imprecise translation), linguistic errors (syntactic errors, semantic ambiguity, improper collocation, redundant words, unnecessary repetition, spelling, etc.) and miscellaneous errors etc. Error tags were added based on the textual analysis and error patterns described in Liao (2010). A corpus tool WordSmith 5.0 was used to analyze the translation data. The findings showed that the most frequently occurring error type was syntactic errors with the second being imprecise translation. For the English-to-Chinese translation, the most recurrent error type was identified as imprecise translation and the error type of misunderstanding the source text came as second. For the Chinese-to-English translation, the most frequently occuring error types were syntactic errors, semantic ambiguity, improper collocation, redundant words and unnecessary repetition. More comprehension problems were found in translating into their mother tongue and more expression problems in translating out of their mother tongue. The findings also suggested that errors caused by Cantonese[16] negative transfer were more frequently spotted in English-to-Chinese translation than in Chinese-to-English translation.

This study holds important implications for translation teaching and learning. The EALTC can provide students not only concrete evidence of grading, but more importantly, objective and constructive feedback to minimize their chances of making mistakes and enhance their translation ability. This corpus could be utilized as an important translation evaluation resources which combined three functions of evaluation, namely, diagnostic (or prognostic) function, summative function and formative function (Martínez Melis and Hurtado Albir 2001). Translation students can learn from the errors made by their peers, identify their own problems and make improvements using this corpus. Besides, the EALTC can also serve as a useful reference for teachers, trainers and researchers alike.

References

Baker, M. 1993. "Corpus linguistics and translation *studies* – implications and applications". In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology – In Honour of John Sinclair*. Amsterdam: John Benjamins.

Bowker, L. 2001. "Towards a methodology for a corpus-based approach to translation evaluation". *Meta* 46(2): 345-364.

Colina, S. 2010. "Evaluation/Assessment". In Y. Gambier and L. V. Doorslaer (eds.) *Handbook of Translation Studies (volume1)*. Amsterdam ; Philadelphia : John Benjamins Pub. Co.

Kenny, D. 2011. "Translation Units and Corpora". In A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-based translation studies: research and applications*. London: Continuum.

Laviosa, S. 1998. "The corpus-based approach: a new paradigm in translation studies". *Meta* 43(4): 474-479.

Laviosa, S. 2010. "Corpora". In Y. Gambier and L. V. Doorslaer (eds.). *Handbook of Translation Studies (volume1)*. Amsterdam; Philadelphia: John Benjamins Pub. Co.

Leung, E. and Yip, L. n.d. A Bilingual Corpus of Interpreting Students' Performance. Retrieved January 10, 2014, from http://arts.hkbu.edu.hk/~engester/main.html.

Liao, P. 2010. "An analysis of English-Chinese translation errors and its pedagogical applications". *Compilation & Translation Review* 3 (2): 101-128.

Martínez Melis, N. and Hurtado Albir, A. 2001. "Assessment in translation studies: Research needs". *Meta* 46(2): 272-287.

Saldanha, G. 2011. "Style of translation: The use of foreign words in translations by Margaret Jull Costa and Peter Bush Peter Bush". In A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-based translation studies: research and applications*. London : Continuum.

Uzar, R. and Waliński, J. 2001. "Analysing the fluency of translators". *International Journal of Corpus Linguistics*, 6(S1): 155-166.

[16] Cantonese, a dialect of Chinese, is spoken by 89.2% of population in Hong Kong. Chinese and English are the official languages there. (http://www.gov.hk/en/about/abouthk/facts.htm)

The BUiD Arab Learner Corpus: Explaining second language writing systems within a markedness framework

Yasemin Yildiz
The British University in Dubai
yasemin.yildiz@buid.ac.ae

1 Introduction

This paper has a two-fold goal. The first goal is to contribute to the literature of Second Language Writing Systems (L2WS) by focusing on the British University in Dubai Arab Learner Corpus (BALC). The second goal is to demonstrate the close relationship between phonology and orthography in L2WS and critically address the issue of reform in a script. Unlike previous studies which provide a holistic and descriptive analysis of all possible spelling errors of Arabic-speaking learners of English (e.g. Randall and Groom, 2009; Haggan, 1991; Hassan, 2010) this study is different in two kinds: 1) As a first attempt BALC will be interpreted within a markedness linguistic framework 2) Particular emphasis will be given to the erroneous spelling forms which appear in lexical items with complex onset and coda clusters at phonological level only (e.g. *stamped* [stæmpt]).

The existing theories explaining L2WS range from *Contrastive Analysis Hypothesis* (Lado 1957; herafter CAH), which compared the areas where the L2 differed from the L1 to determine what would be difficult for the learner, to *Error Analysis* (Corder 1967; hereafter EA), which advocates looking only at the developing grammar of the learner to ascertain where difficulties exist. Moreover, although both of these theories may be able to foresee or account for the linguistic difficulties of the learners, they exhibit two shortcomings. First, CAH relies only on native language transfer. Second, Error Analysis misses the relationship between L1 transfer and universal processes. As an alternative model, this study attempts to explain how the *Markedness* framework, can also be a useful tool in modelling the first language and universal constraints in L2WS. In fact, according to Spolsky (1989) the *markedness condition* is necessary as a linguistic ground for language learning.

2 Theoretical framework

Trubetzkoy and Jakobson were the first linguists to introduce the idea of ‘markedness’ in the 1930s and is treated as a language-particular phenomenon. Trubetzkoy approached the term markedness within a descriptive framework and it was initially confined to phonetics. Jakobson (1968), however, approached the term markedness within the perspective of language acquisition. The underlying principle of Jakobson’s theory is that there is a universal order of acquisition, largely based on phonological oppositions and phonetic properties of segments. Based on the structural contrasts in his theory, Jakobson suggested that the unmarked forms would be the earliest acquired and would also occur in all the world’s languages.

3 The study

The BUiD Arab Learner Corpus (BALC) consists of 1,865 texts written by either first year university students or secondary school students (year/grade 12 – the last year of schooling). It comprises 287,227 word tokens and 20,275 word types. The texts themselves fall into three types: texts collected by MEd students in secondary schools, retired first year university test essays, and texts sourced from the Common Educational Proficiency Assessment (CEPA) examinations (All school students in the United Arab Emirates need to take CEPA as a university entrance exam). The scripts were all hand written and then converted into text files for incorporation into the corpus.

4 Instrumentation and procedure

The misspelling data which exhibit consonant clusters will be identified and categorized by using the Wmatrix3 program (Rayson 2003, Rayson 2005), which is an online integrated corpus linguistic software environment in which texts can be loaded and analyzed for word frequency profiles and concordances, annotated in terms of part-of-speech (using the well-known CLAWS tagger, see Garside et al. 1997) and word-sense (semantic content and word sense tagger). The semantic content component, named the UCREL Semantic Analysis System (or USAS), contains a multi-tier structure with 21 major discourse categories.

These 21 categories are further refined and categorized. A particular refinement within the 'Z' category identifies the unmatched items (or those items not recognized by the system) and is categorized as 'Z99'. The data elicitation will be sourced from the Z99 category, as this category can identify all the spelling errors and provide the frequency distribution. The quantitative analysis will be conducted by using the findings from the Z99 category. A further qualitative analysis will be conducted within the markedness framework.

5 Research questions

This study takes up the following three questions for investigation:

1) What modification strategies do the learners use in the production of consonant clusters?
2) To what extent are L2 syllables constrained by allowable L1 syllable structure and to what extent do universal principles apply or even prevail?
3) What is the role of markedness for the production of consonant clusters?

References

Corder, S. P. 1967. “The Significance of Learners` Errors”. *International Review of Applied Linguistics* 5: 161-169.

Garside, R., Leech, G. and McEnery, T. 1997. (eds). *The Computational Analysis of English*. London: Longman.

Gnanadesikan, A. E. 2004. “Markedness and faithfulness constraints in child phonology”. In R. Kager, J. Pater and W. Zonneveld (eds.) *Constraints in Phonological Acquisition*. Cambridge: Cambridge University Press. pp. 73–108. [ROA-76]

Haggan, M. 1991. “Spelling errors in native Arabic-speaking English majors. A comparison between remedial students and fourth year students”. *System* 19(1): 45-61.

Lado, R. 1957. *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press: Ann Arbor.

Spolsky, B. 1989. *Conditions for Second Language Learning: Introduction to a General Theory*. Oxford University Press.

Trubetzkoy, N. 1939. Grundzüge der Phonologie (Principles of Phonology). *Travaux du cercle linguistique de Prague* 7.

Randall, M. 2007. *Memory, psychology and second language learning*. Philadelphia: Benjamins Publishing Company.

Randall, M. and Groom, N. 2009. “Introducing the BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English spelling”. In M. Mahlberg, V. González-Díaz and C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference CL2009, University of Liverpool, UK, 20-23 July 2009*.

Rayson, P. 2003. Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison. Ph.D. thesis, Lancaster University. Available online at http://ucrel.lancs.ac.uk/people/paul/publications/phd2003.pdf

Rayson, P. 2005. *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster University. Available online at http://www.comp.lancs.ac.uk/ucrel/wmatrix/

Yildiz, Y. and Ozek, Y. 2009. The Role of Markedness in Vocabulary Learning. In *Proceeding of the International Conference of Technology, Education and Development (ICERI 2009).*

An investigation into the use of dative alternation by L1 and Arab L2 users of English

Abdalkarim Zawawi
Lancaster University
a.zawawi@lancaster.ac.uk

1 Introduction

This paper evaluates the use of the English dative alternation by native English speakers and Arab learners of English as a foreign language (EFL). It is based on corpus data where English ditransitive verbs, that may or may not take a prepositional phrase (PP) as their indirect objects are explored. The dative construction refers to speakers' grammatical choice of using a PP with the preposition 'to' or 'for' as an indirect object to a noun phrase (NP). The alternative construction to the dative case is to use an NP which consists of two objects as a complement to one ditranstive verb like in: 'Michael gave Maria a book' versus 'Michael gave a book to Maria'.

2 The present study: The use of dative alternation

The study investigates the extent to which Arab EFL learners' choices differ from the native English norm in using the dative case. A sub-corpus of 208,645 words of conversation and interviews from the spoken component of the British National Corpus (BNC) was compared to a sub-corpus of 154,754 words from the Arabic first language (L1) component of the Longman learner corpus.[17]

These naturally occurring native and learner data allowed the identification of the typical dative alternation use by automatically retrieving occurrences of a wide range of ditransitive verbs such as (give, offer, post, sell, show, throw, send, explain, design, open, ask, buy, contribute, refuse, offer). These verbs accept (a) mainly the dative case (e.g., buy, explain), (b) mainly the double object construction (e.g., wish, refuse), or (c) both constructions (e.g., give, offer) as their complements (Berk, 1999). Importantly, some of these verbs behave differently in Arabic grammar. For example, while it is grammatical to use the double object to complement a ditransitive verb like 'buy' in English, Arabic grammar does not allow a double object complement with the verb 'ishtara' (to buy).

[17] Unfortunately, a spoken Arab EFL learner corpus seems to be unavailable to date.

3 Results

First results indicate that English ditransitive verbs are not identical in their degree of reluctance in accepting the {NP, NP} or {NP, PP} object construction. The verb 'give' showed a considerably higher reluctance to accept the dative case than it did in a similar native English corpus-based study by Gries' (2005) on the dative alternation case.

The verb 'send' showed a similar pattern in the L1 and second language (L2) learner corpora in that it is used considerably higher with a double object rather than the dative case construction. Learners seem to overuse the dative case of 'send' – which can probably be attributed to the fact that the dative case is preferred in Arabic as a completment to the ditransitive verb '2rsala' (to send).

The verb 'explain' showed seemed to have a similar behaviour of resorting to the dative case rather than the double-object construction in both corpora. The reason why Arab learners did not confuse the double object with the dative case may be that Arabic grammar does not allow the double object construction as a complement to the verb 'explain'. Noticeably, most of the individual Arab learners who used the ditransitive form of the verb 'explain' were at a proficient stage of learning English.

An interesting finding is that Arab learners appear to overuse the preferred, 'unmarked' alternative in Arabic grammar, while they underuse the 'unusual' marked ones. This behaviour is apparent in overusing the dative case where Arabic does not allow a double object complement such as the verbs 'send', 'buy' and 'design'.

4 Conclusion and future steps

This paper presents an argument that Arab learners deviate in their use of dative alternation from the English norm. It can be argued that they tend to overgeneralise what is the preferred and grammatical norm for dative constructions in their L1 Arabic to their L2 English – a form of negative transfer which refers to erroneous usage that results from a given language interference (Gilquin et al., 2008).

By exploring spoken L1 and L2 data future studies can widen the scope of research on linguistic interference to include more grammatical categories and the extent to which learners rely on their L1 to make their L2 grammatical choices.

I will further investigate Arab learners' English spoken interactions to better understand the role of syntactic priming, i.e: "… the tendency for a speaker to produce a syntactic structure that occurred in the recent discourse rather than an alternative structure" (Kim & McDonough, 2008).

References

Berk, L. (1999). *English Syntax from Word to Discourse.* New York/Oxford: Oxford University Press.

Gilquin, G., Papp, S., & Diez-Bedmar, M. (2008). *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(issue number?), 365–399. Doi available for this one?

Kim, Y. & McDonough, K. (2008). Learners' production of passives during syntactic priming activities. *Applied Linguistics,* 29(issue number), 149–154. doi:10.1093/applin/ amn004

Poster presentations

Combining two corpus tools for easier and effective DDL

Kiyomi Chujo
Nihon University
chuujou.kiyomi@nihon-u.ac.jp

Laurence Anthony
Waseda University
anthony@waseda.jp

Shiro Akasegawa
Lago Institute of Language
lagonist@gmail.com

Kathryn Oghigian
Waseda University
oghigian@gmail.com

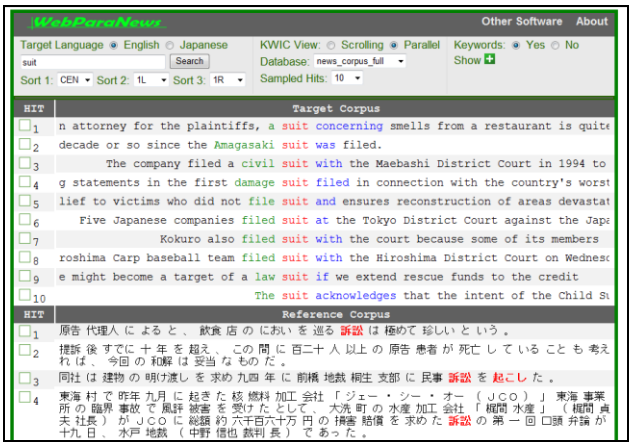
1 Two web-based parallel corpus tools

This poster demonstrates combining two newly developed web corpus tools to promote more effective DDL in the foreign language classroom.

We developed a KWIC concordance tool, WebParaNews (see Figure 1), and a lexical profiling tool, the LagoWordProfiler (LWP) for ParaNews (see Figure 2). Both are freeware and are based on the same parallel corpus which consists of newspaper texts in English along with their aligned translations in Japanese. [18] Although typical applications of a parallel corpus include translator training, bilingual lexicography, and machine translation (O'Keeffe, et al. 2007), we use it for L2 classroom applications of DDL.

The merits of the KWIC format are drawing the learners' attention to the target items,finding patterns in surrounding words, and showing multiple examples and contexts to make generalizations from sets of evidence (Murphy 1996; Barlow 2004; Mishan 2004; Boulton 2009). Even with truncated but color-coded concordance lines, the KWIC can "reduc[e] the information load – especially important perhaps for lower levels" (Boulton 2009) and it helps to highlight the target word usage visually so learners can form hypothesis inductively. However, there are limits to finding patterns in word usage from concordance lines, because extensive searching is often required to find a comprehensive analysis of a word behaviour.

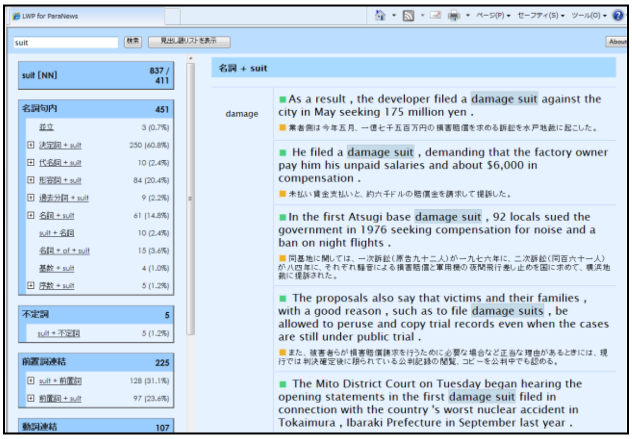
On the other hand, lexical profiling systems such as LWP for ParaNews are tools which show a comprehensive analysis of how words behave. LWP is a browsing system that provides corpus-derived summaries of collocation/colligation information by entry word. For example, in Figure 2, the summary of collocation/colligation usage of the noun *suit* is summarized into four grammar categories, i.e., noun phrase, infinitive, preposition concatenation, and verb concatenation, with their respective frequencies; and each category has subcategories, for example, noun phrase has subcategories such as determiner + *suit*, noun + *suit*, adjective + *suit*, and a user can choose the particular subcategory and view the sentences included in that subcategory. Thus, after learners conduct the inductive DDL activities through a KWIC presentation, they can attempt the deductive DDL activities using the second lexical profiling tool by checking the rules in the corpus-derived summary and in actual example sentences to test their hypotheses about the target items.



HIT	Target Corpus
□ 1	n attorney for the plaintiffs, a suit concerning smells from a restaurant is quite
□ 2	decade or so since the Amagasaki suit was filed.
□ 3	The company filed a civil suit with the Maebashi District Court in 1994 to
□ 4	g statements in the first damage suit filed in connection with the country's worst
□ 5	lief to victims who did not file suit and ensures reconstruction of areas devastat
□ 6	Five Japanese companies filed suit at the Tokyo District Court against the Japa
□ 7	Kokuro also filed suit with the court because some of its members
□ 8	roshima Carp baseball team filed suit with the Hiroshima District Court on Wednesc
□ 9	e might become a target of a law suit if we extend rescue funds to the credit
□ 10	The suit acknowledges that the intent of the Child Su

HIT	Reference Corpus
□ 1	原告 代理人 に よる と 、 飲食 店 の におい を 巡る 訴訟 は 極めて 珍しい と いう 。
□ 2	提訴 後 すでに 十 年 を 超え 、 この 間 に 百二十 人 以上 の 原告 患者 が 死亡 し て いる こと も 考え れ ば 、 今回 の 和解 は 妥当 な もの だ 。
□ 3	同社 は 建物 の 明け渡し を 求め 九四 年 に 前橋 地裁 桐生 支部 に 民事 訴訟 を 起こし た 。
□ 4	東海 村 で 昨年 九月 に 起き た 核 燃料 加工 会社 「 ジェー ・ シー ・ オー （ ＪＣＯ ） 」 東海 事業 所 の 臨界 事故 で 風評 被害 を 受け た として 、 大洗 町 の 水産 加工 会社 「 梶間 水産 」 （ 梶間 貞 夫 社長 ） が ＪＣＯ に 総額 約 六千百六十万 円 の 損害 賠償 を 求め た 訴訟 の 第 一 回 口頭 弁論 が 十九 日 、 水戸 地裁 （ 中野 信也 裁判 長 ） で あっ た 。

Figure 1. WebParaNews



suit [NN]	837 / 411
名詞句内	451
並立	3 (0.7%)
⊞ 決定詞 + suit	250 (60.8%)
⊞ 代名詞 + suit	10 (2.4%)
⊞ 形容詞 + suit	84 (20.4%)
⊞ 過去分詞 + suit	9 (2.2%)
⊞ 名詞 + suit	61 (14.8%)
suit + 名詞	10 (2.4%)
名詞 + of + suit	15 (3.6%)
基数 + suit	4 (1.0%)
⊞ 序数 + suit	5 (1.2%)
不定詞	5
suit + 不定詞	5 (1.2%)
前置詞連結	225
⊞ suit + 前置詞	128 (31.1%)
⊞ 前置詞 + suit	97 (23.6%)
動詞連結	107

名詞 + suit

damage

■ As a result , the developer filed a damage suit against the city in May seeking 175 million yen .
■ 業者側は今年五月、一億七千五百万円の損害賠償を求める訴訟を水戸地裁に起こした。

■ He filed a damage suit , demanding that the factory owner pay him his unpaid salaries and about $6,000 in compensation .
■ 未払い賃金支払いと、約六千ドルの賠償金を請求して提訴した。

■ In the first Atsugi base damage suit , 92 locals sued the government in 1976 seeking compensation for noise and a ban on night flights .
■ 同基地に関しては、一次訴訟（原告九十二人）が一九七六年に、二次訴訟（同百六十一人）が八四年に、それぞれ騒音による損害賠償と軍用機の夜間飛行差し止めを国に求めて、横浜地裁に提訴された。

■ The proposals also say that victims and their families , with a good reason , such as to file damage suits , be allowed to peruse and copy trial records even when the cases are still under public trial .
■ また、被害者らが損害賠償請求を行うために必要な場合など正当な理由があるときには、現行では判決確定後に限られている公判記録の閲覧、コピーを公判中でも認める。

■ The Mito District Court on Tuesday began hearing the opening statements in the first damage suit filed in connection with the country 's worst nuclear accident in Tokaimura , Ibaraki Prefecture in September last year .

Figure 2. LWP for ParaNews

2 Example task

In this type of combined-resource DDL, learners first use WebParaNews to find general patterns and tendencies in the usage of the target items. For example, they search *suit* in the newspaper corpus, as shown in Figure 1. They learn that *suit* is used in NPs such as "determiner + *suit*" as in *a suit*, or *the suit*; "determiner + noun + *suit*" such as *the Amagasaki suit*, or *a law suit*, "determiner + adjective + *suit*" such as *a civil suit*, and, as the objects of a verb such as *[subject] filed suit*.

[18] The Japanese-English News Article Alignment Data, comprised of 180,000 sentence pairs from *The Yomiuri Shimbun* (6.1 million Japanese morphemes) and *The Daily Yomiuri* (4.9 million English words), and automatically aligned.

Next the learners can look at *suit* (NN) in LWP and view examples of DET + *suit*, N + *suit*, Adj + *suit*, and verb + *suit*. Numerous examples of these collocation/colligation patterns are provided with Japanese translations.

3 Case study

We are conducting a case study combining these two tools in the spring semester of 2014 and will be collecting student feedback to add modifications to improve LWP for ParaNews.[19]

We believe using different types of information from two corpus tools can provide useful insights to learners. Firstly, using the information from the KWIC presentation allows learners to discover and form their own hypotheses about the language, and secondly the information from the profiling summary supports hypothesis testing. We hope to determine if this combined-resource approach may be more helpful for recall and long-term retention than traditional DDL approaches.

References

Akasegawa, S. 2014. LagoWordProfiler. Shiga, Japan: Lago Institute of Language. Available online at http://www.lagoinst.com/LPN/LWPforParaNews.html

Anthony, L. 2012. WebParaNews. Tokyo, Japan: Waseda University. Available online at http://www.antlab.sci.waseda.ac.jp/

Anthony, L., Chujo K. and Oghigian, K. 2011. “A novel, web-based, parallel concordancer for use in the ESL/EFL classroom”. In J. Newman, H. Baayen and S. Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation.* Amsterdam/New York: Rodopi Press.

Barlow, M. 2004. “Software for corpus access and analysis”. In J. Sinclair (ed.) *How to use corpora in language teaching*. Amsterdam: John Benjamins Publishing Co.

Boulton, A. 2009. “Testing the limits of data-driven learning: language proficiency and training”. *ReCALL* 21(1): 37-54.

Chujo, K., Akasegawa, S., Nishigaki, C., Yokota, K. and Hasegawa, S. 2012. “LagoWordProfiler ni yoru Eigo graded reader corpus no collocation/colligation hindo bunseki [LagoWordProfiler frequency analysis of collocations and colligations in an English graded reader corpus]”. *Journal of the College of Industrial Technology*, Nihon University 45: 1-17.

Murphy, B, 1996. “Computer corpora and vocabulary study”. *Language Learning Journal* 14: 53-57.

Mishan, F. 2004. “Authenticating corpora for language learning: a problem and its resolution”. *ELT Journal* 58(3): 219-227.

O’Keefe, A., McCarthy, M. and Carter, R. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge, England: Cambridge University Press.

Utiyama, M. and Isahara, H. 2003. “Reliable measures for aligning Japanese-English news articles and sentences”. In E. Hinrichs and D. Roth (eds.) *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo: The Sapporo Convention Center.

[19] In Figure 2, the collocation/colligation terms are written in Japanese because the students are beginner level and they prefer L1 terms. We have an English version as well.

A corpus-based analysis of English language learning needs in Nigeria

Alexandra Esimaje
Benson Idahosa University
aesimaje@biu.edu.ng

Ulrike Gut
University of Münster
gut@wwu.de

1 English in Nigeria

Nigeria is a highly multilingual country whose roughly 140 million inhabitants speak an estimated 500 languages (Grimes 2000). The major indigenous Nigerian languages include Hausa, Yoruba and Igbo.

Due to Nigeria's colonial past, English has a geographical spread throughout the country and is spoken by an estimated 20% of the population (Jowitt 1997). English plays a key role in the Nigerian education system both as the language of instruction as well as the language of textbooks, students' written assignments and examinations. English is also used in contexts such as government, media, literature, business, commerce and as a lingua franca among the educated élite.

Some syntactic features of Nigerian English such as the omission of articles, nonstandard plural formation of nouns and mismatches in subject-verb concord have been described in Alo and Mesthrie (2008), but corpus-based descriptions are rare so far (but see Gut and Fuchs 2013).

This paper reports on preliminary results of an ongoing corpus-based study of English learner language in Nigeria and Cameroon. The aim of the project is to describe syntactic features of the language productions of secondary school and university students in both countries and to identify their particular learning needs.

2 Corpus compilation

For this purpose, a learner corpus of written language productions by Nigerian and Cameroonian learners is being compiled. The final corpus will comprise a total of 300,000 words and include class-based essays, letters written by students as a class activity and end of term examinations, covering the period from 2010 to 2013.

So far, data has been collected from 411 Nigerian secondary school students aged 16-17, who produced 160,692 words, and from 268 final-year university students aged 18-23, who produced 139,949 words. The students are of different ethnic backgrounds and come from six different geopolitical zones of Nigeria.

This poster first reports on the corpus compilation process, which is carried out using Pacx[20] (*P*latform for *A*nnotated *C*orpora in *X*ML), a recently developed platform that supports the compilation, archiving, annotation, distribution and searching of large amounts of language data (see Gut 2011). It is based on the Eclipse platform[21] and extends it by the addition of several plug-ins: the XML editor Vex, the image viewer QuickImage and the client for the version control system Subversive.

For the annotation of written data, the transcriber marks a word and selects a tag from a predefined list. We annotated all spelling errors and all syntactic structures that do not conform to British or American standard grammar rules.

3 Preliminary results: Syntactic constructions in Nigerian learner language

First results are being presented that are based on an analysis of a sub-corpus of approximately 50,000 words comprising 76 examination papers written by 66 Nigerian university students with different ethnic backgrounds.

The corpus was first searched for the syntactic features of Nigerian English described by Alo and Mesthrie (2008). Results show that Nigerian learners omit definite and indefinite articles in noun phrases in about 2% of all cases. Both plural marking of non-count nouns and mismatches in subject-verb concord occur with a rate of 1%. Further, we report on some specific non-standard uses of the modal auxiliaries *would* and *will*, of the progressive and of the use of the reflexive pronouns.

References

Alo, M. and Mesthrie, R. 2008. "Nigerian English: morphology and syntax". In R. Mesthrie (ed.) *Varieties of English. Africa, South and Southeast Asia*. Berlin: Mouton de Gruyter.

Grimes, Barbara 2000 *Ethnologue. Languages of the World*. Dallas: SIL International.

Gut U. 2011. "Language documentation and archiving with Pacx, an XML-based tool for corpus creation and management". In N. David (ed.) Workshop on Language Documentation and Archiving, London.

Gut U. and Fuchs, R. 2013. „*Progressive aspect in Nigerian English". Journal of English Linguistics* 41(3): 243-267.

Jowitt, D. 1997. "Nigeria's national language question: Choices and constraints". In A. Bamgbose, A. Banjo and A. Thomas (eds.) *New Englishes: A West African perspective*. Trenton, NJ: Africa World Press.

[20] http://pacx.sourceforge.net/

[21] http://www.eclipse.org

Lexico-phraseology profile in English logistics corpus: an investigation into register features in written academic logistics language

Yuying Hu
Nottingham University Ningbo China & Guang Dong University of Foreign Studies
katehyy@163.com

John McKenny
Faculty of Education, British University in Dubai
john.mckenny@buid.ac.ae

The present research aims at exploring the lexico-phraseological profiles of an English logistics corpus focusing on written academic data. By doing so, it is expected to reveal features of registers manifested by linguistic features in forms of single word levels (i.e. words and keywords) and combined-word units of both continuous and non-continuous phrasal levels (i.e. lexical bundles, collocational frameworks and concgrams). This research will bridge the gap between corpus linguistics studies and lexico-phraseological profiles across English logistics corpus data, particularly with respect to pervasive language use. Following the three-step research framework proposed by Biber and Conrad (2009), both quantitative and qualitative methods are used to present a full picture of vocabulary and phraseology aided by a corpus-driven approach. More precisely, words, keywords, lexical bundles, collocational frameworks and concgrams construct the lexico-phraseological profiles in the target corpus, which consists of 4 corpus datasets including textbooks, journal articles, theses and monographs. The ideal size of corpus is one million words with a focus on written logistics materials. In accordance with Biber and Conrad (2009), register analyses in the present study centre on three core components: the situational/communicative contexts, the linguistics features, and the functional relationships between the first two components. This is because these three perspectives are helpful for an observation of the linguistic features as well as their functions in situational contexts. Because the research at present stage is just a pilot study focusing on the comparison between logistics textbook corpus (238,664 tokens) and FLOB (1 million tokens), partial results of the pilot study reveal that patterns of language use are different across logistics textbooks, and that the observed variation is related to the situational contexts and the functions of language use in logistics textbooks and FLOB texts. That is to say, the results of the pilot study indicate the various characterized linguistic features are not only content-related (Rǒmer 2009), but also function-related (Grabowski 2013). The findings of the research could be beneficial for the teaching practice of English for Specific Purpose (ESP) regarding vocabulary teaching, writing tutoring as well as optimizing syllabus designs. Moreover, findings in the present study would also be helpful data for dictionary compilation and writing instructions for logistics researchers and professionals. Lastly, the comprehensive method of register study in this research could be transferrable to similar specialized corpora studies in other disciplines such as law, science, engineering and agriculture.

References

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

Grabowski, L. 2013. “Register Variation across English Pharmaceutical Texts: A Corpus-Driven Study of Keywords, Lexical Bundles and Phrase Frames in Patient Information Leaflets and Summaries of Product Characteristics”. *Procedia-Social and Behaviour Science* 95: 391-401.

Rǒmer, U. 2009. “English in Academia: Does Nativeness Matter”? *Anglistik: International Journal of English Studies* 20 (2): 89-100.

Selected syntactic features of the Czech learner corpus of spoken English

Šárka Ježková
University of Pardubice
Czech Republic
sarka.jezkova@upce.cz

The poster presents one part of a bigger multidimentional project called “Aspects of English Language Acquisition of Czech Students on the Onset of Teacher Education”, which aims at identifying external and internal factors influencing the process of learning English as a foreign language by Czech learners, particularly the achieved level of communicative competence in speaking. This particular part of the research is focused on the analysis of Czech university student’ performances (including both monologue and dialogue), with respect to the specific features of spoken discourse.

Grammar of speech has been a subject of interest for a couple of decades and some authors argue that writing and speech are two different systems (Carter and McCarthy 1995). For second / foreign language acquisition process such differences are quite important because they should be reflected in productive skills of speaking and writing. However, in the Czech educational background, as in many other countries English language learners are usually instructed on the basis of written discourse (McCarthy and Carter 2001). Thus the research has been motivated by the question if the students acquire the syntactic and discourse structures which are observable in native speakers’ discourse (Biber 1988, Biber et al. 1999).

Besides others, the team set up the objectives: to create a corpus of learner English of spoken communication, to make an analysis of selected grammatical, discourse and pronunciation features with conclusions for second language acquisition processes, to obtain and to analyze quantitative and qualitative data regarding students’ individual learning histories.

The corpus now comprises recordings of first year students from three Czech universities which were transcribed by students themselves and checked by the research team later. Even though the concepts of English as lingua franca and of learners’ English language significantly differ, some of the tendencies in the usage of language can be similar (Seidlhofer 2011). Thus, based on the previous studies of English native speakers’ conversation (Biber et al. 1999) and non-native speakers’ discourse (Mauranen 2012, Götz and Schilk 2011, Housen 2002, Meunier 2002), certain grammatical features have been chosen, explored in learners’ corpus and considered within the concept of English as lingua franca. It has been revealed that similar processes like simplification, diversification, regularization and productivity are also observable in the learners’ English language.

All the chosen grammatical structures (e.g. verb forms: concord, tense, aspect, irregular forms; nominal forms: singular vs. plural, nominative vs. accusative; relative pronouns, assertive vs. non-assertive pronouns, word classes and word formation, double negation, double comparison, etc.) are analysed with respect to the second language acquisition processes and considered within the background of systemic differences between English and Czech languages.

References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

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

Carter, R. and McCarthy, M. 1995. “Grammar and the Spoken Language.” *Applied Linguistics* 16 (2): 141–158.

Götz, S. and Schilk, M. 2011. “Formulaic sequences in spoken ENL, ESL, EFL: Focus on British English, Indian English and learner English on advanced German learners.” In Mukherjee, J. and Hundt, M. (eds.) *Exploring Second-Language Varieties of English and Learner Englishes*. 79–100. Amsterdam: John Benjamins.

Housen, A. 2002. “A Corpus-based Study of the L2-acquisition of the English Verb System.” In Granger, S., Hung, J. and Petch-Tyson, S. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. 77–116. Amsterdam: John Benjamins.

Mauranen, A. 2012. *Exploring ELF. Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

McCarthy, M. and Carter, R. 2001. “Ten criteria for a spoken grammar.” In Hinkel E. and Fotos, S. *New Perspectives on Grammar Teaching in Second Language Classrooms*. 51-75. Mahwah, N.J.: Lawrence Erlbaum Associates.

Meunier, F. 2002. “The Pedagogical Value of Native and Learner Corpora in EFL Grammar Teaching.” In Granger, S., Hung, J. and Petch-Tyson, S. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. 119–41. Amsterdam: John Benjamins.

Seidlhofer, B. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Creation of CATE and a Corpus-based Study on L3 Acquisition of the Spanish Past Tense

Hui-Chuan Lu
NCKU, Taiwan
huichuanlu1@gmail.com

An-Chung Cheng
University of Toledo, USA
accheng99@gmail.com

Sheng-Yun Hung
Research Assistant,
NCKU, TAIWAN
yunbe0811@gmail.com

1 Introduction

This presentation features the creation of a learners' corpus in Spanish and the study of written and oral production of Taiwanese learners of Spanish as a third language (L3), whose first language (L1) is Chinese and second language (L2) is English.

2 Corpus-based research on third language acquisition

Among different types of corpora, the construction of learners' corpus benefits research in language acquisition (e.g., Granger, 2003, 2009; Myles, 2005). 25 of the 360 exiting corpora[22] are learners' corpora. 92% (23/25) of them are related to English, 80% (20/25) are written corpora and 12% (3/25) are oral ones, while only two corpora are related to Spanish. This corpus CATE (Corpus de Aprendices Taiwaneses de Español) fills the gap in the field.

The purpose of the CATE was to construct a learners' corpus of Chinese-speaking learners of L3 Spanish in order to inform teaching and advance the research on multi-language acquisition.

Previous studies show that there were differences between learners' oral and written productions (Dickerson & Dickerson, 1977; Tarone, 1979; Hsieh, 2005; Larsen-Freeman, 2006; Ellis, 2008). Cortés (2002) and Blanco Pena (2013) indicate that similar error patterns of written production can be observed in the oral development but occur more frequently in oral development of acquisition process. Thus, the data of the CATE includes both written and oral productions.

Based on the data from CATE, this study focuses on the acquisition of the Spanish past tense under the framework of Lexical Aspectual Hypothesis (Andersen 1986, 1991). The research questions are: (1) Do written and oral productions of Chinese-speaking learners of L3 Spanish follow the same developmental stages with respect to the Lexical Aspectual Hypothesis (LAH)? (2) Do they show the same developmental pattern as learners with other language backgrounds in previous studies?

3 CATE: a L3 learners' corpus

CATE includeds two sub-corpora: CEATE (Corpus Escrito de Aprendices Taiwaneses de Español) and COATE (Corpus Oral de Aprendices Taiwaneses de Español). The construction phases involve (1) data collection, (2) error correction and annotation, (3) programming and design of user interface. The written corpus, CEATE, includes 2,425 texts and 446,694 words with participants from 15 universities in Taiwan from 2005 to 2011. Between 2009 and 2013, a 45-minute Wisconsin Placement Test was conducted to assess participants' Spanish proficiency. CEATE IIB (2010-2011) includes elicited narration of a fairy tale "Caperucita Roja/Little Red Riding Hood". The oral corpus, COATE (2013), includes 846-minute orally recorded data of 68 participants from four universities with the same narration topic of CEATE IIB. EXMARaLDA Partitur-Editor[23] was used to facilitate the transcription of oral data. In the annotation phase, Spanish native speakers corrected the errors of learner productions and the research team annotated lexical aspects of verbs in the corpora. Then, the FreeLing was utilized to POS-tag the learners and revised data. Finally, NCKU CSIE WMMKS Laboratory provided professional technical support for programming and designing the user interface and search functions with MySQL and Perl. The website now is open for searchers[24].

4 Language acquisition study

With the help of CEATE and COATE, we investigated the past test usage of Taiwanese learners of Spanish based on Lexical Aspectual Hypothesis proposed by Anderson (1986, 1991).

In total, we have analyzed 132 written texts from CEATE 2010-2011 and 65 oral texts from COATE 2013. The written data includes 103 texts from lower intermediate level, and 29 from intermediate level. The oral one covers 46 and 19 texts from the above mentioned levels.

The preliminary findings show that the acquisition of the Spanish past tense in written production was earlier than that in oral production. In addition, the accuracy rate of the preterite form was higher than that of the imperfect form in the oral and written productions of lower intermediate learners. Third, they also shared the similar tendency

[22] Lee, D. 2010. http://www.uow.edu.au/~dlee/CBLLinks.htm [2014-1-12]

[23] http://www.exmaralda.org/

[24] http://corpora.flld.ncku.edu.tw/.

of developmental stages with respect to LAH in comparison with English-speaking learners. In terms of the preterite form, learners used accomplishment and achievement verbs more correctly than activity and stative verbs. The usage of the imperfect form followed this sequence: stative>activity>accomplishment+achievement, which corroborated with the pattern of English-speaking learners of Spanish (e.g., Anderson, 1986 & 1991).

In the process of multi-language acquisition, learner's knowledge of their first and second language could play different roles to certain extents. Future research will examine possible linguistic factors in Chinese (L1) and English (L2) that might affect the learning of the Spanish past tense for L3 learners based on contrastive analysis.

References

Andersen, R. W. 1986. "El desarrollo de la morfología verbal en el español como segundo idioma". In Meisel, J. M. (Ed.), *Adquisición del lenguaje – Acquisição da linguagem* 115–138. Frankfurt: Klaus-Dieter Vervuert Verlag.

Andersen, R. 1991. "Developmental sequences: The emergence of aspect marking in second language acquisition". In Huebner, T. & Ferguson, C. A. (Eds.), *Crosscurrents in second language acquisition and linguistic theories* 305–324. Amsterdam: John Benjamins.

Blanco Pena, J. M. 2013. "Escollos lingüísticos de los principiantes chinos de español como lengua extranjera: Causas y sugerencias pedagógica" *Hispania*: 96(1): 97-109.

Cortés, M. 2002. "Dificultades linguísticas de los estudiantes chinos en el aprendizaje del ELE". Carabela 52: 77-98.

Dickerson, L. & Dickerson, W. 1977. "Interlanguage phonology: current research and future directions". In Corder, P. & Roulet, E. (Eds.), *Interlanguages, Prdgins and their Relation to Second Language Pedagogy* 18-29. Librairie Droz, Neufchâtel: Faculté des Lettres and Genève.

Ellis, R. 2008. "Interlanguage variability in narrative discourse: Style shifting in the use of the past tense". *Studies in Second Language Acquisition* 9 (1): 1-19.

Granger, S. 2003. "The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research". *Tesol Quarterly* 37 (3): 538-546.

Granger, S. 2009. "The contribution of learner corpora to second language acquisition and foreign language teaching". *Corpora and Language Teaching* 33: 13-2.

Hsieh, H.-H. 2005. A study of communication strategies in Taiwan EFL college learners' spoken language and written language. National Chengchi University. MA thesis.

Larsen-Freeman, D. 2006. "The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English". *Applied Linguistics* 27 (4): 590-619.

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

Tarone, E. 1979. "Interlanguage as chameleon". *Language Learning* 29: 181-191.

Content words and function words in nuclear science English corpus

Daehyeon Nam
Ulsan National Institute of
Science and Technology
dnam@unist.ac.kr

1 Introduction

Drawing upon the recent attention to EAP (English for Academic Purposes) and ESP (English for Specific Purposes) research and education using corpora (Belcher, Ed, 2009; Gledhill, 2000; Hyland, 2011; Parkinson, 2013, Römer & Wulff, 2010), the current research explores the lexical and grammar patterns of English for nuclear science. Nuclear science English, one of the English styles used in a specialized discipline, may render idiosyncratic characteristics of its own discipline. In addition to the linguistic features of the language, in S. Korea, the discourse around the discipline has been a crucial news topic both politically and diplomatically. In this regard, research on the nuclear science English is worth linguistic and discourse analysis research.

As a preliminary analysis towards a full-scale study of nuclear science English, the current study examines the characteristics of the nuclear science English in a research journal. From the keyword analysis of the nuclear engineering English corpus, it is expected to construct a set of disciplinary characteristics of the nuclear engineering English. Specifically, a basic frequency list pulled out of a small corpus of research articles, notable keywords compared against general English word usage, and collocational combinations which may suggest how certain words are idiosyncratically used in the nuclear science ESP community.

2 Research methods

To examine the idiosyncratic characteristics of nuclear science English, it is necessary to compare the word usages in the nuclear science English corpus and a general English corpus: one being the collection of nuclear science English texts and the other, a large reference corpus, respectively. To analyze and identify any salient characteristics of nuclear science English, simple steps of corpus linguistic approach was utilized. First, it is necessary to compare the frequency lists of nuclear science English corpus and BNC-Baby. One of the notable characteristics differences between the nuclear science English corpus and BNC-Baby is its use of pronouns. For example, BNC-Baby contains uses personal pronouns such as he, she, I, you with high frequency. In the nuclear science English frequency lists none of these pronouns in the section of with high frequency. However, as one can expect easily, both the frequency list contains function words in a higher rankings. Further, the most content words in the nuclear science English corpus are the discipline specific terms that may represent the whole idea of the corpus. Next, to statistically identify the words that are used saliently from the general English usages, log-likelihood of each word in the nuclear science English corpus is calculated. The calculation provides a different word list, a list often called a keyword list, and depending on the saliency of words, this list may or may not contain the words in the frequency list. Once certain words that are saliently used in the nuclear science English corpus, then it is necessary to locate how these salient words are uses in the context of nuclear science discipline. To understand how the keywords are used in the discipline context, the relationship between the keywords and their neighbouring words should be examined. The relationship can be measured using a statistical technique called mutual information (MI) or cubic mutual information (MI3) depending the emphasis on higher frequency words or lower frequency words. The current research explored different statistics to examine high frequency words, such as function words and the lower frequency words, for example content words at the same time.

3 Results and discussion

Of the generated keywords, the discussion can be made regarding the salient function words and content words of the nuclear science English. A total of 157 instances of the phrase *of these* were found in the nuclear science English corpus. Many of the cases were found at the beginning and the end of the research articles. The following words of the phrase are the words like *components*, *factors*, *elements*, and *codes*, which express the parts of what is described or explained in the earlier context.

For the phrase *of this*, a total of 112 instances were located. Unlike the phrase *of these*, the current structure is frequently found at the beginning of the research articles. In addition, the words following the phrase are *study*, *paper*, *research*, *approach*, *work*, which thereafter explains the 'purpose' of the research article.

There are content words expressing certain connotation that they may mean differently in general English usage. One of the keywords is the word *human*. The word *human* in the nuclear science English 'doubtful' or 'imperfect' considering the words that if modifies: *human error(s)*, *human performance*, *human reliability*, etc. On the other hand, the same word goes well with words denoting neutral: *human being(s), human rights, human behaviour*.

The series of analyses above has shown that the nuclear science English words can be characterized by (1) the organization of the written texts, (2) collocations of certain phrase structures, and (3) the connotation in the discipline. It is expected that the results of the study can be a useful source for understanding the nuclear engineering English stylistics. Further, the results can be utilized useful resources for developing an ESP education programs in a given discipline.

References

Belcher, D. (Ed.). (2009). *English for Specific Purposes in Theory and Practice*. Ann Arbor, MI: University of Michigan Press.

Gledhill, C. J. (2000). *Collocations in science writing*. Tübingen, Germany: Gunter Narr Verlang.

Hyland, K. (2011) Disciplinary specificity: discourse, context and ESP. In D. Belcher, A.M. Johns & B. Paltridge (Eds.), *New Directions in English for Specific Purposes Research* (pp. 6-24). Ann Arbor, MI: University of Michigan Press.

Parkinson, J. (2013). English for Science and Technology. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 155-174). New York: Blackwell.

Römer, U. & Wulff, S. (2010). Applying corpus methods to writing research: Explorations of. MICUSP. *Journal of Writing Research, 2*(2), 99-127.

Automated vocabulary quiz creation using online and offline corpora

Ralph Rose
Waseda University
Faculty of Science and Engineering
rose@waseda.jp

1 Summary

Word Quiz Constructor is a Java application designed to create a large number of quizzes from word lists by drawing test materials from large online or offline corpora. Pilot tests show that the generated quizzes are close to acceptable levels of reliability but further development is needed.

2 Background

Several tools have been created to generate vocabulary test questions for learners (e.g., Aist 2001; Brown et al. 2005; Kunichika et al. 2003; Lee et al. 2013). These tools can reliably provide a variety of question types that result in scores that correlate well with human-generated questions. The present work seeks to add to this body of work by developing a tool that can construct vocabulary quizzes en masse. This was motivated by the need for such quizzes for a large-scale, highly-managed university English language program taught by many teachers in many different time slots, where consistency across sections but also quiz security was desired. Word Quiz Constructor (hereafter, WQC) was designed to meet this need.

3 Basic design

WQC is a Java application which takes a list of target vocabulary items (e.g., academic word list of Coxhead 2000), a set of user-defined parameters for the desired quiz (e.g., number of items, number of options for multiple choice items, difficulty, source corpus) and generates a quiz, drawing random target words from the vocabulary list. Currently, WQC can generate two question types, as follows.

Multiple-choice questions provide a stem in which the target word is replaced by a blank and is provided as one answer option along with three distractor items. The distractor items are chosen by tri-gram analysis of the target word’s context in the original stem: In short, high-frequency contexts for the target word are first chosen, and then the target word is replaced by random vocabulary items (controlling for part-of-speech) in the context to find distractor words with a tri-gram frequency of zero; hence, words that are presumed to be highly unsuited to the context. Tri-gram frequencies are based on the British Academic Written English (BAWE) corpus (Gardner and Nesi 2012).

Definition questions provide two independent stems in which the target word has been replaced by a blank along with a definition of the word drawn from WordNet 3.0 (Miller 1995). Test-takers are required to write the target word that matches the definition and fills in both blanks.

Sample sentences used in the stems of both question types are drawn from Wikipedia, using the MediaWiki API to query the Wikipedia database for random pages containing the target words. The texts are processed using the Stanford parser (Klein and Manning 2003) and the difficulty level of the stems is controlled by using the automated readability index (ARI: Smith and Senter 1967). Although WQC has been primarily tested with Wikipedia as the source corpus, it is capable of working with other on-line or off-line corpora.

Quiz questions can be output in plain text format for use as paper-based materials, or in csv or xml formats for uploading to question banks in course management environments (e.g., Moodle, Blackboard). WQC can also run in a batch mode for the mass production of quizzes.

4 Performance

WQC is still under development, but has been informally piloted with nine classes comprising over 400 students at Waseda University in Japan. Results from various administrations of quizzes constructed by WQC show an average Cronbach's alpha of 0.55, not quite reaching the typically regarded acceptability threshold of 0.7. This may be a result of the limited tri-gram coverage of the BAWE corpus, which meant that the difference between the high-frequency and low-frequency thresholds for target words and distractors was rather small so that some distractors may have actually been plausible options. Furthermore, because of the academic writing style used in Wikipedia, the ARI threshold could not be set too low or no items could be constructed successfully. Thus, several items may have been too difficult, with an ARI at a level higher than university level).

5 Future work

Future plans for WQC include changing the n-gram data from that derived from BAWE to the Google n-gram corpus which is more robust, as well as adding more question types, and a graphical user interface. Furthermore, more formal validation tests are planned in order to verify the reliability of the question items produced.

References

Aist, G. 2001. Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment, *International Journal of AI in Ed* 12: 212-231.

Brown, J., Frishkoff, G. and Eshkenazi, M. 2005. Automatic question generation for vocabulary assessment. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 819-826. Association for Computational Linguistics.

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

Gardner, S. and Nesi, H. 2012. A classification of genre families in university student writing. *Applied Linguistics* 34 (1): 1-29.

Klein, D. and Manning, C.D.. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Kunichika, H., Katayama, T., Hirashima, T. and Takeuchi, A. 2003. Automated question generation methods for intelligent English learning systems and its evaluation. *Proceedings of ICCE2004*.

Lee, K., Kweon, S., Kim, H. and Lee, G. 2013. Filtering-based Automatic Cloze Test Generation. *Proceedings of Speech and Language Technology in Education (SLaTE)*, pp. 72-76.

Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39-41.

Smith, E.A. and Senter, R.J. 1967. Automated Readability Index.. Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, USA. AMRL-TR-6620.

Students’ perception of implicit acquisition of medical terms in foreign languages

Ariane Ruyffelaert
Ghent University, Belgium;
University of Granada, Spain
ariane.ruyffelaert@ugent.be

Víctor Carriel
Ghent University, Belgium;
University of Granada, Spain
vcarriel@ugr.es

1 Introduction

The knowledge of foreign languages (FL) is an important competence in medical and scientific careers, because most of the scientific information is in English or French in Europe. In addition, the acquisition of scientific terms and teaching of FL for specific purposes is difficult and it is not adequately addressed in the basic and university curricula of medical students. In this sense, we evaluated medical students’ perception of the implicit acquisition of medical terms in English and French.

2 Methodology

This study was performed with voluntary medical students enrolled in their first year at the School of Medicine of the University of Granada (Spain). For this propose, multiple choice histological tests in English (n=230), French (n=201) and Spanish (n=252) were performed. Each test was composed by ten representative histological images with their respective questions and 4 choices. These tests were performed according to the practical course of histology (dictated by V.C.). The English and French tests were performed using the application of www.thesistools.com. In the case of Spanish, the test was the final practical exam of the course (using exactly the same methodology). For each test, we calculated the scores reached by each student from 0 to 10, and the results were expressed as mean ± standard deviation and the utility of this approach was questioned by a voluntary survey.

3 Results

The tests were not performed by all the students (n=275). 92% performed the Spanish test, 84% the English test and 73% the histological test in French. The perception of the students differed between the English and French tests. The results were more favorable for the English histological test, where 71% of the students found this novel approach useful and 9% found the test useless. We observed that 3% of the students did not answer the question about the usefulness of the English test and 16% of the students did not perform the histological test in English (see table 1).

Survey answers	N	%	Mean scores English/Spanish
Useful	196	**71,27**	**6,52 / 6,88**
Useless	25	9,09	5,88 / 6,85
Usefulness without answer	9	3,27	4,11 / 7,23
Total performed test	230	83,6	6,44 / 6,88
Test not performed test	45	16,36	- / **4,97**
Total of students	275	100	

Table 1. English test scores and perceptions

The perception of the students about the histological test in French was less favorable. 52% of the students found this test useful and 16% useless. 4% of the students left this question blank in the survey and 26% of the students did not perform the histological test in French (see table 2).

According to the score of the histological tests, the means were 6.44±1.93 in English (see table 1), 4.45±2.32 in French (see table 2) and 6.63±1.5 in Spanish.

Survey answers	N	%	Mean scores French/Spanish
Useful	145	**52,73**	**4,79 / 6,93**
Useless	44	16	3,70 / 6,96
Usefulness without answer	12	4,36	3,17 / 7,03
Total performed test	201	73	4,45 / 6,94
Total not performed test	74	26,91	- / **5,56**
Total of students	275	100	

Table 2. French test scores and perceptions

4 Discussion and conclusion

The progressive application of this type of tests could be useful for the implicit acquisition of scientific terms in FL. This preliminary study demonstrated that medical students are open-minded to follow innovative approaches for the acquisition of medical terms in FL, especially in English. In addition, the poor scores reached in French, suggest that these students have a poor background in this language, and could explain their little motivation to perform this test. In conclusion, this approach had a positive impact on the histological knowledge of the participating students who obtained higher scores in their final exam. In addition, the students demonstrated a positive perception to this novel method, which could be a useful tool for the implicit acquisition of medical terms in FL. However, more analyses are needed to demonstrate the efficacy of this novel approach.

Acknowledgments

This study was financed by the Department of French Linguistics, Ghent University, Belgium. The authors are grateful to the Professors P. Hadermann (UGent, Belgium), J. Suso and M. Alaminos (UGR, Spain).

How does a corpus influence learning L2 collocations?

Yoshiho Satake
Aoyama Gakuin University
yoshiho.satake.sugitani@gmail.com

1 Introduction and literature review

While the strength of the use of corpora in language teaching has been stated, more empirical research needs to be conducted on the effectiveness of the use of corpora on language learning (Flowerdew 2010).

According to Flowerdew (2010), interacting with corpora helps learners acquire phraseological patterning (i.e. collocations, colligations and semantic preferences and prosodies) because these features are not easily accessible in either dictionaries or grammar books.

Satake (2014) states that dictionary users tend to look up and memorize more collocations and corpus users tend to output more collocations. However, since she used the Japanese-English translation test to evaluate learners’ collocational knowledge, their L1 could have influenced the results favorably for the dictionary users who mainly used English-Japanese dictionaries.

More empirical research is needed to judge whether corpus use is effective in improving L2 collocational knowledge, and how DDL works if it is effective.

2 Research questions

The aim of this study was to investigate the effects of corpus use on learning L2 collocations. In order to investigate how corpus use influenced learning L2 collocations, the effects of corpus use were compared with the effects of dictionary use. The following research questions were addressed:

(1) Do a corpus and dictionaries produce different effects on memorizing collocations?

(2) Do a corpus and dictionaries produce different effects on learners’ word associations?

(3) Do a corpus and dictionaries produce different effects on output of collocations?

(4) Do corpus and dictionary users access and process different information?

3 Methods

The two group (experimental vs. control) pre-post design was used to analyse the effects of corpus use on learning collocations.

Two classes of Japanese undergraduate students (in total, sixty students) at a private university in Tokyo participated in the study. They were upper intermediate English learners and reached level B1 to B2 in the Common European Framework of Reference for Languages (CEFR).

The target word was “marrow” and how students learned collocational knowledge of the word through a corpus or dictionaries was investigated. The word “marrow” was used because it is not a high frequent word and thus students would not have enough collocational knowledge of it. When students looked up collocations, one class with twenty-nine students used Corpus of Contemporary American English (COCA) and the other class with thirty-one students used dictionaries. The students who used COCA were given instruction on how to use it.

The following procedure was taken in the present study.

(1) Pre-test (five fill-in-the blank questions with four choices and an association test for the target word, 5 minutes)

(2) Treatment (looking up collocations of the target word “marrow” in COCA or dictionaries, 10 minutes)

(3) Post-test (almost the same as the pre-test, except that the post-test also asks the students to write phrases using “marrow” as many as possible, 5 minutes)

4 Results

The results were as follows.

(1) Both the corpus and the dictionaries were significantly effective in memorizing collocations and their effects were not significantly different.

(2) The corpus was significantly more effective in improving learners’ word associations than the dictionaries.

(3) The corpus was significantly more effective in promoting learners’ output of collocations than the dictionaries.

(4) Although the corpus users looked up fewer collocations than the dictionary users, the corpus users looked up more frequent collocations than the dictionary users.

5 Discussion and conclusion

While both the corpus and the dictionaries were significantly effective in memorizing collocations, the corpus was significantly more effective in improving learners’ word associations and output of collocations than the dictionaries.

Considering the corpus users looked up fewer but more frequent collocations within the time limit and produced more word associations and collocations than the dictionary users, the results admit of two interpretations: (1) the information the corpus users collected through the corpus was highly useful. (2) the corpus users could use the information they collected more efficiently than the dictionary users. This could be because the corpus users had more time to process each information than the dictionary users. The difference of the time they could spend for each information would have affected how deeply students processed each information, how their word associations were improved and how their output of collocations was promoted.

The strength of using corpus for learning collocations lies in its effectiveness in promoting learners’ word associations and output of collocations. However, it remains unclear why the use of corpus has such effectiveness and more research is needed.

References

Flowerdew, L. 2010. “Using corpora for writing instruction”. In A. O’Keeffe and M. McCarthy, (eds.) *The Routledge Handbook of Corpus Linguistics*. Abingdon, Oxon: Routledge.

Satake, Y. 2014. Corpora vs. dictionaries: their effects on learning English collocations in L2 writing tasks. The paper will be read at the second Asia Pacific Corpus Linguistics Conference, at the Hong Kong Polytechnic University, from March 7th to 9th, 2014.

A corpus-based German language course for civil engineering students

Sigrun Schroth-Wiechert
Leibniz University of Hannover
schroth-wiechert@fsz.uni-hannover.de

This presentation aims to contribute to the question of how to implement a corpus-based teaching and learning approach in a language classroom at university level. Typical questions that I am faced with every day would be:

1. Do you know the rules regarding a space between a number and a unit or symbol, for example 100 %?

2. Do you know the rules regarding a hyphen between a combination of words, for example Pin-on-Disk-Verfahren?

3. Are both forms correct: nicht mobil – unmobil?

4. Is there a difference between „erwünschten“ (desired) and „gewünschten“ (favored) results?

As a specialist in German as a second language with a focus on didactics, I have no engineering background. Nonetheless, I agreed in 2003 to give a course for technical German at the FSZ, which was supposed to cover reading, listening and writing comprehension.

Very quickly it became obvious that writing is one of the major problems for foreign students of civil engineering, and this lead me to establishing the so called “writing-mentor” approach. It was also obvious that the writing-problem was not the specialised lexical terminology or semantics but rather the general scientific language.

After spending many years correcting a lot of technical reports in a university context written by foreign students from all over the world, I became familiar with the kinds of writing difficulties these students face. As time went by, many mistakes and many questions from the students recurred.

I began to look for textbooks or other material to support my students’ writing process but there was nothing available. This lack of material then lead me to start defining the typical categories of writing goals or communicative purposes (e.g. to compare, to introduce the subject) and to illustrate these with authentic examples taken from technical reports written by Germans. This approach culminated in the book „Deutsch als Fremdsprache in den Ingenieurwissenschaften – Formulierungshilfen für schriftliche Arbeiten in Studium und Beruf“ published in 2011 by Cornelsen[25]. The goal is for the students to transfer those structures into their own writing and complete them with the specific terminology of their subject.

There is still a very long list of questions from the students like the one I presented above. It is not possible to answer all of them professionally; some just with my native-speaker intuition, which might be satisfactory for the students – but not for myself.

My every day work shows that there is a necessity to continue the investigation of the German scientific language of engineering in order to find conventions and rules and finally to answer the kind of the questions mentioned above.

Due to a lack of time and sheer necessity, I involved students in an investigation of the German scientific language of engineering in order to find conventions and rules. In the winter semester 2012/13 I carried out a so called “research course” at the Centre for Applied Linguistics and Special Languages with the title: “Technical German: research course scientific language”.

The main characteristic of this course concept is maximizing learner autonomy in the classroom. The focus for the teacher is on the compilation of the corpus before the course starts and on the post-processing of the material developed by the students during and after the course. To allow the material to be used by other student generations, the results have been published on the FSZ-website[26]. In essence it is a win-win situation once the students have understood the meaning and the content of the so-called "research course".

[25] http://www.cornelsen.de/technik-daf/; 24.04.2014

[26] http://www.fsz.uni-hannover.de/materialien.html; 24.04.2014

A learner corpus-based study on relative clause constructions as criterial features for the CEFR levels

Yuka Takahashi
Tokyo University of Foreign Studies
busagi0v0@gmail.com

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

1 Introduction

In foreign language learning and teaching, it is desirable to set the learning objectives clearly so that both materials designers and teachers in the classroom have clear images of what should be in the course materials and what should be taught in the classroom. In the last decade, especially in Europe, an effort has been made to make this goal as explicit as possible by offering the common reference framework for languages, called the Common European Framework of Reference for Languages (CEFR). Since the Council of Europe officially announced the use of the CEFR for designing and evaluating foreign language syllabus and materials designs in each EU country in 2002, the use of the CEFR has been constantly expanding not only within Europe but also to the other parts of the world.

This framework is generic and language-independent. Thus it is underspecified as to what kind of grammar and lexis should be taught for each CEFR level. To supplement the framework, the procedure called Reference Level Descriptions (RLDs) has been undertaken, in which grammar points and lexical items are identified for each CEFR level.

Projects such as the English Profile Programme (EPP) (Hawkins and Filipovic 2012) use corpus data intensively in order to identify criterial features. The EPP especially is quite ambitious in the sense that they use both native and learner corpora to determine to what extent certain linguistic features serve as criterial for particular CEFR levels.

In the same vein, we have been investigating the nature of criterial features for Japanese learners of English, using our own corpus resources (Tono 2012; 2013). One of the features we focused on in this study is a relative clause construction, which is said to be one of the most difficult grammar items for learners of English (Hawkins and Buttery 2010) and also very frequently mentioned in SLA literature (cf. Ellis 2008: 562ff). By closely examining the state of acquisition of relative clauses, we hope to discover the path of identifying criterial features not just by quantitative, statistical methods, but also by looking at the process of acquisition in more detail.

2 Method

The corpus used in this study is the Japanese EFL Learner Corpus (JEFLL) (Tono 2007). It consists of written compositions by 10,038 Japanese secondary school students (669,304 running words). Originally the corpus was classified by school years, but in a new government-funded project, the entire JEFLL Corpus has been re-classified into CEFR levels.

The entire corpus was tagged for POS using TreeTagger. Extraction of relative clause constructions was done by writing pattern matching queries using regular expressions for the parts of speech of antecedents and each relative pronoun. The zero relative pronoun, which is common in producing contact clauses, was not covered in the present study.

All the instances of relative clauses were classified into the following categories:

- Categories based on the Noun Phrase Accessibility Hierarchy (NPAH) (Comrie & Keenan 1979) Hypothesis: S/DO/IO/GEN/OBL/OCOMP
- Categories based on the SO Hierarchy Hypothesis (Hamilton 1994) : SS/SO/OS/OO

Also each sentence was judged in terms of grammaticality and annotated for errors based on the following criteria:

- wrong selections of relative pronouns
- resumptive pronouns
- wrong matrix positions

The present study aims to answer the following research questions:

- RQ1: Does the use of relative clauses increase along the CEFR levels, thus serving as criterial features?
- RQ2: Does the distribution of the use of relative pronouns across the CEFR levels confirm the NPAH Hypothesis?
- RQ3: Does the distribution of the use of relative pronouns across the CEFR levels confirm the SO Hierarchy Hypothesis?
- Are there any cases where errors were uninterpretable? If so, what seems to be the problem?

3 Results

The results show that basically the number of relative pronouns used in the essays was found to be increasing across the CEFR levels. Therefore, the first research question was confirmed.

Regarding the two hypotheses related to RQs 2 and 3, overall, while the SO Hierarchy Hypothesis was largely supported, the NPAH Hypothesis was partially supported due to the lack of evidence in GEN, OBL and OCOMP. These occurrences are also relatively infrequent in native corpora, compared to S and DO, so it seems that the results are reasonable.

Finally, we noted very interesting cases for L2 interlanguage state. There are cases in which the use of relative pronouns seemed to trigger more errors in the embedded sentences. For instance, errors such as tense/aspect, a confusion of intransitive verbs as transitive seemed to be more frequent in embedded clauses than in ordinary clauses. We will report this result using special interlanguage annotation schemes.

References

Comrie, B. and Keenan, E. 1979. Noun phrase accessibility revisited. *Language* 55: 649-664.

Ellis, R. 2008. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Hamilton, R. 1994. Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning* 44: 123-157.

Hawkins, J. and Buttery, P. 2010. Criterial features in learner corpora. *English Profile Journal* 1 (1), e5.

Hawkins, J. and Filipovic, L. 2012. *Criterial Features in L2 English*. Cambridge: Cambridge University Press.

Tono, Y. 2007. *Nihonjin 1-mannin no Eigo Corpus: JEFLL Corpus*. Tokyo: Shogakukan.

Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.) 2012. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins.

Tono, Y. 2013. Automatic extraction of L2 criterial lexico-grammatical features across pseudo-longitudinal learner corpora: Using edit distance and variability-based neighbour clustering. In C. Bardel, C. Lindqvist & B. Laufer (eds), *L2 Vocabulary Acquisition: Knowledge and Use: New perspectives on assessment and corpus analysis* (pp.149-176). EuroSLA monographs. EuroSLA.

Using corpus annotation for teaching contrastive linguistics

Beata Trawinski
Institut für Deutsche Sprache (Mannheim, Germany)
trawinski@ids-mannheim.de

The use of corpora in (foreign) language learning and teaching has grown increasingly over the past decades. Large electronic collections of written and spoken texts allow learners to access natural language data of any type and any complexity (parallel corpora are particularly useful for these purposes). There is also a wide range of corpus-based reference works for teaching and learning languages, such as dictionaries, grammars and other teaching materials and tools. In this paper, we suggest how the national corpora of the Slavic languages and, more precisely, their morphosyntactic annotations, can be used for teaching contrastive Slavic linguistics.

Currently, there exist the following (national) corpora of Slavic languages: the Bulgarian National Corpus (http://www.ibl.bas.bg/BGNC_bg.htm), the Croatian National Corpus (http://www.hnk.ffzg.hr), the Czech National Corpus (http://ucnk.ff.cuni.cz), the Slovak National Corpus (http://korpus.juls.savba.sk), the National Corpus of Polish (http://www.nkjp.pl), the FidaPLUS corpus of Slovenian (http://www.fidaplus.net) and the Russian National Corpus (http://www.ruscorpora.ru). All of them are POS-tagged and morphosyntactically annotated. However, the sets of the morphosyntactic tags used in the particular corpora differ. While some of these differences are purely notational (cf. the POS labels assigned to nouns / substantives: S in the corpora for Russian and Slovak languages, SUBST in the National Corpus of Polish and N in the Bulgarian, Croatian, Czech and Slovenian corpora, which use the Multext-Eeast tagset), others reflect the morphosyntactic peculiarities of the languages in question. For example, the morphosyntactic information provided in the Bulgarian National Corpus does not include the grammatical category of case, which is included in the remaining Slavic corpora. Conversely, the Bulgarian Corpus provides the category of definiteness, identifying the values 'indefinite' and 'definite', which are not available in the other Slavic corpora. Another example can be found in the National Corpus of Polish, which provides five values for the grammatical category of gender: 'human masculine', 'animate masculine', 'inanimate masculine', 'feminine' and 'neuter', not found in that form in any other Slavic corpus. The corpus of

Slovenian, on the other hand, offers three possible values for the category 'number': 'singular', 'plural' and 'dual', the last of which does not appear in the other corpora.

We see interesting potential here for comparing empirically motivated morphosyntactic corpus annotations in the contrastive (Slavic) linguistics classroom. Based on the classical inventory of parts of speech and grammatical categories, we can discuss together with our students the empirical motivation for adopting specific tags in a given corpus, and in this way discover systemic differences and similarities between the (Slavic) languages. What is distinctive about this way of introducing (contrastive) theoretical linguistic issues to students is that it is based on authentic language data.

How oral history helps pupils become researchers

Marina Tzakosta
University of Crete
martzak74@gmail.com

Angelos Patsias
Fourfouras Primary School
angelpats@gmail.com

Anna Sfakianaki
University of Ioannina &
Aristotle University of Thessaloniki
runna07@gmail.com

1 Introduction

Linguistic varieties are complete language systems just like standard languages (cf. Kontosopoulos 1997, Ntinas & Zarkogianni 2009). Specifically for Greek, while the national curriculum (2002, 2003) promotes literacy as well as the communicative approach of language teaching in kindergarten and primary school, the teaching of dialects and dialectal variants is absent from the Greek school. Nevertheless, the teaching of different language varieties and forms of standard Greek gives pupils the possibility, on the one hand, to be acquainted with the treasures of the expressive means of their mother language as well as embody it in a broader cultural and historical context. On the other hand, dialect teaching helps pupils discover the grammatical adjacency of linguistic varieties that pupils acquire together with the standard language. Dialect teaching further facilitates the cultivation of the pupils’ metalinguistic capacity, i.e. the conscious knowledge and successful manipulation of the standard language and the dialects at all their grammatical levels (phonology, morphology, syntax, semantics, pragmatics). We argue that none of the above can be achieved if there is no active interaction of pupils and educators during the educational process.

2 Aims

Aim of the project presented here is the design, construction and organization of the Digital Museum of Greek Oral History (DiMuGOHi) as a research and educational tool available to both pupils and educators. Side goals of the project are, on the one hand, the training of Primary and High School pupils on methods of language data collection and processing and, on the other hand, the collection, processing, filing, preservation, and diffusion of dialectal linguistic data which will be available in audiovisual files. DiMuGOHi will have the form of a digital platform. It will be useful as a tool for teaching language, geography, social sciences, local history, familiarize pupils with environmental education and any topic relevant to education through dialectal speech. DiMuGOHi will a) contribute to the sensitization of pupils regarding dialectal issues, b) exempt the latter as important linguistic systems and c) investigate topics of local history (cf. Thompson 1978).

Within the context of the same project, we will proceed to the indexing and investigation of major phonetic and phonological characteristics of the cretan dialect in order to explore the ways through which specific phonetic and phonological aspects of the dialect influence speech production, word formation and vocabulary development and enrichment. The dialect of western Crete will constitute the core of DiMuGOHi, however, the museum will be designed in such a way so that it will be able to host dialectal material from various regions of Greece. In our presentation we will also display the major axes on the bases of which the platform will be designed as well as the principles underlying the suggested activities.

3 Conclusion

DiMuGOHi will contribute to the educational procedure in direct and indirect ways. A first direct outcome is pupils' sensitization regarding dialects, their linguistic properties and structural adjacency with standard Greek as well as the role of the dialects in the preservation and diffusion of local and national cultural heritage. In addition, it will accentuate issues of environmental education, social sciences, geography, local history. Some indirect outcomes are that, first, it will facilitate the improvement of metalinguistic awareness regarding the cretan dialect, second, it will enhance knowledge that the dialect has a complete linguistic system, just like the standard language, third, it will improve stylistic and sociolinguistic awareness, namely the conscious knowledge of the linguistic contexts in which dialectal material is used, and, fourth, it will preserve the linguistic treasures of linguistic varieties.

References

Kontosopoulos, N. 1997. *Topics of the cretan dialect. Reproduction of studies [Θέματα Κρητικής Διαλεκτολογίας. Αναδημοσίευση μελετών]* [in Greek]. Athens: Anastasakis Publ.

Ministerial Decision Γ2/21072β - ΦΕΚ 304/vol. Β΄/13-3-2003. *Cross-thematic unified framework of study curricula for preschool education*.

Ntinas, K.D. & E.H. Zarkogianni. 2009. *Didactical development of modern Greek dialects. The case of the variety of Afantou Rhodes [Διδακτική αξιοποίηση των νεοελληνικών Διαλέκτων. Η περίπτωση του Ιδιώματος Αφάντου Ρόδου]* [in Greek]. Thessaloniki: University Studio Press.

Thompson, P. 1978. *The voice of the past: oral history*. Oxford: O.U.P.