Learner Corpus Research 2013

Book of Abstracts

Programme Committee Chairs:
Kari Tenfjord, Anne Golden, Fanny Meunier, Koenraad De Smedt
Bergen, Norway
September 27–29, 2013

UNIVERSITETET I BERGEN

Acknowledgements

The Program Committee chairs acknowledge the financial support to the conference provided by the University of Bergen and the Research Council of Norway. We thank the Program Committee for the 294 reviews that they produced. The abstracts, titles, author names and keywords were provided by the respective authors and were collected with the help of the EasyChair system. Last but not least, we thank the local organization led by Victoria Rosén for the practical organization of the conference.

Kari Tenfjord, Anne Golden, Fanny Meunier and Koenraad De Smedt

UNIVERSITAS BERGENSIS

Norges forskningsråd
The Research Council of Norway

Program Committee

Andrea Abel
Theodora Alexopoulou
Ulrike Altendorf
Marcus Callies
Meilin Chen
Koenraad De Smedt
Pieter De-Haan
Liesbeth Degand
Markus Dickinson
María Belén Díez Bedmar
Signe Oksefjell Ebeling
John Flowerdew
Lynne Flowerdew
Thierry Fontenelle
Gaëtanelle Gilquin
Anne Golden
Stefan Gries
Jone Grigaliuniene
Nicholas Groom
Hilde Hasselgård
Marlise Horst
Shin'Ichiro Ishikawa
Jarmo-Harri Jantunen
Sofie Johansson Kokkinakis
Tomoko Kaneko
Natalie Kübler
Marie-Aude Lefer
María José Luzón
Anke Lüdeling
Amaya Mendikoetxea
Fanny Meunier
Florence Myles
Joanne Neff
Nadja Nesselhauf
Magali Paquot
Pascual Pérez-Paredes
Paul Rayson
Juan-Pedro Rica-Peromingo
Victoria Rosén
Ute Römer-Weyhofen
Kari Tenfjord
Yukio Tono
Sylwia Twardo
Bertus Vanrooy
Annelie Ädel

Table of Contents

Linguistic features discriminating between native English speakers and East Asian learner groups with different proficiency levels

Mariko ABE; Yuichiro KOBAYASHI; Masumi NARITA
Chuo University; Ritsumeikan University; Tokyo International University
abe.127@g.chuo-u.ac.jp; kobayashi0721@gmail.com; mnarita@tiu.ac.jp

In recent years, the application of large computational databases of written and spoken samples produced by language learners has developed as a way to create a general picture of interlanguage use. However, fewer learner corpus-based studies have targeted East Asian learners of English, because a sufficiently large-scale learner corpus with proficiency level information based on an objective rubric was not available to the public. Additionally, despite early work on Second Language Acquisition (SLA), relatively few researchers have been concerned with describing interlanguage development using multiple linguistic features (Biber, Conrad, & Reppen 1998).

In order to address these problems, this study aims to profile the language use characteristics of East Asian learners of English from multiple linguistic aspects. The following research question is investigated to accomplish this purpose: What linguistic features distinguish native and non-native speakers of English and the proficiency levels of different L1 groups?

The present study is based on the methodology originally developed by Biber (1988) to analyze the differences in the spoken and written language of native speakers of English. As in Biber's (1988) study, we used a large amount of linguistic data showing multiple linguistic features to identify the linguistic features characterizing learner language. We used the International Corpus Network of Asian Learners of English (ICNALE), the largest East Asian composition database in which each composition was coded with the participant's CEFR level. Written compositions by 2,000 EFL learners and 400 native speakers of English were analyzed in terms of the automatically extracted 58 linguistic features in Biber's (1988) list.

We used correspondence analysis and cluster analysis, since they are more suitable for investigating similarities among variables (McEnery & Hardie 2012). Correspondence analysis is suitable for reducing the complexity of the data, and cluster analysis is useful for classifying large groups into meaningful clusters that are similar to each other. Correspondence analysis was used to initially identify whether various linguistic features cluster to form CEFR levels and then cluster analysis was used to specify groupings of CEFR levels by different L1s in more detail.

The results indicated that "native speakers of English and EFL learners from Hong Kong" and "EFL learners from Japan, Korea, and Taiwan" clearly formed two different groups. Learners tended to use nouns more frequently than native speakers of English. Most interestingly, Japanese EFL learners used (1) first person pronouns, (2) the present tense, and (3) independent clause coordination more frequently than the other L1 groups, although they showed less frequent use of (1) attributive adjectives, (2) emphatics, (3) other adverbial subordinators, and (4) predictive modals. In contrast, native speakers of English used (1) the perfect aspect, (2) split auxiliaries, (3) adverbs, and (4) the relative subject more frequently than the learner groups.

English proficiency of four learner groups was classified into three clusters: (1) native speakers of English, CEFR B1- and C1-level learners in Hong Kong, and C1-level learners both in Japan and in

Korea, (2) A2-, B1-, and B2-level learners in Japan, and (3) the rest of the learners. It is noteworthy here that learners from Hong Kong seem to have higher proficiency than the others, and the more frequent use of first personal pronouns found in A2- to B2-level learners in Japan seems to discriminate these proficiency levels from their C1-level peers.

In conclusion, frequency patterns of key linguistic features were identified to distinguish both learner language variation and English proficiency levels among different L1 groups. To enhance our understanding of the nature and characteristics of learner language, further studies are necessary to fully examine the criteriality of major distinguishing features of proficiency levels that have emerged from the present analysis, as strongly suggested by Hawkins and Filipović (2012).

References

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

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

Hawkins, J. A., & Filipović, L. (2012) *Criterial features in L2 English*. Cambridge: Cambridge University Press.

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

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

A Trilingual Learner Corpus illustrating European Reference Levels

Abel, Andrea*; Nicolas, Lionel*; Hana, Jirka♦; Štindlová, Barbora✦; Bykh, Serhiy▪; Meurers, Detmar▪
European Academy Bolzano/Bozen*; Charles University Prague♦; Technical University Liberec✦; University of Tübingen▪
andrea.abel@eurac.edu; lionel.nicolas@eurac.edu; jirka.hana@gmail.com; barbora.stindlova@tul.cz; katrin.wisniewski@tu-dresden.de; detmar.meurers@uni-tuebingen.de; sbykh@sfs.uni-tuebingen.de

Since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has gained a leading role as an instrument of reference for language teaching and certification and for the development of curricula (cf. CEFR 2001). At the same time, there is a growing concern about CEFR reference levels being insufficiently illustrated, leaving practitioners without comprehensive empirical characterizations of the relevant distinctions. This is particularly the case for languages other than English (cf. e.g. Hulstijn 2007, North 2000).

The MERLIN project addresses this demand for the first time for Czech, German and Italian by developing a didactically motivated online platform that enables CEFR users to explore authentic written learner productions that have been related to the CEFR levels in a methodologically sophisticated and rigorous way. The core of the multilingual online platform is a trilingual learner corpus relying on a cross-linguistic design and composed of roughly 2300 learner texts produced in standardized language certifications (as used by telc, DE, UJOP-Charles University in Prague, CZ as well as UNIcert, DE) validly related to the CEFR, covering the levels A1-C1.

The aim of this paper is both to present the MERLIN project and its corpus and to discuss its current state. In addition to providing preliminary statistics, we detail three key aspects: (1) the data collection and transcription, (2) the creation of the annotation schemata and (3) the technical design of the workflow devised to compile and query the corpus. For each of these aspects, we highlight its requirements and the means employed for the underlying tasks. We also discuss the challenges induced by the cross-linguistic nature of the project: particularities of languages from three different families (Slavic, Germanic and Romance) have to be covered, both on a linguistic and a technical perspective.

(1) We explain how data was collected under standard test conditions during accredited and audited tests from highly representative testing institutions. We also explain how texts were rated by specially trained experts on the basis of guidelines created for the three languages (CEFR scales used: vocabulary range, vocabulary control, grammatical accuracy, coherence & cohesion, sociolinguistic appropriateness, orthographic control) and how statistical quantitative and qualitative measures, e.g. Rasch analyses to estimate inter-/intra-rater reliability, were devised to ensure quality. To ensure test intra-rater reliability, we double-rated part of the corpus.

Regarding transcription, we detail how the hand written texts were transcribed and encoded in XML using a set of transcription tags (<greeting>, etc.), double-checked following guidelines we have developed, and prepared for additional manual and automatic processing.

(2) We explain how indicators of L2 proficiency were identified to be later used for analyzing the L2 productions. These indicators were selected through a large study of four different perspectives: (a) inductively derived indicators on the basis of the linguistic analyses of the performance samples, (b) CEFR indicators derived on the basis of the operationalisation of the CEFR scale descriptors, (c) deductively derived indicators on the basis of the relevant literature on SLA and language testing, (d) experientially derived indicators on the basis of textbook analyses and the projects questionnaire study. We thus explain how the meaningful and robust indicators for describing L2 competences for Czech, German and Italian were identified and how harmonised annotation schemata taking both

common and language-specific features into account (e.g. gender/article in German, reflexive possessive pronouns in Czech, pronoun particles in Italian) were established (cf. e.g. Wolfe-Quintero, Inagaki & Kim 1998, Ortega 2003, Read 2007, Rimrott & Heift 2008, Housen & Kuiken 2009, Lu 2011, Granger & Bestgen 2011, Mellor 2011, Vajjala & Meurers 2012).
(3) We detail how we split the workflow into five fundamental tasks: transcription, manual annotation, automatic annotation, linguistic analysis and conversion of the data to the required formats. Corpus representation decisions taken at the beginning have important and long-lasting implications in terms of the broad and sustained usefulness of a corpus. Fundamental decisions on basic units, such as tokenization and anonymisation, influence all later annotation steps and the ability to retrieve the data appropriate for a particular use case. We discuss the fundamental representation issues and the tools we developed or adapted (e.g. Schmidt 1994, Müller & Strube 2006, Zeldes et al. 2009).
Currently, the corpus consists of about 200 texts for each of the included CEFR level and for each language. The texts are enhanced with standardized metadata such as the learners L1, test and task descriptions. The devised annotation schemata are being tested in a pilot manual annotation to validate their coverage and validity for all three languages. Challenging aspects are detected and improved and annotation guidelines are being adapted consequently.

References

[CEFR 2001] Council of Europe (2001) *The Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Granger, Sylviane & Bestgen, Yves (2011) Categorizing spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life Long Learning* 21(2–3): 235–252.

Housen, Alex & Kuiken, Folkert (2009) Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4): 461–473.

Hulstijn, Jan H (2007) The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91: 663–667.

Lu, Xiaofei (2011) A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45(1): 36–62.

Mellor, Andrew (2011) Essay Length, Lexical Diversity and Automatic Essay Scoring. *Memoirs of the Osaka Institute of Technology*, Series B, 55(2): 1–14.

Müller, Christoph & Strube, Michael (2006) Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, Sabine & Kohn, Kurt & Mukherjee, Joybrato (eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods* (pp. 197–214). Frankfurt: Peter Lang.

North, Brian (2000) *The Development of a Common Framework Scale of Language Proficiency*. Oxford: Peter Lang.

Ortega, Lourdes (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4): 492–518.

Read, John (2007) Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies* 7(2): 105–125.

Rimrott, Anne & Heift, Trude (2008) Evaluating automatic detection of misspellings in German. Language Learning & Technology 12(3): 73-92.

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44-49). Manchester.

Vajjala, Sowmya & Meurers, Detmar (2012) On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In: Tetreault, Joel & Burstein, Jill & Leacock, Claudia (eds.) *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7) at NAACL-HLT* (pp. 163-173). Montreal, Canada: Association for Computational Linguistics,.

Wolfe-Quintero, Kate & Inagaki, Shunji & Kim, Hae-Young (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Zeldes, Amir & Ritz, Julia & Lüdeling, Anke & Chiarcos, Christian (2009) Annis: A search tool for multi-layer annotated corpora. In: Mahlberg, Michaela & González-Díaz, Victorina & Smith, Catherine (eds.) *Proceedings of Corpus Linguistics 2009*. Liverpool.

The use of *that* in noun complement and relative clauses in learner corpora

BALLIER, Nicolas ; KANTÉ, Issa
Université Paris 7 – Diderot ; Université Paris 13
nballier@free.fr ; kanteissa2609@yahoo.fr

That as a complementizer and a relativizer is one of the most common ambiguous grammatical categories in learners' syntactic analyses. In using *that*, they encounter series of syntactic and semantic difficulties in which their L1 and L2 are significant factors. Previous studies (like Nomura 1993 and Khalifa 2004) focused on difficulties for learners to distinguish *that* relativizer from *that* complementizer in syntactic analyses and constructions.

Using part of the LONGDALE project[1] (Longitudinal Database of Learner English), we extracted data from 110 files containing about 68,200 words to examine the use of *that* in learners' discourse. In our study, we explore how effectively learners use *that* in noun complement and relative clauses. Using AntConc[2], we have analyzed the syntax and the frequency of these two structures. The data show that for the expression of restrictive specification, learners use *that* relative clauses than *that* nominal complement clauses more extensively: 90 occurrences of the former vs. 18 of the latter. Among the 18 occurrences, there are 6 with the head noun *fact*, 4 with *problem*, 2 with *thing*, 2 of *impression* and 1 of *feeling, moral, stereotype* and *sense,* (see Schmid 2000, Ballier 2004 / 2007, Kanté 2010 / 2011, for structural variants and a discursive analysis of noun complement clauses).

In this paper, we firstly show that when we address the issue in learners' discourse, the syntactic distinction is not necessarily the main issue, but rather the tendency to use a limited number of head nouns and to resort to L1 syntactic pattern transfers. In using the complementizer *that*, the challenge for learners seem to be more a matter of restricted usage rather than struggling with errors. In other words, learners tend to use a limited number of complement clauses with a very limited number of head nouns such as *fact*, *impression, problem, feeling,* etc. Most of the patterns, as far as the noun and the subordinator are concerned, are collocationally correctly used. The few errors are mostly found at other syntactic levels.

Secondly, along with Biber and Reppen's (1998: 148-150) observation of transfer effects between L1 and L2 in the use of **verbal *that*-clauses**, we observe a similar phenomenon with ***that* noun complement clauses** too. For instance, the head noun *moral* in the example below is very significant in this respect, since it is almost a direct translation of the French left dislocation construction (*la morale de l'histoire* (*c'*)*est que*, example 1). Interestingly, instead of using the noun *story* in English the speaker uses *history,* which is rather unexpected in this context.

(1) ... and some people (er) conduce conduces us to a an hospital and (em) (er) **the moral (er) in (er) that history is that** (er) today I have my (er) my (er) driving li= licence and (er) I know that (er) that (er) roads are very very dangerous [DID0080-S001.txt, L.765]

[1] LONGDALE is an international project involving ten partner universities around the world (https://www.uclouvain.be/en-cecl-longdale.html) / http://www.clillac-arp.univ-paris-diderot.fr/projets/longdale)

[2] http://www.antlab.sci.waseda.ac.jp/antconc_index.html

In a contrastive perspective, the comparison with a native corpus will allow us to further analyze this phenomenon and find out more semantic-syntactic sources of errors.

Thirdly, in addition to transfer effects between L1 and L2, we further hypothesize that L2 learners seem to restrict some nouns for *that*-nominal clauses and others for relative clauses. Even though any noun can basically be used as an antecedent for relative clause, L2 learners tend to reserve common head nouns like *impression, problem, feeling, story,* etc. to *that*-complement clauses. They barely use those nouns as a relative antecedent even though they are not subject to any selectional constraint in this construction. In our data, the only case in which the same noun complement *that*-taking noun occurs in a relative structure is the noun *thing* (cf. Aijmer 2004).

(2) <B>(em) **the first thing is that** I travel a lot with my parents<B/> <A>(uhu) (uhu)<A/> <B>so I enjoy speaking speaking in english<B/> [DID0028-S001.txt, L.225]

(3) <B>(em) (..) because (em) it's maybe **the only thing that** I like I don't (laugh) I don't really like mathematic... [DID0026-S001.txt, L. 215]

Finally, as Biber and Reppen (1998: 151-156) observe "lexical associations" in ***that*-clause controlled by a verb** in learners discourse, a similar phenomenon is reflected in recurrent collocational patterns for the use of ***that* noun complement clauses**.

References

Aijmer, Karin 2004. "The fact is – an emergent discourse marker?" In *An International Master of Syntax and Semantics, papers presented to Aimo Seppänen on the occasion of his 75th birthday*. *Gothenburg studies in English,* n° 88, pp. 1-9.

Anthony, Laurence 2005. AntConc: "A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning", pp. 7-13.

Ballier, Nicolas 2004. "Deverbal nouns as heads of noun complement clauses in English". Paper given at the international conference on deverbal nouns, Université Lille 3, France, 23-25 September 2004. In *Dossier en vue de l'Habilitation à diriger des recherches,* Université de Paris X, volume 2 *Travaux*, pp. 4-35.

Ballier, Nicolas 2007. "La complétive du nom dans le discours des linguistes". In D. Banks (éd.), *La coordination et la subordination dans le texte de spécialité*. Paris : l'Harmattan, 55-76.

Biber, Douglas and Reppen, Randi 1998. "Comparing native and learner perspectives on English grammar: a study of complement clauses". In Granger (ed.) *Learner English on Computer*. London: Addison Wesley Longman, pp. 145-158.

Kanté, Issa 2010. "Mood and modality in finite noun complement clauses: A French-English contrastive study." In M. Stefania; K. Heylen and G. De Sutter (eds.), *Corpus Studies in Contrastive Linguistics*. *IJCL* 15: 2. Amsterdam: John Benjamins, 267-290. Republished in *Benjamins Current Topics 43,* 2012, pp. 117-140.

Kanté, Issa 2011. *La complétive nominale finie entre syntaxe et sémantique : une étude contrastive anglais-français.* Unpublished PhD, Université Paris 13.

Khalifa, Jean-Charles 2004. *Syntaxe de l'anglais : Théories et pratique de l'énoncé complexe au concours.* Paris : Ophrys.

Nomura, Masuhiro 1993. "The semantics of the content clause construction in English". *English Linguistics,* 10, 184-210.

Schmid, Hans-Jörg 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition.* Berlin: Mouton de Gruyter.

Investigating the influence of multi-L1 learner corpora variables on native language identification

Julian Brooke and Graeme Hirst
Department of Computer Science, University of Toronto
{jbrooke,gh}@cs.toronto.edu

Native language identification (NLI) is the task of automatically identifying the L1 background of a non-native writer (Koppel et al. 2005). NLI is receiving an increasing amount of attention both as a natural language processing (NLP) text classification task (Swanson and Charniak, 2012; Wong et al., 2012) as well as an approach to studying language transfer within the field of second language acquisition (Jarvis and Crossley, 2012). The standard method for NLI is to apply some form of machine learning, i.e. training statistical classifiers on multi-L1 learner corpora. Until recently, nearly all NLI research relied entirely on the well-known International Corpus of Learner English or ICLE (Granger et al., 2009), in part because there have been few other multi-L1 corpora available. However, some NLI researchers have suggested that training and testing in a single corpus is problematic due to within-corpus biases (Brooke and Hirst, 2012), and cross-corpus testing (that is, training on one corpus and testing on another) has recently become a fairly standard alternative (Tetreault et al., 2012; Bykh and Meurers, 2012). At the same time, a number of other (English L2) multi L1-corpora have become available to researchers, including the First Certificate of English (FCE) portion of the Cambridge Learner Corpus (Yannakoudakis et al. 2011), the Lang-8 web journal corpus (Lang-8) (Brooke and Hirst 2012), the International Corpus of Crosslinguistic Interlanguage (ICCI) (Tono et al. 2012), the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa 2011), and the TOEFL11 corpus (Tetreault et al., 2012). There are significant differences, however, across these various corpora with regard to variables such as topic, genre, and learner proficiency, and the cross-corpus research done so far suggests that these variables may be having a significant effect on NLI performance; in this paper we will make use of the variety of new corpora and the cross-corpus approach to isolate the influence of particular corpus variables on the NLI task.

We first provide analysis of our six multi-L1 corpora with respect to the variables we are interested in, in particular topic, genre, and learner proficiency. With regards to topic, our corpora range from those which are very tightly controlled for a very small range of topics (e.g. the ICNALE, which has just two topics) to one with no such restrictions (the Lang-8). Most of the corpora are college-level texts, but three of them (TOEFL11, FCE, and ICNALE) are tagged for proficiency differences, and the ICCI includes much younger students. Essays are the most popular genre, but across the corpora there is a clear distinction between argumentative and descriptive essays, and two corpora (the FCE and the Lang-8) have a very different composition, with the former in particular consisting mostly of letters and stories. Of practical interest here are also the L1s represented, since we cannot compare corpora which do not have overlapping L1s; fortunately, all the corpora represent students from Japan and China, and there are much larger overlaps for some smaller groupings of corpora.

Our central methodology is to select sets of corpora where one of two (or more) training corpora has the same property as a third test corpus while the second training corpus is markedly different; we then perform cross-corpus NLI to see if the difference seems to be having an effect on performance. For example, to test the effect of genre, we use the ICLE as a test set and compare the effectiveness of training in the TOEFL11 (same genre) to the FCE (different genre). Variables such as text length and overall token count are strictly controlled for, and other variables as much as is possible within the limitations of the corpora. For features, we use two well-established options: raw lexical *n*-grams (unigrams and bigrams) and abstracted POS/function word *n*-grams (where lexical words are replaced with their parts-of-speech). We found clear effects for all three of the variables. We were surprised to find that differences in proficiency appear to have greater effect on performance than genre. For instance, when we test in the ICLE, we find that by far the best training set is the very similar TOEFL corpus, followed by the FCE, which differs only in genre, followed by the ICCI, which differs in proficiency, and finally the Lang8, which differs in both genre and proficiency. But when testing in the FCE, the ICNALE showed much worse performance than the Lang-8; its lack of variety in topic seems to severely limit its effectiveness as an NLI training corpus, at least on its own. If we consider the possibility of combining multiple corpora for training, we find that only the ICCI consistently lowers performance when included, perhaps because its low-proficiency European texts confuse the classifier, which may be relying partially on proficiency to discriminate Asian and European texts.

References

Brooke, J. and Hirst, G. (2012). Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING ’12). Mumbai, India.

Bykh, S. and Meurers, D. (2012). Native language identification using recurring *n*-grams – investigating abstraction and domain dependence. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING ’12). Mumbai, India.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, and K. Poonpon, editors, *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow, UK.

Jarvis, S., and Crossley, R. (eds.) (2012). *Approaching Language Transfer through Text Classification*. Multilingual Matters.

Koppel, M., Schler, J., and Zigdon, K. (2005). Determining an author’s native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (KDD ’05), pages 624–628, Chicago, Illinois, USA.

Swanson, B. and Charniak, E. (2012). Native language detection with tree substitution grammars. In P*roceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (ACL ’12), Jeju, Korea.

Tetreault J., Blanchard D., Cahill, A., Chodorow, M. (2012). Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING ’12). Mumbai, India.

Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.) (2012). *Developmental and Cross-linguistic Perspectives in Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins.

Wong, S.-M. J., Dras, M., and Johnson, M. (2012). Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL ’12), Jeju, Korea.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189, Portland, Oregon.

Phraseological units and second language acquisition: A corpus study on learner Finnish

Brunni, Sisko
University of Oulu
sisko.brunni@oulu.fi

The present paper discusses lexical priming (Hoey 2004, 2005) in the context of second language acquisition. Lexical priming is a central notion in language acquisition and language use, and the mastery of phraseological units is based on it. Lexical priming refers to the language user's ability to form typical and correct expressions in their native language effortlessly. It involves collocations and colligations, as well as semantic and pragmatic association concerning both cotext and wider context of the word (see also *lexical item* in Sinclair 1998).

Phraseology and lexical priming on the whole pose challenges for language learners. Several studies (Granger 1998; Nesselhauf 2005; Jantunen 2007) have shown that even advanced language learners have great difficulty with nativelike collocations and idiomaticity; many grammatical constructions produced by language learners sound unnatural and foreign. It is also shown (Jantunen & Brunni 2013) that when the focus is on the synthetic languages morphological primings in learner production differ from the ones that native speakers produce.

According to the basic hypotheses of Hoey's postulation the priming process of polysemous words alters from meaning to meaning. Utilising the polysemous verb ANTAA, (e.g. 'to give', 'to pass', 'to allow', 'to assign', 'to offer'), I have used Contrastive Analysis to compare and contrast the phraseological units of learner language of different proficiency levels and native language. In this study the focus is mainly on the frequently occurring phenomenona in the cotext of word. These include collocations, colligations as well as semantic preferences (i.e. every word is primed to occur in association with particular semantic sets). In such a morphophonologically rich language as Finnish, in addition to the examples stated above, morphological priming also takes place: the core and the words in cotext are primed to occur in and therefore favour certain inflectional forms both in the core and in the cotext. All these four phenonenona are analysed using WordSmith Tools version 4 (Scott 2004). With this information and data I have compared it with the reference material produced by native speakers.

The data is a part of the International Corpus of Learner Finnish (Jantunen 2011). The texts are produced by university students from 22 mother tongue backgrounds and different proficiency levels evaluated using the Common European Frame of Reference for Languages. The data is from the following proficiency levels: A1: 1050 tokens, A2: 59 459 tokens, B1: 363 622 tokens, B2: 303 595 tokens, C1: 94 901 tokens and C2: 16 632 tokens. The reference material is from The Finnish Language Text Collection, which is a selection of electronic research material that contains written Finnish from the 1990's (180 million tokens).

The main result of the paper is that the phraseological units in learner's language differ from the ones occurring in native language: 1) the morphological primings differ from each other both in variation and in frequency, 2) in the native language the collocates are more often related to fixed phrases (e.g. ANTAA <YMMÄRTÄÄ> 'to give an impression', <TUKENSA> 'to lend support to sb'), 3) in the learner language colligate <ALLATIVE> case is overused and <A-INFINITIVE> underused and 4) there are some differences in semantic preferences (e.g. the preferences *answer* and *cook* are underused in the learner language). All this relates to the fact that some specific meanings over-represent ('to pass') and same under-represent ('to allow') themselves in the learner language. The second result is that as the level of proficiency increases the phraseological units advance towards the level of native speakers, however, for example, the morphological priming and collocational selection do not reach the same level.

References

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. In A. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145–160.

Hoey, M. (2004). The textual priming of lexis. In G. Aston, S. Bernardini & D. Stewart (eds) *Corpora and Language Learners* (Studies in Corpus Linguistics 17). Amsterdam & Philadelphia: John Benjamins, 21–41.

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

Jantunen, J.H. (2007). Oppijansuomen piirteitä korpusvetoisesti [A corpus-driven study on learner Finnish]. In P. Muikku-Werner, O. Kokko, & H. Remes (eds) *Virsu 3*. *Suomalais-ugrilaisia kohdekieliä ja kontakteja* [Finno-Ugric Target Languages and Contacts] (Studies in Languages 42). Joensuu: University of Joensuu, 69–83.

Jantunen, J.H. (2011) Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi [The International Corpus of Learner Finnish (ICLFI): typology, background variables and annotation]. - *Lähivertailuja 21*, 86–105

Jantunen, J.H .& Brunni, S. (2013): Morphology, lexical priming and second language acquisition: a corpus-study on learner Finnish - S. Granger, G. Gilquin & F. Meunier (eds) (2013) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses Universitaires de Louvain.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: John Benjamins.

Scott, M. (2004*)*. *WordSmith Tools Version 4*. Oxford: Oxford University Press.

Sinclair, J. (1998). The lexical item. In E. Weigand (ed.*)* *Contrastive Lexical Semantics*. Amsterdam & Philadelphia: John Benjamins, 1–24.

High-frequency nouns and their verbal collocates in academic learner writing

Callies, Marcus
University of Bremen, Germany
callies@uni-bremen.de

The study of various kinds of multi-word units (MWUs) or formulaic sequences (e.g. collocations and lexical bundles) has been on the research agenda of learner corpus research since the very beginnings. However, research into the use of MWUs in learner language is characterized by a high degree of heterogeneity in data, methods and findings (Paquot & Granger 2012). Collocations as a special kind of word combination are notoriously difficult to acquire but of crucial importance to achieve native-like command in an L2 in terms of idiomaticity and fluency. Corpus-based research has shown that learners of English as a foreign language (EFL) at all proficiency levels produce far fewer collocations than native speakers, and that interference errors and other forms of deviations persist even at advanced levels (e.g. Nesselhauf 2005; Laufer & Waldman 2011; see also Paquot & Granger 2012 for a recent overview). Corpus studies in the field of English for Academic Purposes (EAP) have also examined the specific lexis and phraseology of that register, and corpora have played a key role in the compilation of academic wordlists to highlight the most frequent items of a general academic vocabulary (e.g. Coxhead 2000; Paquot 2010). MWUs have been found to vary in structure and function across spoken and written registers, as well as disciplines and genres (e.g. Hyland 2008), and therefore, the concept of English for (General) Academic Purposes has been questioned (e.g. Hyland & Tse 2007). On the other hand, there are studies suggesting that, for instance, a considerable number of verbs are shared across disciplines (e.g. Granger & Paquot 2009a, b; Paquot 2010; Schutz 2013). More recently, attention has been drawn to the importance of extending academic wordlists to incorporate collocations and formulae (Coxhead 2008; Durrant 2009; Simpson-Vlach & Ellis 2010; Ackermann, Biber & Gray 2011; Liu 2012),

In contrast to corpus studies in EAP, research on the phraseology of academic writing produced by EFL learners has lagged behind to date. This is largely due to the fact that many existing and widely-used learner corpora are general-purpose corpora that include learner texts of a general argumentative, creative or literary nature. The large majority of these texts do not represent academic writing in a narrow sense because they differ from academic prose in some important aspects (see Callies & Zaytseva 2013). Fortunately, there are now a growing number of language-for-specific-purposes learner corpora that include L2 academic writing (Granger & Paquot 2013, Nesi 2013: 410). This paper investigates collocations in the pilot version of a specialised learner corpus that comprises a variety of academic text types (research papers, abstracts, reading reports and summaries) produced by German EFL learners in university content courses. In contrast to many previous studies that have focused on verb-noun collocations of high-frequency 'light' verbs in general-purpose learner corpora, I will examine the verbal collocates of a set of highly frequent nouns that can be assumed to be part of a general academic vocabulary across disciplines as included in Paquot's (2010) Academic Keywords List: *analysis, (hypo)thesis, experiment, explanation, research, result, study* and *theory*. The findings will be compared to subsets of corpora that contain novice native-speaker writing of a similar kind: the *Michigan Corpus of Upper-Level Student Papers* (MICUSP; Römer & Brook O'Donnell 2011; Brook O'Donnell & Römer 2012) and the corpus of *British Academic Written English* (BAWE; Alsop & Nesi 2009). The results will be discussed in terms of frequency and degree of restriction of the collocations, their L1-L2 congruence and potential effects of cross-linguistic influence.

References

Ackermann, K., Biber, D. & Gray, B. (2011) *An academic collocation list*. Paper presented at *Corpus Linguistics 2011*, 20-22 July 2011, Birmingham/UK.

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

Brook O'Donnell, M. & Römer, U. (2012) From student hard drive to web corpus (part 2): the annotation and online distribution of the *Michigan Corpus of Upper-level Student Papers* (MICUSP). *Corpora* 7(1): 1-18.

Callies, M. & E. Zaytseva (2013) The *Corpus of Academic Learner English* (CALE) – A new resource for the assessment of writing proficiency in the academic register", *Dutch Journal of Applied Linguistics* 2(1), 126-132.

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

Coxhead, A. (2008) Phraseology and English for academic purposes. In: Meunier, F. & S. Granger (eds.) *Phraseology in Foreign Language Learning and Teaching* (pp. 149-161). Amsterdam: Benjamins.

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

Granger, S. & Paquot, M. (2009a) Lexical verbs in academic discourse: a corpus-driven study of learner use. In: Charles, M., D. Pecorari & S. Hunston (eds.) *Academic Writing. At the Interface of Corpus and Discourse* (pp. 193-214). London & New York: Continuum.

Granger, S. & Paquot, M. (2009b) In search of a general academic vocabulary: A corpus-driven study. In: Katsampoxaki-Hodgetts, K. (ed.) *Options and Practices of LSP Practitioners* (pp. 94-108). University of Crete Publications.

Granger, S., & Paquot, M. (2013) Language for specific purposes learner corpora. In: Chapelle, C.A. (ed.) *The Encyclopedia of Applied Linguistics* (pp. 3142-3146). New York: Blackwell.

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

Hyland, K. & Tse, P. (2007) Is there an 'academic vocabulary'? *TESOL Quarterly* 41(2): 235-253.

Laufer, B. & Waldman, T. (2011) Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2): 647-672.

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

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

Nesi, H. (2013) ESP and corpus studies. In: Paltridge, B. & S. Starfield (eds.) *The Handbook of English for Specific Purposes* (pp. 407-426). Malden/MA: Wiley-Blackwell.

Paquot, M. (2010) *Academic Vocabulary in Learner Writing*. London: Continuum.

Paquot, M. & Granger, S. (2012) Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32: 130-149.

Römer, U. & Brook O'Donnell, M. (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.

Schutz, N. (2013) How specific is English for Academic Purposes? A look at verbs in business, linguistics and medical research articles. In: Andersen, G. & K. Bech (eds.) *English Corpus Linguistics: Variation in Time, Space and Genre* (pp. 237-257). Amsterdam: Rodopi.

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

Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis

Campillos-Llanos, Leonardo
Computational Linguistics Laboratory
Autonomous University of Madrid
leonardo.campillos@uam.es; leonardo.campillos@gmail.com

This study aims at understanding the acquisition of oral expression by different groups of learners of Spanish at low-intermediate level: A2 (N=20) and B1 (N=20) (Council of Europe 2001). For that goal, their *interlanguage* (Selinker 1972) was analyzed (both the use of categories and the errors made). The data collection method is a semi-structured interview, and a total of forty (N=40) learners were interviewed. Participants were clustered into ten groups according to their mother tongues (henceforth, L1), each group gathering four learners. There are nine homogeneous groups whose L1 is Portuguese, Italian, French, English, Dutch, German, Polish, Chinese, and Japanese. A heterogeneous group of participants have other L1s (Hungarian, Korean, Finnish, and Turkish). Four interviews with native speakers (control group) were also gathered as a benchmark to compare oral phenomena in both groups (e.g. gender agreement, use of prepositions, or discourse markers). Further details are explained in Campillos-Llanos (2011; 2012).

One of the justifications of the study is the absence of rigorous research about the acquisition of oral expression at low-intermediate levels. In particular, the objectives of the research were to:

- Specify those points that are more problematic for particular groups of students in our corpus (e.g. the use of the article by Chinese or Polish learners), and what degree of difficulty they pose, with a view to improving pedagogical materials.
- Understand in what degree certain contents of the acquisition of oral expression are difficult for every learner (e.g. the use of prepositions or personal pronouns).
- Fulfil the lack of computerised resources for Learner Corpus Research. The interface for the corpus is available at: http://cartago.lllf.uam.es/corele/index.html

The analysis examines Grammar, Lexis, Pronunciation and Pragmatics-Discourse, with the intention of understanding the acquisition of different linguistic levels. Even though only using an oral interview to gather the data may have hindered the rigorous analysis of certain aspects, a global approach was chosen. Departing from Corder's Error Analysis procedure (1971) and the methodology of Computer-aided Error Analysis (Dagneaux, Dennes & Granger 1998), the interviews were manually transcribed and computationally processed. The errors were categorised according to an error taxonomy that was designed considering previous error classifications (Granger 2003; Nicholls 2003; Tono 2003; Díaz-Negrillo & Fernández-Domínguez 2006). In order to normalise the error frequencies, word counts for each morphological category were obtained, which enabled to perform a Contrastive Interlanguage Analysis (Granger 1996; Gilquin 2008) regarding errors and usage of categories between learners and native speakers.

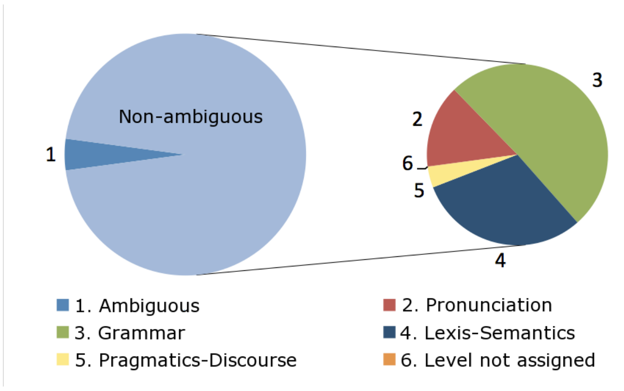
The analysis of the corpus unveiled the following difficulties:

- Progress from A2 to B1 shows a diminution of errors, with a mean (M) of 191.30 non-ambiguous errors at A2 (SD=99.82), and M=135,40 at B1 (SD=53.25). Still, these data only partially reflect the acquisition process, since they can be related, for example, to the avoidance of difficult structures. Likewise, learners at upper levels would be expected to make more errors, as they are trying to practise new structures.
- Errors most frequently affect Grammar (48.61%) and Lexis (29.37%), and there are fewer errors in Pronunciation (14.19%) and Pragmatics-Discourse (3.58%) (Figure 1 and Table 1). Around 4.45% are ambiguous, and 49.21% would be interference errors.
- Lexical errors are more abundant in formal than in semantic aspects (which would be more difficult to acquire). At A2, the most frequent errors are misformations and borrowings. Despite errors decrease at B1, other deviations persist (e.g. semantic relation and gender).
- The most frequent and generalised grammar errors affect articles, sentence structure, agreement, and past tense. Errors registered in pronouns, prepositions or subordination tend to persist at B1. The characteristics of spoken discourse may explain the high number of omission, agreement and word order errors, and the overuse of present tense.
- Interference phenomena tend to strongly persist in B1 learners' pronunciation, where maybe the L1 has the strongest influence—although learners from every language background commit certain errors (e.g. the articulation of /r/).
- Errors related to Pragmatics-Discourse show a wide individual variability, maybe due to the fact that every learner's rhetoric skills in his/her L1 cause the results.

Table 1. Relative frequency (%) of errors in each level linguistic

Errors		**Total**	
	Linguistic level	**Total**	**(%)**
Non-ambiguous regarding the linguistic level	Pronunciation	970	14.19%
	Grammar	3324	48.61%
	Lexis-semantics	2008	29.37%
	Pragmatics-Discourse	245	3.58%
	Not classified	3	0.04%
Ambiguous	-	288	4.21%
Total		6838	

Figure 1. Ambiguous and non-ambiguous errors (these are broken down into linguistic levels)



There exist several limitations in our study. Using only oral data, it is difficult to diagnose the type or the linguistic level of certain deviations, or whether they are due to competence or performance. Moreover, with a small number of participants in each L1 group, and only from low-intermediate level, neither we can generalise results nor we can infer conclusions as to the possibility of acquiring an almost bilingual proficiency. Yet some results are similar to error analyses of written learner corpora of Spanish (Fernández-López 1997) and English (Díez-Bedmar 2011), in which the most frequent errors affected Grammar (especially, articles, verbs, and pronouns), followed by Lexis.

Acknowledgements

This project was founded by Madrid Regional Government and the European Social Found (doctoral grant).

References

Campillos-Llanos, L. (2011) A XML-tagged Spanish Learner Oral Corpus for Learner Corpus Research. Poster presented at *Learner Corpus Research 2011 Conference. Louvain-la-Neuve, Belgium.* www.lllf.uam.es/ESP/Publicaciones/XMLLearner.pdf

Campillos-Llanos, L. (2012) *La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus.* PhD Thesis. Autonomous University, Madrid

Council of Europe. (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge Cambridge University Press.

Corder, P. (1971) Idiosyncratic Dialects and Error Analysis. *International Review of Applied Linguistics*, 9(2): 147-60.

Dagneaux, E., Dennes, S., & Granger, S. (1998) Computer-aided error analysis. *System*, 26(2): 163-174.

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

Díez-Bedmar, Mª. B. (2011) Spanish pre-university students' use of English: CEA results from the University Entrance Examination. *International Journal of English Studies,* 11(2): 141-158.

Fernández López, S. (1997) *Interlengua y Análisis de Errores en el aprendizaje del español como lengua extranjera*. Madrid: Edelsa.

Gilquin, G. (2008) Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In Gilquin, G., Papp, S. & Díez-Bedmar, Mª. B. (eds.) (2008) *Linking up contrastive and learner corpus research* (pp. 3-33). Amsterdam/New York, NY: Rodopi. Studies in Practical Linguistics Series, 66.

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

Granger, S. (2003) Error-tagged Learner Corpora and CALL: a promising synergy, *CALICO Journal*, 20(3): 465-480

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for Lexicography and ELT. In Archer *et al.* (eds) *Proceedings of the Corpus Linguistics 2003 Conference (CL2003)*. Lancaster University (pp. 572-581).

Selinker, L. (1972) Interlanguage. *International Review of Applied Linguistics*, 10(3): 209-301.

Tono, Y. (2003) Learner corpora: design, development and applications. *Proc. of the 2003 Corpus Linguistics Conference*. Lancaster University (pp. 800-809). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.6849&rep=rep1&type=pdf

MAZEA: Multi-label Argumentative Zoning for English Abstracts

Candido Jr, Arnaldo[1]; Dayrell, Carmen[2]; Schuster, Ethel[3]; and Aluísio, Sandra[1]
[1] NILC/University of São Paulo, [2] UNINOVE, [3] Northern Essex Community College
arnaldoc@icmc.usp.br, carmengc@uninove.br, eschuster@necc.mass.edu, sandra@icmc.usp.br

The analysis of *rhetorical moves* has proven to be useful in genre-based pedagogies since they play a key role in text organization and structuring. By *move*, we refer to "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004: 228). By teaching learners to recognize the move structure of their target genre and the key linguistic features within each move, we can provide them with a better understanding of how genres are organized and of the language they are expected to write.

This paper focuses on abstracts of research papers written in English. Abstracts are of special interest for their relevance in most academic contexts, even where English is not the official language, as it is the case of Brazil. However, as Swales & Feak (2009:xiii) point out, constructing an efficient, clear abstract is a fairly difficult task, even for experienced and widely published writers. Abstracts are 'highly polished and condensed texts' (Gledhill 2005:41) in which authors have to capture readers' interest and convince them about the relevance and main claims of the paper (Hyland & Tse 2005).

One challenge in the study of rhetorical moves is that manual annotation of large corpora tends to be extremely time-consuming. Within English for Academic Purposes (EAP), many systems were developed to automatically identify rhetorical moves in scientific texts. For abstracts, although categories may vary from one system to another, these are usually considered: (i) background, (ii) gap, (iii) purpose, (iv) method, (v) result, and (vi) conclusion. Examples of systems that identify rhetorical moves in English abstracts include: Anthony & Lashkia, (2003), Mcknight & Arinivasan (2003), Shimbo *et al.* (2003), Ito *et al.* (2004), Yamamoto & Takagi (2005), Wu *et al.* (2006), Lin *et al.* (2006), Genovês *et al.* (2007), Ruch *et al.* (2007), Hirohata *et al.* (2008). These systems work at the sentence level and establish a one-to-one relationship between sentences and moves, that is, a given sentence can only match one single move. If there is more than one move in the sentence, the most prominent move is considered. However, this approach has a fundamental limitation: it fails to adequately reflect actual language use since a move is "is a functional, not a formal, unit" that can be realized by a clause, a sentence, or even several sentences (Swales (2004:229).

In an attempt to address this critical issue, we developed MAZEA (*Multi-label Argumentative Zoning for English Abstracts*). It is a machine learning classifier which automatically identifies rhetorical moves in English abstracts, enabling a given sentence to be assigned to as many categories as appropriate. This includes "no label" for those cases when the classifier cannot decide which category the segment refers to. Dayrell *et al.* (2012) presents a full description of the system, including its training corpora and the process of manually annotating all abstracts, working environment (algorithms), and a discussion on our tests and the system's performance.

In its initial version, MAZEA focused on two broad fields: physical sciences and engineering (PE) and life and health sciences (LH). Its overall performance has been regarded as reasonably satisfactory, considering that MAZEA is the first of its kind. The best classifier reached 69% of chance of assigning the correct category to a given sentence or part of it. Such level of accuracy is much higher than our baselines: the expected accuracy for a random categorization would be 16.66% and, if the most frequent move (method) was assigned, in a mono-label classification, the level of accuracy would be 33.7% for the two corpora altogether. Multi-label sentences accounted for 16.5% of all LH sentences (1,082 out of 6,544) and for 11.3% of all PE sentences (445 out of 3,933).

MAZEA has been successfully used in English academic writing courses as a pedagogical resource tool. Learners are first asked to write an abstract for their research projects and then identify the moves they have used and how they were organized. These abstracts are then submitted to MAZEA and learners can compare their categorization with the system's. The idea is to have students reflecting on the move structure of their own abstracts and raise their awareness of key aspects related to text organization.

In this demo presentation, we introduce a refined version of MAZEA, adapted to different disciplines with different conventions and ways of expressing their ideas and arguments. Thus, the system now processes abstracts based on the discipline it matched to, as defined by the user. The disciplines included are: biology, bioengineering, biophysics, computing, dentistry, earth sciences, electrical engineering, industrial engineering, mechanical engineering, pharmaceutical sciences, and physics. It is important to mention that an online demo version of MAZEA will be made publicly available together with our annotated corpora.

References

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

Dayrell, C., A. Candido Jr., G. Lima, D. Machado Jr., A.Copestake, V. D. Feltrim, S. Tagnin and S. Aluísio (2012) 'Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora'. *Proceedings of LREC 2012 (The 8th International Conference on Language Resources and Evaluation)*, Istanbul (Turkey), 21-27 May 2012. Available at -http://www.lrec-conf.org/proceedings/lrec2012/pdf/734_Paper.pdf

Genovês Jr., L.; Feltrim, V.D.; Dayrell, C. and Aluísio, S. (2007). Automatically detecting schematic structure components of English abstracts. In *Proceedings of the RANLP´2007, Workshop on Natural Language Processing for Educational Resources*. Borovets, Bulgaria, pp. 23--29.

Gledhill, C. (2005), *Collocations in Science Writing.* Tübingen: Gunter Narr Verlag.

Hirohata, K.; Okazaki, N.; Ananiadou, S. and Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008).* Asian Federation of Natural Language Processing, pp. 381--388.

Hyland, K. & P. Tse (2005), 'Hooking the reader: A corpus study of evaluative *that* in abstracts'. *English for Specific Purposes*, 24: 123-139.

Ito, T.; Simbo, M.; Yamasaki, T. and Matsumoto, Y. (2004). Semi-supervised sentence classification for medline documents. In *IPSJ SIG Technical Report*, v. 2004-ICS-138, pp. 141--146.

Lin, J.; Karakos, D.; Demner-fushman, D. and Khudanpur, S. (2006). Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, Association for Computational Linguistics (ACL), pp. 65--72.

McKnight, L., Arinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA 2003 Symposium Proceedings*, American Medical Informatics Association, pp. 440--444.

Ruch, P.; Boyer, C.; Chichester, C.; Tbahriti, I.; Geissbuhler, A.; Fabry, P.; Gobeill, J.; Pillet, V.; Rebholz-Schuhmann, D.; Lovis, C. and Veuthey, A. L. (2007). *Using argumentation to extract key sentences from biomedical abstracts*. International Journal of Medical Informatics, v. 76, pp. 195--200.

Shimbo, M.; Yamasaki, T. and Matsumoto, Y. (2003). Using sectioning information for text retrieval: a case study with the medline abstracts. In *Proceedings of the 2nd International Workshop on Active Mining (AM'03)*, Maebashi, Japan; 2003 , pp. 32--41.

Swales, J. (2004). *Research Genres: Exploration and applications.* Cambridge University Press, Cambridge.

Wu, J.; Chang, Y.; Liou, H. and Chang, J. S. (2006). Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 41--44.

Yamamoto, Y., Takagi, T. (2005). A sentence classification system for multi-document summarization in the biomedical domain. In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDEW'05), IEEE Computer Society Washington, DC, USA,* pp. 90--95.

Economy is a living organism: metaphorical expressions in a learner corpus of English

Castaño, Emilia; Verdaguer, Isabel; Ventura, Aaron; Laso, Natalia Judith
University of Barcelona
e.castano@ub.edu; i.verdaguer@ub.edu; venturaa@ub.edu; njlaso@ub.edu

Cognitive linguistics has shown that metaphorical language is pervasive in everyday conventional language (Lakoff & Johnson 1980; Kövecses 2005), as human beings usually understand and conceptualize abstract concepts (target domain) in terms of concrete or more structured domains (source domain) through mappings that take the form TARGET DOMAIN IS SOURCE DOMAIN. In this view, metaphors are more than stylistic devices; instead, they are cognitive templates that license an open-ended number of metaphorical expressions, which are the surface realization of cross-domain mappings (Lakoff 1993).

One particular field where we can find a wealth of metaphors is business discourse. Newspapers and magazines strongly rely on metaphorical expressions to describe the state of the economy, markets and business issues in general. Thus, it is not strange to find that business is presented as a WAR in which companies face an economic battle, and the economy and markets as LIVING ORGANISMS that may grow and develop or fall sick and face a slow recovery.

While there are studies dealing with metaphorical language in business English (Burcea 2010; Kovács 2006; White 2003; Kövecses 2002; Eubanks 1999), less attention has so far been devoted to metaphors in the language of learners or nonexperts (Castaño *et al.* 2011; Chapetón 2010; Chapetón & Verdaguer 2012; Golden 2012). The use of metaphors in a learner corpus of business English is, thus, worth investigating. The aim of this paper is to explore if learners also use metaphors in business English and, if so, what metaphorical conceptualizations are the most frequent.

The learner corpus analysed in the present study consists of a sample of 34 essays (30,000 tokens, approximately) written by undergraduate students of Economics, as part of the course "Business English II". Their level of proficiency in English was B1, according to the Common European Framework of Reference for Languages (CEFR). Learners were allowed to write a free essay on a business-related topic. A qualitative analysis of these essays was carried out by first identifying what metaphorical expressions (if any) were commonly used and then they were classified and compared against those already identified in the literature based on native production. Finally, networks of relations were established among them, on the basis of the ontological correspondences that structure the metaphorical conceptualisations identified in business texts.

Our research has shown that learners do make use of metaphorical language as conceptual metaphor is a basic cognitive process, in line with previous research on cognitive linguistics (Lakoff 1993). The metaphorical expressions identified in our learner corpus can be grouped into four main conceptual metaphors: ECONOMY/BUSINESS IS A LIVING ORGANISM, BUSINESS IS WAR, BUSINESS IS A RELATIONSHIP, and ECONOMIC SUCCESS AND FAILURE ARE MOVEMENTS ON A VERTICAL AXIS, which match those used by native speakers (Kövecses 2002; White 2003; Kovács 2006; Burcea 2010; among others). In addition, we found that even in learner's language metaphors are connected and establish larger networks within the same text, which contribute to enhancing the unity of the text, as pointed out by Semino (2008). So we believe that the nature of these relationships as well as the discourse functions of metaphorical language should be further explored.

Finally, we suggest the need to develop pedagogical and lexicographic applications which increase the speakers' and learners' awareness of the potentiality of metaphorical language given its pervasive use. Not all languages have the same metaphorical models. As Kövecses (2002) states, at a generic level a given metaphor is very similar across cultures. However, at a specific level we can notice important cross-cultural differences. A cognitive approach to teaching English for specific purposes would thus help learners from diverse cultural backgrounds to become aware of any cross-linguistic differences that they may encounter; avoiding thereby problems in linguistic choices and possible errors in the L2.

References

Burcea, Raluca (2010) Metaphorical conceptualizations of marketing. *Annals of Spiru Haret University, Economic Series* 1(3): 153-160.

Castaño, Emilia; Hilferty, Joe; Verdaguer, Isabel; Comelles, Elisabet & Laso, Natalia Judith (2011) The metaphorical basis of coherence relations, paper presented at the *Learner Corpus Research 2011*. Louvain-la Neuve (15-17 September 2011).

Chapetón, Marcela (2010) Metaphor in argumentative writing, a comparative corpus-driven study (Unpublished doctoral dissertation). Barcelona: University of Barcelona.

Chapetón, Marcela & Verdaguer, Isabel (2012) Researching Linguistic Metaphor in native, Non-Native, and Expert Writing. In: MacArthur, Fiona, José Luis Oncins-Martínez, Manuel Sánchez-García & Ana María Piquer-Píriz (eds.) *Metaphor in Use: Context,Culture and Communication* (pp. 149-174). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Eubanks, Philip (1999) Conceptual metaphor as rhetorical response. A reconsideration of metaphor. *Written* Communication 16: 171–99.

Golden, Anne (2012) Metaphorical expressions in L2 production. Researching Linguistic Metaphor in native, Non-Native, and Expert Writing. In: MacArthur, Fiona, José Luis Oncins-Martínez, Ana María Piquer-Píriz & Manuel Sánchez-García (eds.) *Metaphor in Use: Context, Culture and Communication* (pp. 135-148). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kovács, Eva (2006) Conceptual Metaphors in Popular Business Discourse. *Publicationes Universitatis Miskolcinensis. Sectio Philosophica* Tomus XI (Fasciculus 3): 69-80.

Kövecses, Zoltan (2002) *Metaphor: A practical introduction*. New York: Oxford University Press.

Kövecses, Zoltan (2005) *Metaphor in culture: Universality and variation.* Cambridge, UK: Cambridge University Press.

Lakoff, George (1993) The contemporary theory of metaphor. In: Ortony, Andrew (ed.) *Metaphor and Thought* (pp. 202-251), 2ª ed. Cambridge, UK: Cambridge University Press.

Lakoff, George & Johnson, Mark (1980) *Metaphors we live by*. Chicago: University of Chicago Press.

Semino, Elena (2008) *Metaphor in Discourse*. Cambridge: Cambridge University Press.

White, Michael (2003) Metaphor and economics: the case of growth. *English for Specific Purposes Journal* 22: 131-151.

Semantic and collocational behaviours of phrasal verbs in Chinese learners' English writing and native English novice writing: A multi-corpora approach

Chen, Meilin

City University of Hong Kong

meilinchen8388@gmail.com

Phrasal verbs (PHVs) are perceived as notoriously difficult for ESL/EFL learners because they are semantically non-compositional, very often polysemous, and syntactically more flexible than other types of phraseological patterns (e.g. variation of particle positions and pronoun or noun insertions are allowed in PHVs.) Previous studies repeatedly found that regardless of their L1 background, learners tend to avoid using PHVs 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 another complex picture of learners' use of PHVs. The avoidance observed in SLA studies is not always found in learners' actual writing. Waibel (2007), for instance, found that German EFL learners use more PHVs than native speakers, while leaners of other L1s (e.g. Italian, Spanish, Swedish, French) use fewer.

This study explores Chinese learners' use of PHVs in comparison with their American and British counterparts. The PHVs in a learner corpus of argumentative essays compiled by the author (188,628 words) was compared to that in four native novice corpora. The first two native corpora include argumentative essays written by American and British novice writers taken from: the *Louvain Corpus of Native English Essays* (LOCNESS-US) and the *General Studies* corpus (GS-UK) (Milton 2001) respectively. The second two native corpora consist of academic papers from the *Michigan Corpus of Upper-Level Student Papers* (MICUSP) and the *British Academic Written English* (BAWE) respectively.

To investigate the number and type of meanings used by different writers, three PHV dictionaries were used. They are: *Collins COBUILD Dictionary of Phrasal Verbs 2nd Edition* (2002), *Longman Dictionary of Phrasal Verbs* (1983), and *Oxford Phrasal Verbs Dictionary for Learners of English* (2001). The PHVs were extracted from the five corpora by using the WordSmith tool 5.0 (Scott 2008) and then the meanings of the PHVs were looked up in the dictionaries and recorded.

The results show that in comparison with their native English coutnerparts the Chinese learners are not only able to produce a sufficient number of PHVs but also use them in a variety of meanings. The meanings of the PHVs in the five corpora were further divided into figurative meanings and literal meanings. It shows that nearly 80% of the PHV meanings in the learner corpus are figurative. However, this is still lower than those in the native novice corpora. For instance, the figurative use of PHVs in BAWE is almost 90% of the total number.

The analysis of the collocations of PHVs reveals the second gap between the learners' knowledge of PHVs and that of their native counterparts. A number of deviant collocations of PHVs were found in the results. However, no specific factors that contribute to such deviations can be identified. The Chinese learners' choices of collocations are random and arbitrary.

The findings from this study indicate that the acquisition of PHVs may not be as difficult as previous studies proposed; the learners did not show obvious intention to avoid using PHVs in writing. However, difficulties do exist. The learners might have acquried PHV forms, but the semantic complexity is still a big barrier for them. Figurative meanings of PHVs are more difficult than literal ones to acquire. The second challenge for the Chinese learners to master PHVs is the collocational behaviours of PHVs and the ability to use them in an idiomatic way.

Dictionaries

COBUILD Dictionary of Phrasal Verbs (2nd Edition). 2002. Sinclair, J., P. Hanks, G. Fox, R. Moon, & P. Stock (Eds.). Edinburgh: HarperCollins.

Longman Dictionary of Phrasal Verbs. 1983. R. Courtney (Ed.). London: Longman.

Oxford Phrasal Verbs Dictionary for Learners of English. 2001. D. Parkinson (Ed.). Oxford: Oxford University Press.

References

Dagut, Menachem & Laufer, Batia (1985) Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition* 7: 73–79.

Hulstijn, H. Jan & Marchena, Elaine (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition* 11: 241–255.

Laufer, Batia & Eliasson, Stig (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, Yan & Fukuya, Yoshinori (2004) Avoidance of Phrasal Verbs: The Case of Chinese Learners of English. *Language Learning* 54 (2): 193–226.

Milton, John (2001) *Research Reports (V2): Elements of a Written Interlanguage: A Computational and Corpus-based Study of Institutional Influences on the Acquisition of English by Hong Kong Chinese Students*. Hong Kong: Hong Kong University of Science & Technology Press.

Scott, Mike (2008) *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Schmitt, Norbert & Redwood, Stephen (2011) Learner knowledge of phrasal verbs: A corpus-informed study. In: Meunier, Fanny De Cock, Sylvie., Gilquin, Gaëtanelle and Paquot, Magali (eds.) *A Taste for Corpora: In Honour of Sylviane Granger* (pp.173-208).. Amsterdam: John Benjamins,

Waibel, Birgit (2007) *Phrasal verbs in learner English: A corpus-based study of German and Italian students*. PhD thesis. Freiburg: Albert-Ludwigs-Universität Freiburg.

Profiling the Language Competence of Learners of Spanish in Irish Secondary Education

Costello, Kerrill
University College Cork
108223139@umail.ucc.ie

This paper will provide a brief outline of the principal objectives of my own Learner Corpus project, before going on to discuss key aspects of methodological design in greater detail.

The Leaving Certificate (LC) is the national, standardised state examination in Ireland necessary for entry to third level education. With samples of the official LC Spanish examination data, this project will compile a digitised corpus of learner Spanish. The corpus will include samples of the written and oral production of 100 candidates, and will be analysed using a specific investigative corpus technique, Computer-aided Error Analysis (CEA, Dagneaux et al, 1998). The objective of this corpus study is to provide the teaching and linguistic research community with an empirical profile of learner competence that may inform foreign language teaching, materials and examination design as well as curriculum policy.

The investigation will focus on the following areas of inquiry:

(1) What language errors are evident in the corpus in general? (Error counts, frequencies and types, as well as errors in context).

(2) In what instances are learners making these errors? Here, I will compare oral and written production as well as looking at performance across specific tasks. This question will also address other learner specific variables: grade, gender, and examination level.

(3) What are the key factors affecting the development of language competence in the LC Spanish classroom; materials, teaching methods, exercises and activities, and target and native language use.

CEA is a powerful apparatus in that it greatly facilitates the quantification and analysis of a large learner corpus in digital format. CEA is informed by the LC Spanish syllabus (Department of Education and Skills, 1995), and a comprehensive collection of key grammatical criteria (Butt and Benjamin, 2011). Theories of Language Competence are informed by Bachman's framework of Communicative Language Ability (1990) and the Common European Framework of Reference for Languages (2001). As this project is compiling and using a new corpus, methodological design is primarily data driven. Specific research questions as well as the error categorisation scheme are influenced by the data emerging. The error scheme is also influenced by Granger's approach for the FRIDA corpus (2003) in that it is a tri-dimensional taxonomy outlining three levels of error annotation: error domain (lexical, grammatical, and production specific), error category and error type.

Once the annotated data is analysed, it is possible to generate quantitative error statistics via the statistical interface of the UAM Corpus Tool (O'Donnell, 2012), a comprehensive and sophisticated linguistic software package. Tests and measurements will using the include: error counts and frequencies at each level of specification, lexical variety, lexical density, morphosyntactic variation, lemmatisation and feature usage in terms of counts, mean and standard deviation. The UAM Corpus Tool allows recording of candidate-specific variables such as grade, examination level (higher or ordinary), task type and gender. Thus, it will allow critical analysis of the corpus as one unit, as separate written and oral sub corpora and also of performance per task, level and gender.

In addition to the CEA, my study will attend to the social and linguistic setting of learners (following Corder, 1975). Thus, my corpus work will be supported by the analysis of prescribed Spanish text books and classroom materials, as well as data gathered via qualitative interviews with key stakeholders in the field.

The LC is a standardised examination taken by thousands of students in Ireland; this presents a massive, raw corpus of data with the potential to yield invaluable insight into the phenomena of learner interlanguage. This investigation does not make a priori assumptions about the data set, the LC Spanish examination, the context of FLs or of any aspect of learner competence. It undertakes to provide the linguistic research community and the domain of Spanish language learning and pedagogy in Ireland with an empirical, descriptive profile of real learner performance, characterising learner difficulty.

References

Bachman, Lyle F. (1990) Fundamental Considerations in Language Testing Oxford: OUP

Butt, John & Benjamin, Carmen (2011) A New Reference Grammar of Modern Spanish Hodder Education Publishers, 5th Edn

Corder, Stephen Pit (1975) 'Error Analysis, Interlanguage and Second Language Acquisition' Language Teaching 8 (4): 201-218

Council of Europe (2001) Common European Framework of Reference for Languages: Learning, Teaching, Assessment Cambridge: CUP

Dagneaux, Estelle et al (1998) 'Computer Aided Error Analysis' System (26): 163-174

Department of Education and Skills (1995) Leaving Certificate Spanish Syllabus [accessed 20/01/2011] http://www.education.ie/servlet/blobservlet/lc_spanish_sy.pdf?language=EN

Granger, Sylviane (2003) 'Error-tagged Learner Corpora an CALL: A Promising Synergy' CALICO Journal 20 (3): 465-480

O'Donnell, Michael (2012) UAM Corpus Tool Version 2.8.12 (Universidad Autónoma de Madrid) http://www.wagsoft.com/CorpusTool/ [accessed December 2012]

Textual patterns and rhetorical moves in English scientific abstracts: comparing student and published writing

Dayrell, Carmen[1] and Candido Jr., Arnaldo[2]
[1] UNINOVE; [2] NILC/University of São Paulo
carmengc@uninove.br; arnaldoc@icmc.usp.br

Academic communication poses major challenges for novice researchers, who have to come to terms with the conventions adopted by their academic discourse community. For those operating in a foreign language, such task may be even harder. In addition to making the appropriate lexical and syntactical choices in the language in question, they also need to handle cultural differences, since genre practices may vary across languages.

This paper focuses on abstracts of research papers and sets out to investigate recurring textual patterns in abstracts written in English by Brazilian graduate students, and hence native speakers of Portuguese, from various disciplines. The focus on abstracts is motivated by their relevance as gatekeeper in a number of academic activities. Abstracts are a highly condensed form of writing, in which writers need to capture readers' interest and attention as well as convince readers of the relevance of the research and main claims of the paper. These aspects may place heavy demands on learners.

Our primary purpose is to identify similarities and differences between abstracts written in English by Brazilian graduate students vis-à-vis abstracts taken from published papers from the same discipline, with respect to the recurring textual/linguistic patterns. By recurring textual patterns, we refer to chunks of language with some kind of lexical and/or syntactical regularity. In this case, in addition to repeated sequences of words (e.g. *the aim of this study is to*) and recurrent syntactical constructions (such as the passive voice, first person pronoun + verb), we also allow for some kind of variation within chunks, be it in terms of different word-forms of the same lemma (*the aim of this study* ***is/was*** *to*) or lexical choice (*the results* ***show/demonstrate/indicate/suggest/etc*** *that* or *BE* ***investigated/studied/presented/ proposed/etc***). Such textual patterns are examined in relation to the "rhetorical moves" (or communicative stages) in which they occur. With regard to abstracts, the following moves have been suggested in the literature: (i) introduction/background, (ii) gap, (iii) purpose, (iv) methodology, (v) results, and (vi) conclusion. Thus, the idea is to examine whether students and published writers make similar linguistic choices when writing abstracts.

The data are drawn from two corpora of English abstracts from the disciplines of biology, biophysics, computing, dentistry, earth sciences, pharmaceutical sciences, and physics. One corpus is made up of 279 abstracts (55,577 tokens) written in English by Brazilian graduate students (master's and PhD). These abstracts were collected between 2004 and 2010 in courses on English academic writing offered by three Brazilian universities. The other corpus comprises 1,395 abstracts (273,815 tokens) taken from papers published by various leading academic journals from the disciplines in question. It has been designed to be five times larger than the student corpus so as to enable the researchers to reach firmer conclusions concerning the preferred patterns of each discipline. The two corpora contain the same proportion of texts by discipline. For comparison, frequencies are therefore normalized (per 10,000 words).

The identification of rhetorical moves within abstracts was conducted in two stages. We first resorted to the AZEA (*Argumentative Zoning for English Abstracts*) system (Genovês *et al.* 2007) to automatically identify the abovementioned rhetorical moves and annotate abstracts accordingly. This automatic annotation was then manually validated so as to correct potential errors. As for the textual patterns within each move, typical patterns – such as first person pronoun (*We verb (that)*), passive voice, *it* patterns (*it BE* verb *that*), *the results* verb *that, the* noun *of this* noun *BE to* – were identified automatically. The remaining instances were then analysed by means of the *WordSmith Tools, version 6*.

We found relevant differences between student and published writing in relation to their preferred patterns within each move across all disciplines. For example, when stating the purpose of the study, students seem to prefer impersonal constructions such as *this study|paper|article|work VERB* and *the aim|purpose|objective|goal of this study|paper|article| work BE to*. The only exceptions are physics and computing as students tend to opt for first person pronoun (*we propose|describe| present|demonstrate|...*). By contrast, with the exception of earth sciences and dentistry, published authors show a clear preference for first person pronoun (*we*) to state the purpose of their study.

All these patterns can be argued to be acceptable for presenting the purpose of the research. However, our point here is that, overall, students do not seem to comply with disciplinary conventions. Thus, our findings therefore can offer valuable contributions to pedagogic practice. Drawing students' attention to the range of options they have at their disposal and the preferred textual patterns of their disciplines can certainly help them move beyond syntactical structure and improve their ability to use language more effectively.

Reference

Genovês JR., L., V. D. Feltrim, C. Dayrell, S. Aluísio (2007) "Automatically detecting schematic structure components of English abstracts". In *Proceedings of the RANLP´2007, Workshop on Natural Language Processing for Educational Resources,* Borovets, Bulgaria, pp. 23-29.

CEFR B2 to C2: charting a long and winding road

Building a corpus for a longitudinal study of spoken proficiency in English students at Radboud University, the Netherlands

de Vries, Rina; de Haan, Pieter
Department of English / CLS, Radboud University Nijmegen
c.devries@let.ru.nl; p.dehaan@let.ru.nl

In the Netherland, students are usually admitted to university on the basis of a *vwo*-diploma (secondary school diploma), which officially sets their English language skills at CEFR B2 (Melissen, 2007). The English department at Radboud University says in its documentation that its graduates have achieved a CEFR C2 in listening, reading, writing, spoken production and spoken interaction. The department believes that the highest CEFR level is appropriate for its graduates, who will go on to work as teachers, editors, translators, language trainers and communication coaches. In order to reach C2, students are required to take various language proficiency courses during their first two years. But as all other teaching in the department also takes place in English, students are exposed to English on a daily basis and are also required to actively use English outside the language acquisition classroom.

However, at the moment both the B2 entry level and the C2 exit level are hardly more than assumptions. There is no rigorous, principled measuring system in place that would enable the department to assess the actual CEFR level its students are at. There is no doubt that the language skills of students improve significantly in the course of their studies (de Haan & van der Haagen, 2012, 2013; Verheijen, de Haan, & Los, 2013), but any link to the CEFR remains tenuous at best. In view of the fact that some of the other English departments in Dutch universities see their graduates at CEFR C1, it is important for Radboud to start charting the proficiency development of its students and relate this to the CEFR and to the proficiency teaching programme.

It was decided in the context of Radboud's participation in the LONGDALE project (Granger, 2009) to first run a pilot research project, aimed at establishing whether it is possible to make any developments in spoken proficiency visible. Earlier studies have suggested that measurements such as type-token ratio, lexical density and lexical profiling are indicative of improving (spoken) proficiency (de Haan & van Esch, 2005; Johansson, 2008). A pilot project of a small corpus of spoken English provided by students at Radboud University at three different points during their three years at university revealed that their improved proficiency is also visible in the changes observed in type-token-ratio, lexical density and lexical profiling. These promising results have inspired researchers in the English department to continue this line of investigation and set up a longitudinal study of the spoken proficiency of its students. Students will be tested at four different points in their university careers: on entry and at the end of their first, second and third year. They will be required to produce samples of English, which will form a longitudinal corpus and will be analysed in terms of language use, text organisation and communicative achievement.

This poster presentation will outline the suite of spoken language tasks. Tasks have been set up for spoken interaction and spoken production separately. The CEFR has been used used to formulate appropriate tasks which will elicit the required kind of language for each level. Parallel to such tests, a small number of nearly-identical problem solving tasks have been formulated, which will be administered at each of the four testing points, so as to monitor their improved handling of these tasks.

We believe that over time we will be able to describe and illustrate the long and winding road from B2 to C2 better than we can now. We will be able to say with some degree of accuracy how C1 proficiency differs from C2, not just in terms of task types and communicative achievement, but also in terms of actual language use. The findings from our study will indubitably have repercussions for the proficiency teaching programme at Radboud University, but in a wider context, our study might also help take away some of the much-reported 'vagueness' of the CEFR as an assessment tool.

References

de Haan, P., & van der Haagen, M. (2012). Modification of adjectives in very advanced Dutch EFL writing: A development study. *The European Journal of Applied Linguistics and TEFL, 1*(1), 129-142.

de Haan, P., & van der Haagen, M. (2013). Assessing the Use of Sophisticated EFL Writing: A Longitudinal Study. *Dutch Journal of Applied Linguistics, 2*(1), 16-27.

de Haan, P., & van Esch, K. (2005). The development of writing in English and Spanish as foreign languages. *Assessing Writing, 10*, 100-116.

Granger, S. (2009). LONGDALE Retrieved 22 October, 2010, from http://www.uclouvain.be/en-cecl-longdale.html

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61-79.

Melissen, M. (2007). *Exameneisen havo-vwo nieuwe stijl 2007* Alphen aan den Rijn: Kluwer.

Verheijen, L., de Haan, P., & Los, B. (2013). Information Structure: The Final Hurdle? (The Development of Syntactic Structures in (Very) Advanced Dutch EFL Writing). *Dutch Journal of Applied Linguistics, 2*(1), 92-107.

Using an annotated L2 Hungarian corpus to study vowel harmony development

Dickinson, Markus; Ledbetter, Scott
Indiana University
md7@indiana.edu; saledbet@indiana.edu

This research describes an error-annotated corpus of learner Hungarian and its uses for research, especially in the fields of second language acquisition (SLA) and computational linguistics. We detail the corpus, its annotation, and a vowel harmony analysis, charting its development in learner interlanguage (IL). The analysis addresses an exploratory question: What is the process underlying the acquisition of vowel harmony in Hungarian? While there is research on vowel harmony and acquisition in the L1, especially for Hungarian (e.g., Hayes & Londe, 2006; Kontra & Ringen, 1986; Kornai, 1991; Ringen, 1975; Vago, 1991), there is very little in the L2, and none involving learner corpora.

The corpus was collected from students of Hungarian at Indiana University. The data are divided into three levels of proficiency (Beginner, Intermediate, Advanced) as determined by course placement in one of three two-semester sequences. The corpus consists of journal entries from each student, a minimum ten sentences in length on a topic selected by the student. The annotation scheme for the corpus labels and corrects learner errors. Importantly, we make a distinction between *errors* and *adjustments*. While errors are deviations from the standard target form, adjustments are secondary emendations that are conditioned on these errors.

(1) *Szeret -ek kávé -t és tea .*
love 1SG.INDEF coffee ACC and tea .
‘I love coffee and tea.’

	TXT	Szeretek		kávét		és	tea		.
	SEG	Szeret	ek	kávé	t	és	tea		.
Error	CHA								
	MOR								
	REL							MSC	
	SNT								
	TGT	Szeret	ek	kávé	t	és	tea	**t**	.
Adjust.	CHA						CL		
	MOR								
	REL								
	SNT								
	TGT	Szeret	ek	kávé	t	és	**teá**	t	.

In the above example, the learner’s case selection error (MSC) is marked and corrected in the error layer. However, the absent lengthened vowel (CL) in the adjustment layer (necessitated by the corrected accusative case) is not an error, but an adjustment.

The corpus presents a valuable opportunity for linguistic research, providing not only a wealth of authentic production data from second language learners of Hungarian, but also a source of annotated and searchable text, segmented by morphemes to allow for fine-grained analysis of L2 usage and function. We illustrate this by examining the IL development of vowel harmony in Hungarian.

In Hungarian, vowel harmony is a property that determines the selection of allomorphs (predominately suffixes), based on assimilation of vowel height, in noun and verb inflection, as well as derivation (Hayes et al., 2009). The general rule is as follows: stems with only back vowels select a suffix with back vowels (e.g., *ház* + *ban* 'in a house') and stems with front vowels select a suffix with front vowels (e.g., *szék* + *ben* 'in a chair'). Within front vowels, there is a further distinction of rounded and unrounded, though this is often neutralized. While vowel harmony is usually straightforward, numerous exceptions can make selection unpredictable and thus present difficulties for learning. Our analysis can shed light on the troubles learners have and possibly inform instructors and researchers as to the most likely problem areas for targeted instruction.

In our case study of the utility of the annotation, we track the development of two morphemes and their allomorphs in the production data of several learners, showing the emerging competition among forms and rules in each learner's IL. For simplicity, we take two of the first morphemes a learner is expected to encounter: the inessive (*-ban*/*-ben*) case ending and an adverbial derivational ending used with language names (*-ul*/*-ül*). In both cases, the distinction is made between front and back vowels but not between rounded and unrounded. By considering all instances of each morpheme pair, we see a more complete picture of particular underlying morphemes in a learner's developing interlanguage, both what they do right and what they do wrong.

We consider many aspects of the learners' productions, including accuracy, innovation, and consistency in allomorph selection. Such measures allow us a number of insights. First, by tracking the progress of individual learners, we see the emergence of the distinction between front and back vowels. Preliminary results suggest an apparent default preference for back vowels in attested suffixes, giving way to a more complex system over time. Second, we find examples of innovation, as the learners apply hypothesized rules from the grammar to create new forms (e.g., *hal*+*ul* 'fish-speak/fish language'). Finally, with individual timepoints for each learners' interlanguage development, the longitudinal nature of the corpus gives us ordered production data. Competition among allomorphs, evidenced by inconsistencies in the selection of suffixes from one entry to the next, helps us see progress between stages of acquisition. As the data from more learners are analyzed in the future with our annotation scheme, we will be able to draw more generalizable conclusions about the processes of acquisition underlying vowel harmony and other processes in Hungarian.

References

Hayes, Bruce, & Londe, Zsuzsa Cziráky. (2006). Stochastic phonological knowledge: The case of hungarian vowel harmony. *Phonology*, 23(1):59-104.

Hayes, Bruce, Siptár, Péter, Zuraw, Kie, & Londe, Zsuzsa. (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*, 85(4):822-863.

Kontra, Miklós, & Ringen, Catherine. (1986). Hungarian vowel harmony: The evidence from loanwords. *Ural-Altaische Jahrbücher*, 58:1-14.

Kornai, András. (1991). Hungarian vowel harmony. In István Kenesei (ed.), *Approaches to Hungarian 3*, pp. 183-240. Attila József University, Szeged.

Ringen, Catherine. (1975). *Vowel harmony: Theoretical implications*. Ph.D. thesis, Indiana University, Bloomington, IN.

Vago, Robert. (1991). Vowel harmony. In Ronald E. Asher (ed.), *The Encyclopedia of Language and Linguistics*, pp. 4954-4958. Pergamon, Oxford.

Diamesia-related bundles across native and non-native written and oral corpora

Dutra, Deise; Mello, Heliana; Orfano, Bárbara; Grondona, Carolina
UFMG; UFMG; UFSJ; UFMG
dpdutra@ufmg.br; hmello@gmail.com; bmalveira@yahoo.com.br;
carolinaboho@gmail.com

Several Corpus Linguistics studies have focused on lexical bundles recently, mainly in academic settings (Biber et al. 2004; 2006; 2009; Hyland 2008; Simpson-Vlach and Ellis 2010). This last paper presents the Academic Formulas List (AFL) which contains 435 bundles divided in 18 subcategories and is the basis for the pragmatic-functional classification followed in this paper. While Simpson-Vlach and Ellis (2010) based their study on academic oral and written corpora compiled with native speaker productions in academic settings, our research investigates how oral discourse might permeate written texts. Several issues should be considered when comparing oral and written discourse; for instance, the specific genre and not only the mode in which the production takes place. Genre studies (Swales, 2004; Wulff et. al., 2012) have shown that there are specific text features that may vary from textual organization to lexical or syntactic choice and, without genre adequacy, users may fail to express their arguments appropriately. Therefore, we hypothesize that frequent oral bundles are used extensively in essay written production by non-native speakers without a consideration for the genre. This paper aims at presenting the analysis of three and four-word bundles extracted from native and non-native oral and written corpora so as to investigate the extent to which oral discourse constructions permeate the production of argumentative essays. Our data consisted of 2 written and one oral learner corpora, namely, the 16 International Corpus of Learner English (ICLE) sub-corpora (Granger et al. 2009) (3,768,527 words) treated as one learner corpus, Br-ICLE, the Brazilian sub-corpus of ICLE (159,182 words), and LINDSEI-BR (40,456 words), a still under-construction subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI) project. (cf. Gilquin, De Cock, Granger, 2010). The recordings covered three different tasks: a narrative about a chosen set topic by the informant, free discussion with the interviewer and the description of a pictured scene. Two native speaker corpora were also part of our investigation as reference corpora: the Louvain Corpus of Native English Essays (LOCNESS) (324,005 words) and the Santa Barbara Corpus of Spoken English (SBC) made up of conversations which consists of 200,000. For this paper we use a subcorpus of SBC of 56,713 words. The written corpora are constituted by argumentative essays while the oral corpora carry quasi spontaneous and spontaneous conversations. The research methodology included the following steps. First, bundles of 3 and 4 words were extracted from each corpus with scripts and also with Collocate 1.0 (Barlow 2004). Second, we identified the bundles that occurred in the written and oral corpora according to AFL, taking into account both its broad categories (referential expressions, stance expressions and discourse organizing functions) as well as its 18 specific subcategories (e.g. intangible and tangible framing attributes and quantity specification). Third, we identified the differences and similarities in frequency and in bundle structure. Results, based on the analysis of 3 and 4-word bundles, show that: a) the most frequent bundles in the N and NN written and oral corpora are referential and stance expressions; b) the most frequent 3-word referential bundle in the SBC expresses the idea of quantity specification (*a lot* of) and is also present in LOCNESS, however, in this N written corpus other bundles that fulfill the same pragmatic function are more frequent (*a great deal of*, *a large number of*); c) LINDSEI presents similar results to the SBC as far as the most frequent quantity specification bundle are concerned, yet, ICLE and Br-ICLE, the written NN corpora, do not present the use of other bundles to express the same meaning as it was noticed in the N written corpus; d) the most frequent stance expression bundles in both oral and written N and NN corpora are formed by the verb *think* which is part of a verb complement clause (e.g. *I think that the*, *I think he just, I think it is*); e) SBC presents more types of stance bundles with the verb *think* than LOCNESS while the oral and written NN corpora present similar bundles with this verb; f) both SBC and LINDSEI have a high frequency of bundles with verb *know* which is not as frequent in written N and NN corpora. These findings show that a) both N and NN written corpora offer similarities, which may reflect the use of negative politeness markers (e.g. bundles formed with *think*) a characteristic of interpersonal involvement (Tannen 1983); b) the N written corpus shows a variety of bundles for the same pragmatic function; however, the most frequent ones are not as frequent in the oral corpus which may reflect more adequacy to the register than observed in the NN corpora. The conclusion is that oral and written corpora are not totally distinct as far as their bundle frequency analysis shows, but portray bundle preferences for the diamesia to which they belong.

References

Biber, D. et al. (2004) *If you look at* ...: lexical bundles in university teaching and textbooks. *Applied Linguistics*, v.25, n.3, p. 371-405.

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

Biber, D. (2009) A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* v.14, n. 3, p. 275-311.

Granger, S. et al. (2009). *International Corpus of Learner English*: *Version 2*.Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. International *Journal of Applied Linguistics* v.18, p. 41-62.

Simpson-Vlach, R; Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistic*, v.31, n.4, p. 487-512.

Swales, John M. (2004). *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Tannen, D. (1983) Oral and Literate Strategies in Spoken and Written Discourse. *Literacy for life: The demand for reading and writing*, Richard W. Bailey and Robin Melanie Fosheim (Ed.). NY: The Modern Language Association.

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

Learners' and native speakers' use of recurrent word-combinations across disciplines

Ebeling, Signe Oksefjell; Hasselgård, Hilde
University of Oslo
s.o.ebeling@ilos.uio.no; hilde.hasselgard@ilos.uio.no

This study investigates the use of recurrent word-combinations in texts produced by novice writers – both learners and native speakers – across disciplines. More specifically we wish to investigate how salient n-grams really are in different academic disciplines and to what extent the same patterns and functions are used by learners and native speakers. In other words:

- What discourse functions do the recurrent word-combinations have?
- Is L1 background or discipline more decisive for the use of recurrent word-combinations and their functions?

In a previous study (Ebeling 2011), comparing the use and functions of n-grams in UK English student (literature) essays and academic prose, it was concluded that both are academic text types in the sense of being highly informational in nature. An additional trait of the student essays, however, is that they are typically evaluative (i.e. interpersonal) as well (ibid. 69).

In the present study, native speaker data will be taken from the British Academic Written English (BAWE) corpus and (Norwegian) learner data from the Varieties of English for Specific Purposes dAtabase (VESPA-NO). At present VESPA-NO contains sufficient material in two disciplines, namely linguistics and business, to make a comparison with the corresponding native-speaker disciplines worthwhile.

Recurrent word-combinations (n-grams) will be extracted by means of WordSmith Tools. We will investigate the 100 most frequent 3-grams and 4-grams in each subcorpus to assess the degree of overlap across disciplines and L1 groups. Moreover, we will analyse the n-grams functionally along the lines of Moon (1998) into informational, interpersonal and textual types. A hypothesis to be tested is that informational n-grams will discriminate between disciplines while interpersonal and textual n-grams may be more stable. Furthermore, it is expected that the types of n-grams may differ between learners and native speakers; for instance Hasselgård (2012) found that n-grams indicating complex phrases were more typical of native speakers. Furthermore, Paquot et al (2013) found that learners are more visible authors in their texts, which may also show up in their recurrent word-combinations. As the learners in question are relatively advanced, we did not expect to find frequent n-grams that represented lexical errors.

A preliminary look at the data shows a complex picture. Informational n-grams are by far the most frequent type and they seem to be not only discipline-specific, but also topic-specific. There seems to be more n-grams with an interpersonal function (evaluative and modalising) in the linguistics than in the business discipline. N-grams with a textual/organizational function have more similar frequencies across the material. However, there is relatively little overlap in the use of individual n-grams with interpersonal and textual functions across the L1 groups. There is a higher degree of similarity between learners and native speakers in the linguistics discipline than in the business discipline. On the other hand, there is some similarity across disciplines within L1 groups as regards interpersonal and textual n-grams.

References

Ebeling, Signe Oksefjell (2011) Recurrent word-combinations in English student essays. *Nordic Journal of English Studies*, 10:1, 49-76.

Hasselgård, Hilde (2012) *Facts, ideas, questions, problems*, and *issues* in advanced learners' English. *Nordic Journal of English Studies*, 11:1, 22-54.

Moon, Rosamund (1998) *Fixed Expressions and Idioms in English. A Corpus-based Approach*. Oxford: Clarendon Press.

Paquot, Magali, Hasselgård, Hilde & Ebeling, Signe Oksefjell (2013) Writer/reader visibility in learner writing across genres. A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In: Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 377-387.

Corpora

BAWE, see http://wwwm.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx
VESPA, see http://www.uclouvain.be/en-cecl-vespa.html

Expletive subjects in L2 English: a corpus based transfer study

Esther Ferrandis
Universidad Autónoma de Madrid, Spain
esther.ferrandis@uam.es

Much of Second Language Acquisition (SLA) research has traditionally relied on elicited data and understated the role of natural language use data. The consequence of this, according to Granger, is that SLA research tends to be based in a limited number of subjects, which raises questions about the generalizability or results (2002:6), However, this situation started to change thanks to the compilation of linguistic databases and learner corpora from a variety of mediums, from different genres and of varying sizes which provide a much wider empirical base than has previously been available. Now, thanks to these large databases, we can find some structures which are rarely found in small studies and discover patterns which may influence the learner, such as the learner's first language (L1) or the L2 itself, or directly in the learners' output (interlanguage patterns) (Myles 2005:5). The main goal of this study is to contribute to the debate about the nature of transfer or cross-linguistic interference by focusing on the acquisition of overt expletives (*it* and *there*) in L1 Spanish-L2 English grammars.

A corpus study was carried out using the WriCLE corpus, a written corpus of academic essays with 700.000 words written by L1 Spanish learners of L2 English and compiled at the Universidad Autónoma de Madrid (Rollinson & Mendikoetxea 2010). 75 texts from 75 different university students of English Studies and English Philology with different proficiency levels were randomly selected and manually annotated using the software UAM Corpus Tool (version 2.7.2) (O'Donnell 2008). The texts were divided into three groups according to the scores the students obtained in the Oxford Quick Placement Test (the correspondences with the CEFRL-Common European Framework of Reference for Languages- are shown between brackets): 25 for the low group (A2-B1), 25 for the intermediate group (B1-B2) and 25 for the advanced group (C1-C2).

Experimental work in the L2 literature has shown that English and Spanish differ as to the setting of the Null Subject Parameter (NSP). My main objective was to analyse the role the L1 (Spanish), a [+pro-drop] language, has in the acquisition of overt expletives in L2 English, a [-pro-drop] language, as studies have pointed out that expletive subjects perform a very important role in L1 acquisition, being the instruments English children use in order to reset their initial [+ pro-drop] value into their final [-pro-drop] one (Ruiz de Zarobe 1986) (for a further discussion see Hyams et al. 1986, Hyams & Wexler, 1991). My main research questions were: (RQ1) Are overt expletives of English (*it*, *there*) problematic for Spanish learners? If so, are they problematic at all proficiency levels? (RQ2) Do Spanish natives initially transfer their L1 parameter value (use of *Ø expletive*) when acquiring English as a second language? (RQ3) If they do so, are learners able to reset completely their L1 parameter [+ pro-drop] to the English one [-pro-drop]? In order to answer these questions and annotation scheme was designed where I accounted for the referentiality of the subject, the type of predicate it appeared with, the word order it appeared in and its grammaticality or ungrammaticality. I also accounted for the reason of this possible ungrammaticality. A total of 681 expletive subjects were found in the texts selected for the study (expletives *there*, *it* and use of *Ø expletive*) and the results confirmed partially RQ1 and RQ2, as only expletive *it* was problematic in all levels, whereas expletive *there* was not. As for RQ3, I did not expect a full acquisition of this parameter and thus not a full resetting of the L1 setting. My results confirmed this last hypothesis, which indicates that though Spanish learners of L2 English acquire referential subjects early in the acquisition process, expletive subjects remain problematic and are never fully mastered, not even in advanced stages of the learning process.

References

Granger, Sylviane. (2002) A bird's-eye view of learner corpus research. In Granger, S., Hung, J., & Petch-Tyson, S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Amsterdam: John Benjamins.

Hyams, Nina. (1986) *Language Acquisition and the Theory of Parameters.* Dordrecht: Reidel.

Hyams, Nina. and Wexler, Kenneth (1991). On the Grammatical Basis of Null subjects in Child Language. *Linguistic Inquiry, 24*: 421-459

O'Donnell, Michael. (2008) The UAM Corpus Tool: Software for corpus annotation and exploration. Paper presented at the *XXVI Congreso de AESLA*, Almería, Spain, 3-5 April 2008.

Myles, Florence. (2005) Review article. Language Corpora and Second Language Acquisition Research. *Second Language Research 21*, 4, 373-391.

Rollinson, Paul. & Mendikoetxea, Amaya. (2010) Learner corpora and second language acquisition: Introducing WriCLE In: J. L. Bueno Alonso, D. Gonzáliz Álvarez, U. Kirsten Torrado, A. E. Martínez Insua, J. Pérez-Guerra, E. Rama Martínez & R. Rodríguez Vázquez (eds.) *Analizar datos>Describir variación/Analysing data>Describing variation*. Vigo: Universidade de Vigo (Servizo de Publicacións), pp. 1-12.

Ruiz de Zarobe, Yolanda. (1998) El parámetro pro-drop y la adquisición del inglés como segunda lengua. *ITL: International Journal of Applied Linguistics 119-120:*49-64.

Patterns of misspellings in L2 English – a view from the ETS Spelling Corpus

Flor, Michael; Futagi, Yoko; Lopez, Melissa; Mulholland, Matthew
Educational Testing Service, Princeton, NJ, USA
{mflor, yfutagi, mlopez002, mmulholland}@ets.org

This paper describes the Educational Testing Service (ETS) Spelling Corpus. The corpus comprises 3000 essays written by examinees on the writing sections of college-level international high-stakes English language assessments – GRE® and TOEFL®. The corpus has a strong international aspect – it includes essays written by native and non-native speakers of English, where the non-native speakers are English language learners from different countries around the world. The corpus has been manually annotated for non-word and real word misspellings and contains more than 24,000 misspelled tokens, with corrections, in context. As such, this is presently one of the largest corpora of its kind.

The corpus was developed for evaluation of spellcheckers, and for research on patterns of misspellings produced by both native English speakers and English language learners. The paper provides details of the annotation scheme and procedure, the methodology of error classification and decisions taken concerning the status of orthographic errors. The paper provides analyses of patterns of misspellings along several different dimensions. These include some traditional typologies of errors, such as competence versus performance misspellings (typos), linguistic subsystem (lexical, morphological or phonological misspellings), orthographic similarity (degree of target modification as measured by edit distance). Additional dimensions of analysis consider frequency, length and structural aspects of target words, such as over-regularized past tense. Analyses are presented with breakdown for native vs. non-native speakers. We also relate the analyses to writing proficiency, by providing data breakdown for high-scoring vs. low-scoring essays and by specific essay-score levels.

The paper also discusses some additional topics. One interesting issue is density of errors – both global density (errors per essay) and local density (distribution of errors in the essay). Another discussion concerns multi-word errors (spelling errors spanning more than one word in a sequence). We also present data on misspellings of names (place names and person names), a topic that is rarely covered in general discussions of spelling errors. Extending dictionary coverage with a large set of names may lead to decreased accuracy of error detection. That is addressed via improving real-word misspelling detection. The paper provides examples from the corpus.

Bibliography

Bestgen Y. and Granger, S. (2011). Categorising spelling errors to assess L2 writing. International Journal of Continuing Engineering Education and Life-Long Learning, 21(2/3):235-252.

Botley, S. and Dillah, D. (2007). Investigating Spelling Errors In A Malaysian Learner Corpus. Malaysian Journal Of ELT Research, Vol. 3, pp. 74-93.

Boyd, A. (2009). Pronunciation Modeling in Spelling Correction for Writers of English as a Foreign Language. In Proceedings of the NAACL HLT 2009 Student Research Workshop & Doctoral Consortium, 31–36.

Cook V. (1997). “L2 users and English spelling”, Journal of Multilingual and Multicultural Development, vol. 18(6), 1997, p. 474-488.

Dagneaux, E., Denness, S., and Granger, S. (1998). Computer-aided error analysis. System, 26(2), 163-174.

ETS. (2011a). GRE®: Introduction to the Analytical Writing Measure. Educational Testing Service. www.ets.org/gre/revised_general/prepare/analytical_writing (last accessed on March 9, 2012).

ETS. (2011b). TOEFL® iBT® Test Content. Educational Testing Service. www.ets.org/toefl/ibt/about/content (last accessed on March 9, 2012).

Flor M. (2013). Four types of context for automatic spelling correction. Accepted for publication in the TAL journal (Traitement Automatique des Langues), 53(3), Special Issue: Managing noise in the signal: error handling in natural language processing.

Flor, M. & Futagi, Y. (2013). Producing an annotated corpus with automatic spelling correction. In S. Granger, G. Gilquin & F. Meunier (eds) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 139-154.

Flor M. and Futagi Y. (2012). “On using context for automatic correction of non-word misspellings in student essays”, The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, BEA-7 (at NAACL HLT 2012), Montreal, Canada, June 3-8 2012, p. 105–115.

Futagi Y. (2010). “The effects of learner errors on the development of a collocation detection tool.” In Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (AND '10), 2010, p. 27-34.

Hovermale, DJ. (2010). An analysis of the spelling errors of L2 English learners. Presented at CALICO 2010 Conference, Amherst, MA, USA, June 10-12, 2010. Available electronically from http://www.ling.ohio-state.edu/~djh/presentations/djh_CALICO2010.pptx

Kukich K., 1992. Techniques for automatically correcting words in text. ACM Computing Surveys, 24:377-439.

Levenshtein, L. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10:707-710.

Lunsford A.A. and Lunsford K.J. (2008). “Mistakes Are a Fact of Life: A National Comparative Study”, College Composition and Communication, vol. 59 no.4, 2008, p. 781-806.

Mitton R. (1996). English spelling and the computer. Harlow, Essex: Longman Group. Available electronically from http://eprints.bbk.ac.uk/469

Mitton R. (2009). Ordering the suggestions of a spellchecker without using context. Natural Language Engineering, 15(2):173–192.

Mitton R. and Okada T. (2007). The adaptation of an English spellchecker for Japanese writers. Presented at: Symposium on Second Language Writing, 15-17 Sept 2007, Nagoya, Japan. Available electronically from http://eprints.bbk.ac.uk/592

Okada T., 2005. Spelling errors made by Japanese EFL writers: with reference to errors occurring at the word-initial and the word-final position. In V. Cook and B. Bassetti (Ed.), Second language writing systems, pages 164-183. Clevedon: Multilingual Matters.

Pollock, J. J., & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. Communications of the ACM, 27(4), 358-368.

Rimrott A., and Heift T., (2008). Evaluating automatic detection of misspellings in German. Language Learning & Technology, 12(3), p.73-92.

Ringlstetter Ch., Schulz K.U., and Mihov S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. Computational Linguistics, 32(3):295-340, 2006.

Tavosanis M. (2007). A Causal Classification of Orthography Errors in Web Texts (AND2007).

From Learner Corpus Research to Pedagogy in EAP: State-of-the-art

Flowerdew, Lynne
HKUST
lclynne@ust.hk

Learner corpus research is now well established. Researchers/practitioners have emphasised the value of the findings for pedagogy and illustrated how learner corpora can usefully supplement native corpora in teaching (Meunier 2002; Gilquin et al. 2007a; Flowerdew 2010, 2012). The purpose of this paper is to review how learner corpus research of EAP writing has informed pedagogy in English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP and to illustrate how the gap between learner corpus-based research and pedagogy has narrowed in recent years. At the same time, key issues in using learner corpora will be raised, including the type of native speaker writing deemed appropriate for the contrastive corpus and the question of delayed vs. immediate pedagogic use of learner corpora.

Since the mid-1990s to the present day a great deal of valuable research has been conducted on the ICLE sub-corpora consisting of EGAP argumentative writing of learners at high intermediate/advanced level with different L1 backgrounds (see Granger 1998, 2003 for studies). Most of this research has been carried out on the overlapping areas of modality and stance and is of a contrastive nature (see Aijmer 2002 for Swedish learners; Hasselgård 2009 for Norwegian learners; Neff et al. 2003 for Spanish learners). Another study on learner argumentative writing (but not using the ICLE sub-corpora) is that by Hyland & Milton (1997) on Hong Kong learners. All these studies offer important pedagogic implications on how to address students' infelicities. For example, Aijmer (2002) and Hyland & Milton (1997) comment on the fact that as textbooks put emphasis on modal verbs at the expense of other categories, learners should be exposed to a much greater range of other linguistic devices. That these devices should be presented in a discourse-analytical context with attention paid to tone, register and meaning is another key recommendation. Thirdly, Hasselgård (2009) and Neff et al. (2003) advocate sensitizing learners to the contrastive aspect of epistemic and pragmatic modal contexts in L1 and L2 and also placing emphasis on the stylistic and communicative effects produced by different epistemic markers. As for direct applications, Milton's (1998) findings on learner corpora have been applied in an online CALL environment with the provision of a 'list-driven' concordancer displaying a set of hedging devices problematic for Hong Kong students.

Vocabulary issues in learner EGAP writing were put on the map by Nesselhauf's (2003, 2007) work on collocations in the German component of ICLE and Paquot's (2010) extensive research across several of the ICLE corpora examining vocabulary from a phraseological and rhetorical perspective. Nesselhauf (2007) outlines suggestions for three major criteria to be considered in the selection of vocabulary for advanced learners, namely frequency, difficulty and degree of disruption. Paquot's research findings have been used to substantially inform writing sections of dictionaries (Gilquin et al. 2007b; Granger & Paquot 2010).

It was not really until the 2000s that attention was paid to ESAP learner corpora. Two studies have investigated learner corpora of technical report writing from a genre-based discourse perspective. Flowerdew (2008) investigated the phraseology of keywords signalling the Problem-Solution pattern in a corpus of student recommendation-based reports. Luzón Marco

(2010), meanwhile, examined a variety of discourse-based errors including signalling nouns, i.e. those nouns such as *reason, conclusion* etc. acting as cohesive devices (see J. Flowerdew 2010). She also found that genre phraseology, i.e. lexical bundles such as *the objective / purpose / aim / goal (of this report) is* was missing from student writing. While these findings are insightful they remain at the level of pedagogic implications. However, two contrastive studies which describe hands-on concordancing activities for dissertation writing using ESAP learner corpora are those by Hewings & Hewings (2002) for MBA students and Eriksson (2012) for biochemistry PhD students (see Granger & Paquot 2013 for further studies).

Granger (2008: 272) has signaled that 'a wider range of learner corpora, in particular, longitudinal, spoken and domain-specific, need to be compiled and disseminated'. With respect to written corpora of an ESAP nature, two large-scale domain-specific projects, VESPA (Varieties of English for Specific Purposes dAtabase) and CALE (Corpus of Academic Learner English), are underway. A longitudinal corpus, the Malmö University-Chalmers Corpus of Academic Writing as Process, is also now under compilation.

References

Aijmer, Karin (2002) Modality in advanced Swedish learners' written interlanguage. In Granger, Sylviane et al. (eds.), pp. 55-76.

Eriksson, Andreas (2012) Pedagogical perspectives on bundles: teaching bundles to doctoral students of biochemistry. In Thomas, James & Boulton, Alex (eds.) *Input, Process and Product. Developments in Teaching and Language* Corpora (pp. 195-211). Masaryk University Press: Brno, Czech Republic,

Flowerdew, John (2010) Use of signalling nouns across L1 and L2 writer corpora. *International Journal of Corpus Linguistics* 15 (1): 36-55.

Flowerdew, Lynne (2008) *Corpus-based Analyses of the Problem-Solution Pattern*. Amsterdam: John Benjamins.

Flowerdew. Lynne (2010) Using corpora for writing instruction. In O'Keeffe, Anne & McCarthy, Michael (eds.) *The Routledge Handbook of Applied Linguistics* (pp. 444-457). London: Routledge.

Flowerdew, Lynne (2012) *Corpora and Language Education*. London: Palgrave Macmillan.

Gilquin, Gaëtanelle, Granger, Sylviane & Paquot, Magali (2007a) Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6 (4): 319-335.

Gilquin, Gaëtanelle, Granger, Sylviane & Paquot, Magali (2007b) Improve your writing skills: writing sections. In Rundell, Michael (editor-in-chief) *Macmillan English Dictionary for Advanced Learners* (2nd edition) (pp. IW4-IW28). Oxford: Macmillan Education,

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

Granger, Sylviane (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, Sylviane (2008) Learner Corpora. In Lüdeling, Anke & Kytö, Merja (eds.) *Corpus Linguistics. An International Handbook* (pp. 259-275). Berlin: Mouton de Gruyter,

Granger, Sylviane, Hung, Joseph & Petch-Tyson, Stephanie (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, Sylviane & Paquot, Magali (2010) The Louvain English for academic purposes dictionary. In Granger, Sylviane & Paquot, Magali (eds.) *eLexicography in the 21st century: New applications, new challenges* (pp. 87-96). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, Sylviane & Paquot, Magali (2013) Language for specific purposes learner corpora. In Chapelle, Carol (ed.) *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell.

Hasselgård, Hilda (2009) Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In Aijmer, Karin (ed.) *Using Corpora in Language Teaching* (pp. 121-140). Amsterdam: John Benjamins.

Hewings, Martin & Hewings, Ann (2002) "It is interesting to note that …": a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, 21(4): 367-383.

Hyland, Ken & Milton, John (1997) Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6 (2): 183-205.

Luzón Marco, Marie Jose (2010) Analysis of organizing and rhetorical items in a learner corpus of technical writing. In Campoy-Cubillo, Mari, Bellés-Fortuno, Begona & Gea-Valor, Marie (eds.) *Corpus-based Approaches to English Language Teaching* (pp.79-94). London: Continuum.

Meunier, Fanny (2002) The pedagogic value of native and learner corpora in EFL grammar teaching In Granger, Sylviane et al. (eds.), pp. 119-142.

Milton, John (1998) Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In Granger, Sylviane (ed.), pp. 172-185.

Neff, J, Dafouz, E., Herrera, H., Martinez, F. and Rica, J. 2003. Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance. In Granger, Sylviane & Petch-Tyson, Stephanie (eds.) *Extending the Scope of Corpus-based Research* (pp. 211-230). Amsterdam: Rodopi.

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

Nesselhauf, N.(2007) The path from learner corpus analysis to language pedagogy: some neglected issues. In Fachinetti, Roberta (ed.) *Corpus Linguistics 25 Years On* (pp. 305-315). Amsterdam: Rodopi.

Paquot, Magali (2010) *Academic Vocabulary in Learner Writing*. London: Continuum.

Comparing French/Spanish L1 transfers in two English learner corpora: the case of indexicals *it, this* and *that*

Gaillat Thomas
University of Paris-Diderot & University of Rennes 1
thomas.gaillat@univ-rennes1.fr

In this paper, we show that it is possible to carry out a fine-grain comparative error analysis of learner English of different L1s. The objective of this work is to better understand how learners make use of *this* and *that* as pro-forms. As they appear to interact closely with the pronoun *it* (Gaillat 2013), a micro-system of errors is revealed and the errors highlight transfers that depend on a particular L1. To compare how errors, and thus transfers, might be different between different L1s, we query two corpora of two different L1s.

The question of annotation schemes is highly significant (Wynne, 2005) as we want to find a way to compare data from different corpora. If we consider PoS tagging used in learner corpora, a variety of error-based (Dagneaux et al, 1998; Diaz-Negrillo et al, 2006) or native-based annotation schemes have been developed over time. This makes cross-corpora data comparison difficult. A solution is to be found in work on data-exchange through a functional re-tagging of the corpus. Research on demonstrative-related errors shows problems with the selection of not only demonstratives (Lenko-Szymanska, 2004), but also of pro-forms including the pronoun *it*. Demonstratives endorse distinct grammatical functions, and their selection is highly dependent on the contexts of use and “certain features of the indexical element's immediate context play a decisive role in its interpretation” Cornish (1999: 83). Errors in the form of substitutions occur within learner language when indexical expressions are constructed by learners. It is therefore paramount to characterise indexicals in their functional nature in order to explore the contexts in which they compete, hence revealing a micro-system of referential errors. This characterisation has been partly achieved previously via a modified PoS annotation scheme that helps to distinguish the functional roles of *this* and *that*. Our hypothesis is that these contexts partly correspond to PoS sequences that may be repetitive, forming L1-specific patterns. We intend to identify these patterns with queries embarking PoS and text information, especially with the supplementary TPRON tag identifying pro-form uses (versus deictic) of *this* and *that* and PRP identifying *it.*

In our experiment, we use two corpora of learner English, PoS tagged with TreeTagger (Schmid 1994) with this modified tagset. Corpus files contain their own XML structure that includes two layers of information (text and PoS), queryable simultaneously with NITE NXT Search. As our linguistic goal is to explore the micro-system of pro-form reference used by learners, we want to see if there are differences in the use of the pronoun *it* and the demonstratives as pro-forms between Spanish and French learners of English. We use NOCE (Diaz Negrillo 2009), a PoS and error-annotated corpus of written English of Spanish speakers, and LONGDALE - DIDEROT (Meunier *et al*. 2008), a corpus of spoken English of French speakers. Replicating a previous corpus comparison (Heid *et al.* 2004) with NXT Search (Carletta *et al.* 2003), we converted the two corpora to the NITE NXT format. Both corpora are converted with the creation of a metadata file that matches each corpus to two distinct NITE *observations*. Both corpora can thus be opened simultaneously. We apply queries to analyse variations on the uses of the pro-form function of the demonstratives both as subjects and objects of verb phrases. Search queries combine text and PoS to search all the occurrences of *this*/*that* pro-forms and *it* pronouns in both corpora. For example the query: ($word) : $w@orth = "this" && $word@pos= “TPRON” allows the extraction of all forms of *this* whose function is that of a pro-form.

Query results allow the comparisons of patterns such as *this* pro-form + verb or *that* pro-form + verb or *it* + verb across L1s. Consequently, pro-form-related errors within the referential micro-system are detected and they suggest that variations exist between L1s. By combining multiple annotation layers with several corpora, queries have levels of complexity that span the texts syntagmatically and paradigmatically, and patterns of use of pro-forms in learners are revealed. This experiment supports the view that automatic-POS tagging can help error analysis (Diaz *et al*. 2010) and finer-grained PoS tagging adapted to functional distinctions (deictic vs pro-form) paves the way to the analysis of cross-linguistic L1 effects.

References

Carletta, Jean, Jonathan Kilgour, Tim O'Donnell, Stefan Evert, and Holger Voormann (2003) The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003)*. Budapest, Hungary.

Cornish, Francis (1999) *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford: Oxford University Press.

Dagneaux, Estelle, Sharon Denness, and Sylviane Granger (1998) Computer-aided error analysis. *System*.163–174.

Diaz-Negrillo, Ana, Jésus Fernandez-Domingez (2006) Error Tagging Systems for Learner Corpora. *RESLA*.83–102.

Díaz Negrillo, Ana (2009) *EARS: A user's manual.* Munich. LINCOM Academic Reference Books.

Diaz Negrillo, Ana, Meurers, Detmar, Valera, Salvador., & Wunsch, Holger (2010) Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *In Language Forum*, 36(1-2), pages 139–154.

Gaillat, Thomas (2013) *This* and *that* in Native and Learner English: From Typology of Use to Tagset Characterisation. In S. Granger, G. Gilquin & F. Meunier (eds) (2013) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.

Heid, Ulrich, Holger Voorman, Jan-Torsten Milde, Ulrike Gut, Katrin Erk, and Sebastian Pado (2004) Querying both time-aligned and hierarchical corpora with NXT Search. *Proceedings of LREC 2004*, 1455–1458.

Lenko-Szymanska, Agnieszka (2004) Demonstratives as Anaphora Markers in Advanced Learners' English. *Corpora and Language Learners*. In Aston, Guy & Bernardini, Silvia & Stewart, Dominic (Eds). Amsterdam / Philadelphia: Benjamins, 84–108.

Meunier, Fanny, Sylviane Granger, Damien Littré, and Magali Paquot (2008). The LONGDALE (Longitudinal Database of Learner English). UCL-CECL. https://www.uclouvain.be/en-cecl-longdale.html. Last consulted 7 June 2013

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 14–16.

Wynne, Martin (2005) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow. Oxford.

EF Cambridge Open Language Database (EFCAMDAT)

Geertzen, Jeroen; Alexopoulou, Theodora; Korhonen, Anna
Dept. of Theoretical and Applied Linguistics, University of Cambridge
jg532@cam.ac.uk; ta259@cam.ac.uk; alk23@cam.ac.uk

We present the EF Cambridge Open Language Database, henceforth, EFCAMDAT, a new open access database of written L2 English. EFCAMDAT was developed at the Dept of Theoretical and Applied Linguistics, at the University of Cambridge in collaboration with EF Education First, an international educational organisation. EFCAMDAT contains writings submitted to *Englishtown* the online school of EF, accessed daily by around 300,000 learners worldwide. The magnitude of EF operations has allowed us to build a resource of considerable size, currently containing 412K scripts from 76K learners summing up 32 million words. As new data come in, we expect to reach 100 million words by 2014 and be able to follow the longitudinal development of even more students.

EFCAMDAT consists of writings submitted to *Englishtown*, the online school of EF Education First, accessed by language learners all over the world (Education First, 2012). A full course in Englishtown spans 16 proficiency levels aligned with common standards such as TOEFL, IELTS and the Common European Framework of Reference for languages (CEFR). When students start a course at EF they are placed at the first level of a stage (levels 1, 4, 7, 10, 13, or 16) after a placement test and may proceed to higher levels through successful progression through coursework. Each of the 16 levels contains eight lessons, offering a variety of receptive and productive tasks. EFCAMDAT consists of scripts of writing tasks at the end of each lesson on topics like those listed in Table 1.

Table 1: Examples of essay topics at various levels. Level and unit number are separated by a colon.

ID	Essay topic	ID	Essay topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you're doing	11:1	Writing a movie review
2:8	Describing your family's eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature

Given 16 proficiency levels and 8 units per level a learner who starts at the first level and completes all 16 proficiency levels would produce 128 different essays. Essays are graded by language teachers; learners may only proceed to the next level upon receiving a passing grade. Teachers provide feedback to learners using a basic set of error markup tags or through free comments on students' writing. Currently, EFCAMDAT contains teacher feedback for 36% of scripts.

The data collected for the first release of EFCAMDAT contain 551,036 scripts (with 2,897,788 sentences, and 32,980,407 word tokens) written by 84,864 learners. We currently have no information on the L1 backgrounds of learners, but metadata on the L1 background of learners is being collected for the second release of the database. Information on nationality is, thus, used as the closest approximation to L1 background. EFCAMDAT contains data from learners from 172 nationalities, with 28 nationalities having more than 100 learners, and 38 nationalities having more than 50 learners. Table 2 shows the spread of scripts across the nationalities with most learners.

Table 2: Percentage and number of scripts per nationality of learners

Nationality	Percentage of scripts	Number of Scripts
Brazilians	36.9%	187,286
Chinese	18.7%	96,843
Russians	8.5%	44,187
Mexicans	7.9%	41,115
Germans	5.6%	29,192
French	4.3%	22,146
Italians	4.0%	20,934
Saudi Arabians	3.3%	16,858
Taiwanese	2.6%	13,596
Japanese	2.1%	10,672

Most learners only complete portions of the program. Nevertheless, around a third of learners (around 28K) have completed 3 full levels, corresponding to a minimum of 24 scripts. Texts range from a list of words or a few short sentences to short narratives or articles. As learners become more proficient they tend to produce longer scripts. On average, scripts count 7 sentences (SD=3.8). Sample scripts are shown in the following figure.

1. Learner 18445817, Level 1, Unit 1, Chinese

Hi! Anna,How are you? Thank you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english toghter! Bye! Anfeng.

2. Learner 19054879, Level 2, Unit 1, French

Hi, my name's Xavier. My favorite days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. Learner 19054879, Level 8, Unit 2, Brazilian

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new prosperous city brought. I really enjoyed this song.

EFCAMDAT scripts have been annotated automatically with with Penn Treebank part-of-speech tags (Marcus et al., 1993) and grammatical relations according to the Stanford Dependency scheme (De Marneffe and Manning, 2008). Details of the automatic annotation and an evaluation of how these tools perform on learner data is presented in (Geertzen et al., 2012).

The database is accessed through a web based interface at `http://corpus.mml.cam.ac.uk/efcamdat/`. The interface supports selection of scripts from different proficiency levels and by learners of different nationalities, search for parts of speech and grammatical relations and export of raw text as well as tagged scripts. EFCAMDAT is freely available to the academic community subject to an end-user agreement protecting copyright.

References

De Marneffe, M. C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proc. of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Education First (2012). Englishtown. `http://www.englishtown.com/`.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2012). Automatic linguistic annotation of large scale l2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *in Proceedings of the 31st Second Language Research Forum (SLRF), Carnegie Mellon.* Cascadillla Press.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

L2 nominalization use: A corpus-based investigation into the interplay of L1 influences, L2 proficiency, and genre knowledge.

Gentil, Guillaume; Meunier, Fanny
Carleton University (Canada) ; Université Catholique de Louvain (Belgium)
guillaume.gentil@carleton.ca ; fanny.meunier@uclouvain.be

According to Flowerdew (2006) nominalizations are problematic for learners. Cross-sectional studies have established a positive correlation between the frequency of nominalizations in L2 texts and rating scores of L2 academic writing ability (e.g., Grant and Ginther, 2000). Yet it is unclear what exactly the frequency of nominalizations is a measure of: register development, language proficiency, or genre sensitivity? Also unclear is the degree to which *greater* nominalization use is a reliable indicator of *better* use.

To investigate these questions, we draw on two corpora: the Longitudinal Database of Learner English (LONGDALE) and the International Corpus of Learner English (ICLE). We compare nominalization use in French and Spanish L1 learners of English as a foreign language. The French subcorpus comes from LONGDALE and is truly longitudinal whilst the Spanish subcorpus is pseudo-longitudinal and comes from ICLE.

Comparisons in nominalization use will be made:

- among L1 backgrounds
- across genres (short opinion pieces vs. longer literary or linguistics EAP essays)
- across proficiency levels

From a theoretical perspective, we combine Tardy's (2009) model of genre knowledge (as adapted in Gentil, 2011), Halliday and Martin's (1993) notion of grammatical metaphor, and Biber's operational definition of nominalizations as abstract nouns formed from verbs or adjectives through derivational morphology (Biber, Conrad, & Leech, 2002, p. 458). This three-pronged framework helps to conceptualize nominalization use in relation to genre/register development and language proficiency. We also touch on the potential influence of the L1 background.

CLAWS part-of-speech tagger and UCREL Semantic Analysis System (USAS) are used to identify nominalizations, while WordSmith Tools facilitates the investigation of use in context. A coding system was also developed to tag nominalizations for grammatical accuracy and rhetorical appropriateness.

Preliminary analyses of nominalization use for the L1 French subcorpus indicate minimal variation between year 1 and year 3 within the short opinion pieces, but significant differences between this genre and the year-4 literary or linguistic term papers. The data analysis is currently being done for the L1 Spanish subcorpus. The full set of results will be available by September 2013.

References

Biber, D., Conrad, S., and Leech, G. (2002). Longman student grammar of spoken and written English. Harlow, England: Pearson Education.

Flowerdew, J. (2006) Use of signalling nouns in a learner corpus. In: Flowerdew, J. & Mahlberg, M. (eds.), Lexical Cohesion and Corpus Linguistics. Special issue of International Journal of Corpus Linguistics 11:3, 345–362.

Gentil, G. (2011). A biliteracy agenda for genre research. Journal of Second Language Writing. 20(1), 6-23.

Grant, L., Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. Journal of Second Language Writing, 9(2), 123-145.

Grant, L., Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. Journal of Second Language Writing, 9(2), 123-145.

Halliday, M., & Martin., J. (1993). Writing science. Pittsburgh: University of Pittsburgh Press.

Tardy, C. (2009). Building genre knowledge. West Lafayette, IN: Parlor Press.

Discriminating CEFR levels in Greek L2: a corpus-based study of young learners' written narratives

Giagkou, Maria; Kantzou, Vicky; Stamouli, Spyridoula; Tzevelekou, Maria
Institute for Language and Speech Processing, "Athena" Research Centre
{mgiagkou; vkan; pstam; mtzevelekou}@ilsp.athena-innovation.gr

In recent years a growing number of research projects focus on combining 'Can Do' descriptors of the CEFR proficiency levels (Council of Europe, 2001) with research findings on the acquisition of various linguistic features by second language (L2) learners (Bartning *et al.* 2010; Hawkins & Buttery 2010; Hawkins & Filipović 2012; Prodeau *et al.* 2012), yielding to the establishment of a correlation between CEFR and the specificities of individual linguistic systems. Such efforts have focused mainly to adult L2 learners.

The present study draws from this line of research and aims at identifying the main characteristics of written language produced by young L2 learners of Greek enrolled in Greek state schools. The research question it attempts to answer is: What are the linguistically informed features that discriminate proficiency levels of young Greek L2 learners?

To this end, a corpus of written productions was compiled. Students attending grades three to six of primary school performed two writing tasks: (a) a narrative based on the Cat Story picture series (Hickman 2003), and (b) a letter or diary entry. Each student was placed at a CEFR level on the basis of the aforementioned written productions by two evaluators. Rating was based on CEFR descriptors and more specifically on the Overall Written Production, Creative Writing and Lexical, Grammatical and Orthographic Competence scales. In order to control for task effects, the letter and diary entry were excluded from further analysis. Narratives placed at the same level by both evaluators were included, resulting in a corpus of 150 scripts (9742 tokens). Levels A2, B1 and B2 are represented in the corpus, with 50 scripts in each level.

Narratives were manually transcribed and annotated with respect to:

(a) type of clause (independent / dependent and type of dependent clause),

(b) clitics within the verb frame, subcategorized into correct and incorrect uses,

(c) adjectives and adverbs, subcategorized into descriptive and evaluative and

(d) discourse markers, and type of discourse marker (additive, temporal, contrastive, inferential, other). These features where selected on the basis of previous research on Greek L2 acquisition, which found differences among proficiency levels with respect to narrative length, use of conjunctions and dependent clauses, clitics and evaluative devices (Kantzou 2010, 2012, Stamouli 2010, Varlokosta & Triantafyllidou 2003).

Moreover, in order to proceed to an accurate measurement of lexical density, a second, error-free, version of the corpus was created by eliminating spelling errors.

Statistical analysis of the data (One-way ANOVA and Bonferroni test) revealed significant differences between all levels in terms of narrative length (number of tokens and types, and number of clauses).

As for type of clause, the percentage of dependent clauses was found to discriminate both A2 from B1 and B1 from B2. It is worth mentioning that zero frequencies of dependent clauses were encountered only in A2 and B1, whereas there was at least one dependent clause in all B2 scripts. In other words, scripts with zero dependent clauses are most likely to belong to A2 or B1.

Neither the ratio of clitics to tokens nor the mean number of clitics per clause was found significantly different, whereas correct uses discriminated all levels. Results showed that when a learner exhibits at least one infelicitous use of a clitic, it is likely that s/he is at level B1 or below.

Number of adjectives and adverbs were found non-discriminatory. By contrast, the type of adverb, i.e. evaluative vs. descriptive, plays a significant role in discriminating levels. Almost half of the adverbs used by B2 learners expressed the learner's evaluative judgments. The number decreases significantly as the level decreases, from 25% in B1 to only 5% in A2.

The total number of discourse markers proved to discriminate B2 from B1. B2 learners seem to have used at least one discourse marker in every two clauses. The analysis of different types of discourse markers showed significant differences between levels, e.g. decreased use of additives as the level increases. This is due to the very frequent use of the common additive *και* (and) in lower levels. As the learner's vocabulary grows, s/he exploits different discourse markers to make more subtle links between utterances.

Finally, the ratio of functional to content words did not differ significantly between levels. This demonstrates that lexical density was a non-discriminatory feature.

On the basis of these results, a number of indices are put forward as criterial features discriminating language proficiency levels in L2 Greek narratives: narrative length, frequency of dependent clauses, correct uses of clitics, use of evaluative adverbs and additive discourse markers. Further research needs to validate these indices as criterial features for CEFR levels, by investigating different discourse types in larger corpora. By linking the CEFR levels to linguistic indices, human rating of L2 written productions will gain in reliability. Moreover, these criterial features can be used in data-driven approaches for the (semi)automatic evaluation of writing.

References

Bartning, Inge; Martin, Maisa & Vedder, Ineke (eds.) (2010) *Communicative development and linguistic development: intersections between SLA and language testing research*. Eurosla Monographs Series 1. Available at: http://eurosla.org/monographs/EM01/EM01tot.pdf. (date accessed 21/05/2013).

Council of Europe 2001. *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Hawkins, John A. & Buttery, Paula (2010) Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* 1(1): 1-23.

Hawkins, John A. & Filipović, Luna (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework (English Profile Studies)*. Cambridge: Cambridge University Press.

Hickmann, Maya (2003) Children's discourse: Person, space and time across languages. Cambridge: Cambridge University Press.

Kantzou, Vicky (2010) *The temporal structure of narrative in the acquisition of Greek as a first and as a second language*. Phd Thesis. Athens: National and Kapodistrian University of Athens. [In Greek]

Kantzou, Vicky (2012) The temporal structure of narratives in second language acquisition of Greek. In: Gavriilidou Zoi, Efthymiou Aggeliki, Thomadaki Evangelia. & Kambakis-Vougiouklis Penelope (eds) *Selected Papers – The 10th International Conference of Greek Linguistics* (pp 354-364) Komotini/Greece: Democritus University of Thrace. Available at: http://www.icgl.gr/files/English/26.Kantzou_10ICGL_pp.354-364.pdf (date accessed 21/05/2013).

Prodeau, Mireille, Lopez, Sabine & Véronique, Daniel (2012) Acquisition of French as a Second Language: Do developmental stages correlate with CEFR levels? *Journal of Applied Language Studies* 6(1): 47–68.

Stamouli, Spyridoula (2010) *Narrative development in Greek L1 and child L2*. Phd Thesis. Athens: National and Kapodistrian University of Athens. [in Greek]

Varlokosta, Spyridoula & Triantafillidou, Leda (2003). *Proficiency Levels in Greek as a Second Language*. Athens: Centre for Intercultural Education, University of Athens. [in Greek]

Evaluating the use of idioms in an L1 learner corpus

Abel, Andrea; Glaznieks, Aivars; Blaschitz; Verena
European Academy of Bozen/Bolzano
andrea.abel@eurac.edu; aivars.glaznieks@eurac.edu; verena.blaschitz@eurac.edu

A typical characteristic of language is its highly formulaic nature: “native speakers of any given language know that there are certain preferred formulations of it” (Wray 2006). Even though semantically transparent, collocations, for example, have to be memorized according to their common use by native speakers (e.g. *to brush one’s teeth*). Such fixed collocations restrict the use of other potential combinations (e.g. *to wash one’s teeth*) which could also transport the same semantic content. As a specific type of collocation, idioms are characterized by the fact that the content of the entire construction is semantically not transparent, i.e. the meaning of the idiom cannot be directly computed from its parts (cf. Burger 2010, Pawley & Syder 1983). This is why deviations from the canonical form often lead to a disruption of the entire idiom, so that the conventionalised meaning of the idiom gets lost. However, depending on the particular idiom, some variant forms are sometimes possible, which increases the difficulty for language learners to comprehend and use idioms properly in a foreign language (Wray 2002, 2006).

However, not only do foreign language learners usually have problems with formulaic language and in particular with idioms, but also native speakers can produce erroneous forms of idioms in spoken and written language when using their native language (L1) (cf. Margewitsch 2006). Errors arise especially in such contexts in which writers have to use a special variety of their L1 that they are about to learn. Academic language (“Bildungssprache”, cf. Gogolin & Lange 2010, Habermas 1981), used in schools and universities, is such a variety. Academic language is not the students’ everyday language but has to be learned for educational purposes. In this context, also native speakers at any age can become learners of a variety of their L1.

The project “KoKo” focuses on writing competences of L1 learners at the end of secondary school. The KoKo corpus contains a collection of more than 1300 argumentative texts on the same topic (more than 700.000 lexical tokens) written in an educational context by native speakers of German. All texts were collected from three different regions of the German speaking area characterized by different diatopic varieties: South Tyrol (Italy), North Tyrol (Austria), and Thuringia (Germany). The analyses comprise different linguistically relevant aspects, such as orthographic, grammatical, lexical, and text competences.

Based on our work on the KoKo corpus, this talk focuses on one particular aspect of lexical knowledge, namely the correct and proper use of formulaic language (cf. Read 2004, Wray 2002, Nation 2001) as it shows in the use of idioms in the KoKo corpus. Our main question is how learner errors can be separated from other forms of variation (diatopic, diachronic etc.). The talk will provide an overview of the recursive approach that was chosen to answer this question. The recursive approach combines an explorative method with an analytic one (Sailer 2007, Moon 2007). Idioms were automatically extracted with the help of a program that allows searching for predefined patterns of collocations (Weller & Heid 2010). The results of the automatic extraction comprise a list of possible formulaic sequences. This list can be sorted for key words of idioms which, in turn, produces a list of all variants of a formulaic sequence in the corpus. To analyse the list of variants manually, a procedure was established that facilitates the discrimination of learner errors from all forms of accepted variants of an idiom, such as diatopic variants, and any other deviation from the common pattern (for example effects of language change). The basis of the procedure is a default hypothesis (H_0), which says that all deviations are learner errors. If any doubt concerning the validity of H_0 arises, e.g. because there is a regional distribution within the corpus, alternative hypotheses (H_1, H_2, …, H_n), such as “the derivation of the canonical form is a diatopic variant”, will be tested in order to falsify H_0 (cf. Abel & Glaznieks in print). Testing an alternative hypothesis may also involve the use of a reference corpus, e.g. the German Reference Corpus (DeReKo).

This approach ensures an accurate testing of possible alternative explanations for deviations from a canonical form. Thus, it helps with making informed decisions about the discussed phenomena and brings forward variation that may not be registered yet in existing reference tools such as dictionaries.

References:

Abel, Andrea. & Glaznieks, Aivars (in print) Wo Sprachkompetenzforschung auf Varietätenlinguistik trifft: Empirische Befunde aus dem Varietäten-Lernerkorpus „KoKo". In: Lenz, Alexandra & Glauninger, Michael (eds) *Variation und Varietäten des Deutschen in Österreich - Theoretische und empirische Perspektiven.* Frankfurt/ Main: Peter Lang.

Burger, Harald (2010) *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt Verlag.

Gogolin, Ingrid & Lange, Imke (2010) Bildungssprache und Durchgängige Sprachbildung. In: Fürstenau, Sarah & Gomolla, Mechthild (eds) *Migration und schulischer Wandel: Mehrsprachigkeit* (pp. 107-127). Wiesbaden: VS-Verlag.

Habermas, Jürgen (1981) Umgangssprache, Bildungssprache, Wissenschaftssprache. In: J. Habermas: *Kleine politische Schriften I–IV* (pp. 340-363). Frankfurt am Main: Suhrkamp.

Margewitsch, Erika (2006) *Formelhafter Sprachgebrauch in Schülertexten.* Oldenburg: Didaktisches Zentrum.

Nation, Paul (2001) *Learning vocabulary in another language.* Cambridge: CUP.

Read, John (2004) Plumbing the depths: How should the construct of vocabulary knowledge be defined? In: Bogaards, Paul & Laufer, Batia (eds) *Vocabulary in a second language.* Amsterdam: John Benjamins, 209-228.

Pawley, Andrew & Syder, Frances (1983): Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, Jack C. & Schmidt, Richard (eds) *Language and Communication* (pp. 191-226). London: Longman.

Moon, Rosamund (2007) Corpus linguistic approaches with English corpora. In: Burger, Harald et al. (eds): *Phraseologie* (pp. 1045-1059). Vol. 2. Berlin: de Gruyter.

Sailer, Michael (2007) Corpus linguistic approaches with German corpora. In: Burger, Harald et al. (eds): *Phraseologie* (pp. 1060-1071). Vol. 2. Berlin: de Gruyter.

Weller, Marion & Heid, Ulrich (2010) Extraction of German multiword expressions from parsed corpora using context features. *LREC* 2010.

Wray, Alison (2002): *Formulaic Language and the Lexicon.* Cambridge: CUP.

Wray, Alison (2006) Formulaic Language. In: Brown, Keith (ed) *Encyclopedia of Language and Linguistics* (pp. 590-597). 2nd edition, Vol. 4. Oxford: Elsevier.

Contrastive Interlanguage Analysis: A Reappraisal

Sylviane Granger
Centre for English Corpus Linguistics
Université catholique de Louvain
sylviane.granger@uclouvain.be

Contrastive interlanguage analysis (CIA) is undeniably the most popular method used to analyse learner corpora. Introduced exactly twenty years ago under the name of 'comparative interlanguage research' (Granger, 1993), it was relabelled 'contrastive interlanguage analysis', abbreviated as CIA, in 1996. Since then it has generated numerous studies involving a highly diversified range of learner populations in an increasingly larger number of languages. In the twenty years since the emergence of CIA, the general approach to language – and English in particular – from an applied perspective has undergone significant changes. It is therefore time to revisit the method and reappraise it in the light of these changes. In my presentation I will tackle the two branches of the CIA method: comparison of native and non-native language and comparison of several non-native varieties. The criticisms that have been levelled at each type will be addressed, with a particular attention to the highly controversial issue of the native speaker norm (Mukherjee, 2005). I will also explore why the CIA approach and learner corpus research in general have failed to have a significant impact on the field of second language acquisition. This critical survey will lead to the presentation of a revised version of the CIA model, referred to as CIA^2. In the last part of my presentation I will broaden the perspective by advocating an integrated approach to 'crosslingual varieties', i.e. varieties of language that have specific characteristics due to the interplay of two or more languages. Foreign/second learner varieties clearly qualify as crosslingual, but so do other varieties, in particular lingua franca varieties (Cogo & Dewey, 2012) and translated language (Olohan, 2004). Investigations of the same linguistic phenomena in different crosslingual varieties could help us identify their shared and unique characteristics, thereby contributing to a better understanding of each variety.

References

Cogo, A., & Dewey, M. (2012). *Analyzing English as a lingua franca: A corpus-driven investigation.* London, UK: Continuum.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan & N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation.* Rodopi: Amsterdam & Atlanta, 57-69.

Granger, S. (1996). From CA to CIA and back: an integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*, Lund Studies in English 88. Lund University Press: Lund, 37-51.

Mukherjee, J. (2005). The native speaker is alive and kicking – Linguistic and language-pedagogical perspectives. *Anglistik* 16/2, 7-23.

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

L2 aqcuisition of temporality: Universal or specific? Findings from a corpus based study of the grammatical encoding of past time in L2 Norwegian

Gujord Ann-Kristin Helland
University of Bergen
Ann-Kristin.Gujord@lle.uib.no

This paper presents findings from a PhD study of the grammatical encoding of past time in L2 Norwegian. In Norwegian, the notion of past is grammaticalised through two categories, the preterite and the perfect, which are the main structures that have been addressed in the study. The overall aim was to explore the grammatical encoding of past time in texts written by Vietnamese (*N*=99) and Somali (*N*=97) learners of Norwegian. The texts are extracted from an electronic learner corpus of Norwegian (ASK), and the texts are assessed to be at two levels of proficiency; A2 and B1.

The learners' grammatical encoding has been explored from two principally different theoretical positions: one which emphasises the universal, common path of the acquisition of tense and aspect morphology in the L2, and one which stresses the importance of influence from previously acquired languages in L2 acquisition. The overall aim of the thesis was firstly to empirically investigate the role of L1 influence in the learners' grammatical encoding of past time in Norwegian. This part of the study relates to Jarvis's (2000) methodological framework for investigating transfer effects. The second aim was to investigate the role of verb semantics as described in The Aspect Hypothesis—in particular, to examine whether the predictions concerning the role of telicity in acquisition of L2 verbal morphology hold for the current interlanguage data. Finally, the study also aimed to investigate whether there is interaction between influence from the learner's L1 and verb semantics, as described in some previous studies (e.g. Collins 2002, 2004). In order to reach these aims, three research questions and associated hypotheses have been examined:

1. *L1-influence:* Do the Vietnamese and the Somali learners display a pattern in their use/non-use of the present perfect and preterite in Norwegian that points to within-group similarities, between group differences and cross-language congruity?

1.1 The Vietnamese-speaking learners will use the present perfect correctly more frequently than the Somali-speaking learners will.
1.2 The Somali-speaking learners will have a higher degree of incorrect use of the preterite in contexts where Norwegian requires the present perfect, and a higher degree of incorrect use of the present perfect in preterite contexts, than will Vietnamese-speaking learners.

2. *Lexical aspect:* Do the learners' use of the preterite and present perfect in Norwegian agree with the earlier findings that support the Aspect Hypothesis?

2.1 The Vietnamese-speaking and Somali-speaking learners will have higher verb type proportion in telic verb phrases (achievements and accomplishments) with preterite and present perfect inflection than in atelic verb phrases (states and activities) with preterite and present perfect inflection.
2.2 The Vietnamese-speaking and Somali-speaking learners will have higher verb type proportion in telic verb phrases (achievements and accomplishments) with correct

preterite and present perfect inflection than in atelic verb phrases (states and activities) with correct preterite and present perfect inflection.

3. *Interaction between L1 influence and lexical aspect:* Do the learners' L1s affect the sequence of development of past morphology as described in the Aspect Hypothesis?

3.1 The Somali-speaking learners will have a higher degree of incorrect use with telic verb phrases, in contexts that require the present perfect or the preterite in Norwegian, than will Vietnamese-speaking learners.

The main findings from the analysis can be summarised as follows. Firstly, transfer effects are detected in the analysis; the patterns of use/non-use of the preterite and present perfect in texts written by the L1groups are quite different and they reflect differences in the encoding of time in the L1s. Secondly, lexical-aspectual influence as predicted in the Aspect Hypothesis, which claims the acquisition of past morphology to be influenced by the telicity in verb phrases, is not revealed. Finally, some kind of interaction of influence between the learners' L1s and the temporal content in inflectional categories is detected; however, the precise type of interaction is difficult to discern. These main findings yielded by the analysis of the 196 texts will be presented and discussed against the backdrop of the theoretical perspectives and previous findings addressed in the study.

References

Bardovi-Harlig, Kathleen. (2000) Tense and aspect in language acquistion: Form, meaning and use. *Language Learning* 50 (Supplement 1):xi-491.

Collins, Laura. (2002) The Role of L1 Influence and Lexical Aspect in the Acquistion of Temporal Morphology. *Language Learning* 52 (1):43-94.

———. (2004) The Particulars on Universals: A comparison of the acquisition of tense-aspect morphology among Japanese and French-speaking learners of English. *Canadian Modern Language Review* 61:251-274.

Common European framework of reference for languages: learning, teaching, assessment. (2001) Cambridge: Cambridge University Press.

Jarvis, Scott, and Terence Odlin. (2000) Morphological type, spatial reference, and language transfer. *Studies in second language acquisition* 22:535-556.

Jarvis, Scott, and Aneta Pavlenko. (2008) *Crosslinguistic influence in language and cognition*. New York: Routledge.

Exploring CEFR classification for German based on rich linguistic modeling

Julia Hancke, Detmar Meurers
Universität Tübingen
{jhancke,dm}@sfs.uni-tuebingen.de

The issue The Common European Framework of Reference for Languages (CEFR) has gained a leading role as an instrument of reference for the certification of language proficiency. At the same time, there is increasing interest in a more comprehensive empirical characterization of the relevant linguistic properties of the CEFR levels.

The research reported on in this paper approaches this issue by studying which linguistic properties reliably support the classification of short essays in terms of CEFR levels. Complementing the work on English criterial features and learner language characteristics that is starting to emerge (Hawkins & Buttery 2010; Yannakoudakis et al. 2011), we focus on identifying learner language characteristics of different levels of German proficiency.

Corpus used The empirical basis of our research consists of 1027 professionally rated free text essays from CEFR exams taken by second language learners of German. Each exam level (A1 to C1) is represented by about 200 texts, varying between 8 and 366 words in length (mean length of 121 words). The data has been collected by the project *MERLIN – Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context* (http://merlin-platform.eu).

Features explored We defined a broad set of 3821 features which can be automatically identified using current NLP tools. We primarily use complexity measures from Second Language Acquisition research to model lexical and morphological richness and syntactic sophistication:

At the ***lexical level***, we started by adapting the features discussed for English by Lu (2012) and McCarthy & Jarvis (2010) for German. To measure the depth of lexical knowledge, we implemented a number of features suggested by Crossley et al. (2011). We extracted frequency scores from the lexical database dlexDB (http://dlexdb.de). We computed features of lexical relatedness using GermaNet 7.0 (http://www.sfs.uni-tuebingen.de/lsd), a lexical-semantic resource for German, similar to WordNet (Miller 1995) for English. We added shallow measures of spelling errors in terms of the number of content word types not found in dlexDB and the misspelled words found by Google Spell Check (version 1.1, https://code.google.com/p/google-api-spelling-java).

Our ***morphological features*** for German capture the learner's use of mood, case, and word formation. We automatically extracted tense patterns from the RFTagger (Schmid & Laws 2008) output and included frequency ratios of these pattern as features for our classifier. The tense features might allow more detailed insights into the tenses the learners used at each of the levels.

At the ***syntactic level***, our features are mostly inspired by the measures used for the analysis of syntactic complexity in English (Lu 2010). However, German syntactically differs from English in several relevant respects. For example, German allows subjectless sentences. Thus, while in general the intention behind the English SLA complexity measures can be expressed in terms of the German syntactic structure and categories, the process of adapting and defining syntactic complexity features for German is far from trivial. As basic syntactic vocabulary for German, we made use of the Negra treebank annotation scheme (Skut et al. 1997).

We added ***dependency-based features*** of syntactic complexity that were previously used in second language writing assessment (Yannakoudakis et al. 2011) and readability assessment (Vor der Brück & Hartrumpf 2007; Vor der Brück et al. 2008; Dell'Orletta et al. 2011).

We automatically extracted parse tree rules from the parse trees produced by the Stanford Parser, following Briscoe et al. (2010) and Yannakoudakis et al. (2011), who used a similar feature based on the output of the RASP parser. We used frequency ratios of these parse tree rules as features for our classifier.

Complementing the linguistic syntactic analysis, we also implemented a number of ***shallower language features***. Unigram, bigram and trigram language model scores provide statistical comparisons to a linguistically simpler model based on a news website for children (http://news4kids.de) and a more complex model based on a news website for adults (http://www.n-tv.de).

NLP tools used To automatically identify the lexical, morphological, and syntactic features, we employ a range of NLP tools and resources including Apache OpenNLP (http://opennlp.apache.org), RFTagger (Schmid & Laws 2008), the Stanford Parser (Rafferty & Manning 2008) with the standard German model trained on the NEGRA corpus (http://coli.uni-saarland.de/projects/sfb378/negra-corpus), the SRILM Language Modeling Toolkit (Stolcke 2002), and the lexical database dlexDB (http://dlexdb.de). For dependency parsing we used the *MATE* dependency parser (Bohnet 2010), with the standard model for German (Seeker & Kuhn 2012) trained on the *TIGER* corpus. Before tagging and parsing, a Java API for Google Spell Check was used to reduce problems caused by spelling errors.

Experimental setup On the basis of the 3821 automatically derived features, we trained a classifier using the Sequential Minimal Optimization (SMO) Algorithm as implemented in the *WEKA* toolkit (Hall et al. 2009). We split the dataset into a training and test set by randomly assigning 2/3 of the samples from each class to the training set (721 samples) and 1/3 to the test set (306 samples). As an additional method for evaluation we used ten-fold cross-validation on the whole dataset.

Results The following table provides an overview of the performance of the classifier for the five level (A1–C1) CEFR classification task:

	Accuracy on test set	Crossvalid. on all data
Random baseline	20%	
Majority baseline	32.9%	33.0%
SMO (all features)	57.2%	64.5%
SMO (best features)	62.7%	

The classifier trained with all features achieves an accuracy of 57.2% with the separate training and test set and an accuracy of 64.5% when using cross-validation on all data. Compared to a majority baseline of classifying all samples as the largest class, this is an improvement of 24.3% and 31.5% respectively.

Investigating the notable difference between the test set and the cross-validation results, we identified two issues. Looking at the results of each individual cross-validation fold revealed that there is considerable variance in the results (10.7% between the best and worst performing fold). However, the worst cross-validation fold still had a better result than our test set. This could be an effect to the slightly larger amount of training data available in the cross validation procedure. Another reason for the comparably poor performance on our test set could lie in the uneven distribution of exam types (as opposed to essay grades) across the different CEFR levels.

Examining the performance of individual feature groups with holdout estimation revealed that the lexical (60.5%) and morphological (56.8%) features were the most successful predictors of the

CEFR level. The syntactic features and language modeling scores were not very successful predictors taken on their own (53.6% and 50.0%), but the syntactic features clearly improved the classification in combination with other features groups. Parse rule features and tense features were the least predictive feature groups (49.0% and 38.5%), however, further experiments showed that their indicative power improves when they are encoded as binary instead of as frequency-based features.

The best model was obtained by combining all feature groups and using WEKA's *CfsSubsetEval*, a correlation-based method for feature selection. It included a set of 34 features consisting of syntactic, lexical, language model and morphological indicators and resulted in a classification accuracy of 62.7% on the test set.

References

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 89–97.

Briscoe, T., B. Medlock & O. Andersen (2010). *Automated assessment of ESOL free text examinations*. Tech. rep., University of Cambridge Computer Laboratory.

Crossley, S. A., T. Salsbury, D. S. McNamara & S. Jarvis (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing* 28, 561–580.

Dell'Orletta, F., S. Montemagni & G. Venturi (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. pp. 73–83.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Hawkins, J. A. & P. Buttery (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* .

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

McCarthy, P. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392. URL https://serifos.sfs.uni-tuebingen.de/svn/resources/trunk/papers/McCarthy.Jarvis-10.pdf.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38(11), 39–41. URL http://aclweb.org/anthology/H94-1111.

Rafferty, A. N. & C. D. Manning (2008). Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*. Stroudsburg, PA, USA: Association for Computational Linguistics, PaGe '08, pp. 40–46. URL http://dl.acm.org/citation.cfm?id=1621401.1621407.

Schmid, H. & F. Laws (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, vol. 1, pp. 777–784. URL http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/COLING08/Schmid-Laws.pdf.

Seeker, W. & J. Kuhn (2012). Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation, 3132–3139. Istanbul, Turkey: European Language Resources Association (ELRA)*.

Skut, W., B. Kreen, T. Brants & H. Uszkoreit (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fith Conference on Applied Natural Language*. Washington, D.C. URL http://www.coli.uni-sb.de/publikationen/softcopies/Skut:1997:ASF.pdf.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*. Denver, USA, vol. 2, pp. 901–904. URL http://www.speech.sri.com/cgi-bin/run-distill?papers/icslp2002-srilm.ps.gz.

Vor der Brück, T. & S. Hartrumpf (2007). A semantically oriented readability checker for German. In Z. Vetulani (ed.), *Proceedings of the 3rd Language & Technology Conference*. Poznań, Poland: Wydawnictwo Poznańskie, pp. 270–274. URL http://pi7.fernuni-hagen.de/papers/brueck_hartrumpf07_online.pdf.

Vor der Brück, T., S. Hartrumpf & H. Helbig (2008). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. *Informatica* 32(4), 429—-435.

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, HLT '11, pp. 180–189. URL http://aclweb.org/anthology/P11-1019.pdf. Corpus available: http://ilexir.co.uk/applications/clc-fce-dataset.

Tracing Transfer in Structural Multiword sequences: What Has Keystructure Analysis to Offer?

Ivaska, Ilmari
University of Turku
itivas@utu.fi

This paper introduces an application of the keystructure analysis (Ivaska&Siitonen 2011; Ivaska: in preparation) as a data-driven method to trace potential transfer in learner language. The method is illustrated in an analysis of advanced learner Finnish. The research questions are twofold: 1) How can potential transfer be detected in a data-driven manner? 2) Which structural multiword sequences show potential transfer effects in advanced learner Finnish?

Cross-linguistic influences are said to constitute the most studied and yet least understood field in the L2 studies (Jarvis 2000). Transfer is usually approached either by analysing the language product or the production process (Sajavaara & Lehtonen 1989). The position taken in this paper follows the corpus-driven approach (e.g. Tognini-Bonelli 2001), and the focus is on the product. The attempt is to find statistically significant correlations between learners' language background and features of the target language (Jarvis 2000). This is essentially the concept of the Contrastive Interlanguage Analysis (Granger 1996), as different L1 groups are compared to detect these correlations. Keystructure analysis applies statistical keyness (Scott & Tribble 2006), and correlations are traced by comparing frequencies of structural multiword sequences in different L1 subcorpora. The sequences found are then analysed in terms of their typical use (cf. Francis 1993) to interpret the possible cross-linguistic influences behind them.

In this paper I trace structural bigrams and trigrams that, based on the aforementioned measures, indicate possible transfer in advanced L2 Finnish. The sequences are defined in terms of the morphological forms of each word. The setting is similar to Aarts & Granger (1998) and Wiersma et al. (2011). The frequencies are counted following the skipgram approach, and the words do not have to follow each other immediately as long as they are in the same order (Guthrie et al. 2006). They are then compared statistically with the help of random forests (Breiman 2001) to find the strongest correlations.

The corpus used, The Corpus of Advanced Learner Finnish (Ivaska & Siitonen 2009; Ivaska: in preparation), has been annotated in terms of lemmas, parts-of-speech, morphological forms and syntactic functions. The L1 groups studied are Czech, Japanese, Lithuanian and Russian, with a reference data of L1 Finnish. There are 26 text units from each L1 group, (overall 130 text units and 65,801 tokens). At the time of collection all subjects were studying in a Master's program of Finno-Ugric languages and cultures. The texts were written as part of their studies, and they were not written for language proficiency evaluation.

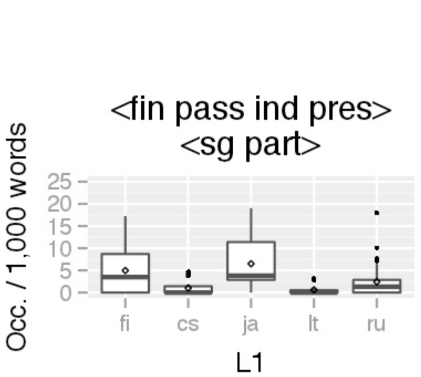


Figure 1. Grequencies of <fin pass ind pres><sg part> bigram.

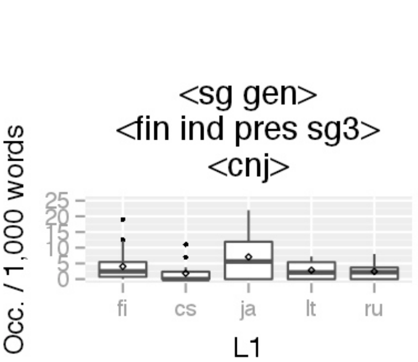


Figure 2. Frequencies of <sg gen><fin ind pres sg3><cnj> trigram.

Preliminary results support the applicability of the method in tracing potential transfer effects in learner language. The statistical analysis indicated several potential keystructures. For example, there is a bigram (consisting of a present indicative passive verb and a singular partitive) and a trigram (consisting of singular genitive, present indicative active verb in singular third person and a conjunction), that are more frequent in the subcorpus of L1 Japanese than in any other subcorpora (bigram means / 1,000 words: fi≈5.8, cs≈1.0, ja≈7.7, lt≈0.6, ru≈2.4; trigram means: fi≈4.1, cs≈1.9, ja≈8.5, lt≈2.8, ru≈2.4) (figures 1 and 2). In the bigrams the frequency of use by L1 Japanese learners is close to that of native speakers, while in the trigrams L1 Japanese learners differ from other learner varieties and native speakers. This may indicate positive or negative transfer, and the possible constructional nature and typical use of these sequences should be analysed and contrasted with the respective L1s to interpret the possible reasons.

References

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In: Granger, S. (ed.) *Learner English on Computer* (pp.132-141). London: Longman.

Breiman, L. (2001). Random forests. *Machine Learning* 45(1): 5-32.

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

Ivaska, I. & Siitonen, K. (2009). Syntaktisesti koodattu oppijankielen korpus: mahdollisuuksia ja ongelmia. In: Eslon, P. & Õim, K. (eds.) *Korpusuuringute metodoloogia ja märgendamise probleemid* (pp. 54-71). Tallinn; Tallinna Ülikool.

Ivaska, I. & Siitonen, K. (2011). Avainrakenneanalyysi: tapa tutkia oppijankielen lauserakennetta korpusvetoisesti. *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 3: 35-47.

Jarvis, S. (2000). Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning* 50(2): 245-309.

Sajavaara, K. & Lehtonen, J. (1989). Aspects of transfer in foreign Language speakers' reactions to acceptability. In: Dechert, H.W. & Raupach, M. (eds.) *Transfer in Language Production* (pp. 35-52). Norwood, NJ: Ablex Publishing Corporation.

Scott, M. & Tribble, C. (2006). T*extual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Wiersma, W., Nerbonne, J. & Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26(1): 107-124.

Transfer in adjective inflection in the interlanguages of English, German and Polish learners of L2 Norwegian – a corpus based research

Janik, Marta Olga
Adam Mickiewicz University in Poznań, Poland; University of Bergen, Norway
martka.janik@gmail.com

The research is a part of my Ph.D.-project which is going to explain the mechanisms behind the acquisition of Norwegian adjective agreement by Polish L1 learners of Norwegian. My project is based on cognitive language theories and it is a part of the ASKeladden-project which aims to explain L1 influences in acquisition of L2 using corpus based methods. In the presentation, however, I shall show how different L1s influence their users' performance in L2 exemplified by adjective inflection in L2 Norwegian performed by speakers of English, German and Polish as L1.

One can say that the adjective has the most intricate inflectional system of all the parts of speech in Norwegian. It must be obligatory inflected in both: attributive and predicative position, and there are 3 inflectional categories in the language: number, gender and definiteness. In spite of the system complexity, there are only 3 morphological endings to choose among: Ø, -e and –t. One would therefore expect no bigger problems in acquisition of the agreement in Norwegian as a second language. It is, however, not a fact and the possible explanation for this can be that there are some additional inflection rules which divide adjectives into different inflectional groups, so the learner must learn which group the adjective belongs to and then use the proper ending.

In addition to this there are probably also occurrences of transfer which influence the performance of users of L2 Norwegian. My study will investigate if the degree of adjective inflection in a language makes its users more or less sensitive (depending on if there is adjective inflection in the language or not) to notice and follow the need of inflecting words in other languages they acquire.

In my investigation I have chosen three quite distinct languages as far as adjective inflection is concerned. English, German and Polish differ namely between each other in the degree of adjective agreement and inflection in general. English has no adjective inflection at all, so there is no grammatical agreement between adjectives and nouns. In German adjectives are inflected in 4 categories: number, gender, definiteness and case, but the agreement appear only in attributive position, and not in predicative position. Polish adjectival agreement is as much complicated as the Norwegian one. Polish adjectives must agree with nouns in both: attributive and predicative positions, and there are 3 inflectional categories of adjectives: number, gender and case.

All data analysed by me are taken from ASK, a language learner corpus of Norwegian as a second language, which contains texts written by speakers of 10 different L1s on 2 proficiency levels. The corpus includes 100 texts of each L1 group on each level (with some exceptions in case of more rare languages). To my investigation I have chosen only three L1 groups: English, German and Polish on both levels; it means that my data comes from 600 texts written by L2 Norwegian learners.

To analyze the data with respect to transfer I have chosen methodology proposed by Jarvis and Pavlenko (2008) and Jarvis (2010). It builds on 4 types of evidence: intragroup homogeneity, intergroup heterogeneity, cross-language congruity and intralingual contrasts. One can state transfer if all of the four types of evidence are confirmed.

Using this method I hope to identify negative transfer in L1 English and German learners of L2 Norwegian. I expect to find most errors in the texts written by English speaking learners. In the case of German speakers I expect to find missing agreement in predicative position. The Polish speakers are expected to perform more correctly than the two other groups as one can reckon with positive transfer which could be explained by similarities in adjective inflection pattern in Norwegian and Polish. Especially I do not expect to find any difference between inflecting adjectives in attributive and predicative position in case of the L1 Polish speaking learners. Such results would constitute evidence of crosslinguistic influence in all of the groups at the same time, but in a different scope.

References

ASK's homepage: http://gandalf.uib.no/ask/

Jarvis, Scott. (2010). Comparison-based and detection-based approaches to transfer research. In: Roberts, Leah, Martin Howard, Muiris Ó Laoire & David Singleton (eds.), *EUROSLA Yearbook 10* (pp. 169-192). Amsterdam: Benjamins.

Jarvis, Scott & Pavlenko, Aneta. (2008). *Crosslinguistic Influence in Language and Cognition*. New York & London: Routledge.

Applying keyword analysis to annotated and CEFR analyzed learner data: positive and negative key items

Jantunen, Jarmo Harri
University of Jyväskylä
jarmo.jantunen@jyu.fi

The present paper links corpus-driven methodology to lemmatized, grammatically annotated and CEFR analyzed learner data. It especially focuses on methodology, but also reveals which lexical and grammatical items are specific to certain levels of L2 proficiency. Keyword analysis has rarely been used to analyze grammatically annotated data, and, especially, to analyze tagged learner data. The present paper illustrates the over- and underused items in learner data; these include grammatical tags, topic keywords, and tentative learner language keywords. The present paper is a follow-up study for an earlier investigation (Jantunen 2011a) in which the data was not yet CEFR analyzed and a pilot study (Jantunen 2012) in which only texts produced by Estonian learners of Finnish were analyzed. Unlike Martin et al. (2010) who studied the frequency, distribution and accuracy of three linguistic items at different proficiency levels, the present study only focuses on frequency and distribution, but also takes several linguistic and lexical items into account.

The data comes from the *International Corpus of Learner Finnish* (ICLFI), which consists of texts produced by Finnish language learners at universities outside Finland. According to the learner corpus typology (Jantunen 2011b), ICLFI is a multi-L1, multi- proficiency-level, multi-genre, monolingual, non-translational, partly diachronic, whole text learner language corpus, which consists of text written in classroom instruction in foreign language teaching context. At the moment, 70% of the data have been assessed according to the CEFR scale from A1 (basic user) to C2 (proficient user) and 25 % is grammatically annotated and lemmatized. The present study makes use of texts written by students whose mother tongues are Estonian, Russian, Swedish and Dutch. In this data the scales vary from A2 to C1. At the moment, the total size of the annotated data is about 254.000 tokens, subcorpus of the Estonian L1 being the biggest (125.000 tokens). The data are annotated and lemmatized semi-automatically using Connexor functional dependency grammar parser (Fi-fdg, Järvinen et al. 2004).

In the analysis, CEFR-analyzed subcorpora are compared with each other (i.e. A2 with B1, etc.), a process which reveals the differences between proficiency levels. At the second stage, the comparison is made between subcorpora of different mother tongues, which should reveal whether learners’ L1 has an impact on the key items in learner data. Finally, the same subcorpora are placed in a comparison with native data in order to complete the picture of learners' L2 development and thereby to illustrate the nature of learner production compared to native production.

The data are analyzed using WordSmith Tools KeyWords program (Scott 2008). The analysis reveals that data yield a complex picture of L2 production: Since the keyword analysis picks up word forms, lemmas, tags and other linguistic meta-information from the learner corpus data, the description of the features at defined developmental stages contains both lexical and grammatical information. This diverse information benefits the description of developmental stages in L2 learning offering both qualitative and statistical information on proficiency levels and L1 transfer. The statistical information reveals that certain grammatical items and lexical elements are either under- or overused at certain proficiency levels; this is illustrated by both positive and negative key items. The analysis also shows that in learner writing there exist certain learner language keywords, i.e. specific lexical items (lexical teddy bears, cf. Hasselgren 1994) that are favored by language learners.

References

CEFR = *Common European Framework of Reference for Languages: learning, teaching, assessment* 2001. Council of Europe.

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

Jantunen, Jarmo Harri (2011a) Avainsana-analyysi annotoidun oppijankieliaineiston tutkimisessa: Alustavia havaintoja [Keyword analysis in investigating learner data. Preliminary findings.] In: Lehtinen, Esa. et al. (eds.) *AFinLA-e: Soveltavan kielitieteen tutkimuksia [AFinLA-e: Applied Linguistics Research]* 3 (pp. 48-61). Available: http://ojs.tsv.fi/index.php/afinla/article/view/4456.

Jantunen, Jarmo Harri (2011b) Kansainvälinen oppijansuomen korpus (ICLFI): Typologia, taustamuuttujat ja annotointi. [International Corpus of Learner Finnish: Typology, variants and annotation.] *Lähivõrdlusi. Lähivertailuja* 21: 86-105.

Jantunen, Jarmo Harri (2012) Linking corpus-driven methodology to annotated and CEFR analyzed learner data: a profitable synergy? *CTAL2012 Conference*, 28-30 June 2012, Suzhou, China.

Järvinen, Timo, Laari, Mikko, Lahtinen, Timo, Paajanen, Sirkku, Paljakka, Pirkko, Soininen, Mirkka & Tapanainen, Pasi (2004) Robust Language Analysis Components for Practical Applications. In: Gambäck, Björn & Jokinen, Kristiina (eds.) *Coling 2004*, *Proceedings of the Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces* (pp. 53-56). Geneva: Coling.

Martin, Maisa, Mustonen Sanna, Reiman Nina & Seilonen Marja (2010) On becoming an independent user. In: Bartning, I. et al. (eds.) *Communicative proficiency and linguistic development*. *Intersections between SLA and language testing research*. (pp. 57-79). Available: http://eurosla.org/ monographs/EM01/57-80Martin_et_al.pdf

Scott, M. (2008) Developing WordSmith. *International Journal of English Studies* 8(1): 95-106.

Signals and clues in detecting crosslinguistic influence: What detectives and detectors can tell us

Scott Jarvis
Ohio University
jarvis@ohio.edu

Language corpora contain clues concerning the identities of the people who produced the texts in the corpus, and they also contain clues about those people's backgrounds, particularly of the L1 and L2 discourse communities they have been a part of. Some of these clues are found in discrete bits of information (e.g., the use of a particular word or a unique type of error), but others are detectable only as part of a larger constellation (e.g., where the use of a word or structure in one part of the text has consequences for the use of other words and structures in other parts of the text). When human judges are asked to identify writers' backgrounds based on their patterns of writing, some human judges are surprisingly good detectives, knowing which clues to focus on and which to ignore. Some human judges also have well-developed intuitions that allow them to detect writers' backgrounds very quickly without an apparent awareness of which clues they have relied on.

Whether human judges rely on detailed analyses or on intuition, high rates of success in identifying writers' backgrounds (such as their L1 backgrounds) serve as evidence of the ubiquity and reliability of the clues in data. High rates of detection accuracy also serve as evidence of the effects of writers' backgrounds on their writing. The advantages of human judges is that they take all available clues into consideration at the same time and flexibly adjust the priority they give to different clues. However, these advantages turn into disadvantages when the researcher wants to determine the strengths of individual clues or classes of clues. For this, it is best to turn to machine-automated analyses—or detectors—which focus on only certain clues at a time, and are blind to all others. In this paper, I illustrate and discuss the complementary advantages of human judges and machine classifiers in the detection of L1 influence in L2 learner corpora.

The development of formulaic repertoires in L2 English at three CEFR levels: a corpus-driven and cross-linguistic comparison

Sylvia Jaworska; Angeliki Salamoura; Fiona Barker
Queen Mary, University of London; Cambridge English; Cambridge English
s.jaworska@qmul.ac.uk; Salamoura.A@cambridgeenglish.org; Barker.F@cambridgeenglish.org

Formulaic language seems to be central to successful foreign language learning. However, it also presents a serious stumbling block. Research examining formulaic sequences in L2 English has demonstrated that learners tend to underuse, overuse or misuse native-like expressions and have, overall, a much smaller repertoire of collocations (De Cock 1998; Granger 1998; Laufer & Waldman 2011). Despite the considerable interest in formulaicity, to date most studies focused exclusively on advanced learners in academic contexts. With the exception of Vidaković & Barker (2010), little is known about the development of formulaic repertoires at other proficiency levels. Moreover, given that lexical choices are particularly prone to L1 transfer (Jarvis 2000), far too little attention has been paid to the influence of L1. The aim of this paper is to report findings from a study concerned with the development of formulaic repertoires at the different proficiency levels as defined by the Common European Framework of Reference for Languages (CEFR). The main objectives that our research addresses are:

R1: Are there any significant distributional and functional differences in formulaicity at the different stages of proficiency?

R2: Can any significant distributional and functional differences be detected depending on the learners' L1.

In doing so, our research intends to offer empirical data which could be used to refine the existing descriptions of the CEFR levels for English and assist teachers and publishers in producing L1-tailored teaching and exam materials. Our methodology follows a corpus-driven design (Biber et al. 2004, Chen & Baker 2010). It is based on a quantitative and qualitative examination of 3- and 4-word sequences produced by learners of different L1s (Arabic, Chinese, German, Greek, Korean and Polish). The data under scrutiny consists of written responses obtained from three widely taken exams FCE (B2), CAE (C1) and CPE (C2), and are from the Cambridge Learner Corpus. We investigated the above research questions in a pilot study, which was based on a smaller sample of our data. It included responses produced by learners of L1s of Indo-European origin (German, Polish and Greek) and at two proficiency levels B2 and C1 (see Table 1).

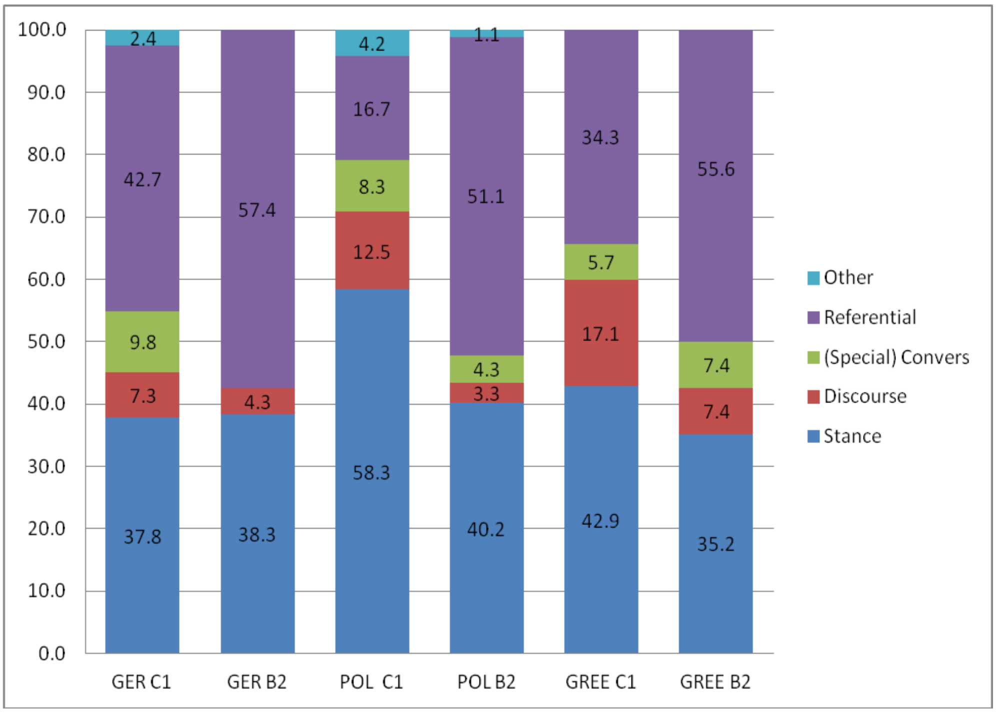
Table 1: Size of the pilot corpus

Learner's L1	Tokens	Types	B2 level (tokens)	C1 level (tokens)
German	12,481	2,083	4,699	7,387
Polish	15,930	2,531	8,127	3,888
Greek	10,967	1,759	6,236	2,782
Total	39,378	6,373	19,062	14,057

For the pilot analysis, we retrieved 4-word sequences that were subsequently categorised according to their main structure (verb-based, noun-based or prepositional phrase) and function (referential, discourse-structuring, stance or special conversational expression) (Biber et al. 2004).

Our preliminary findings suggest that there are considerable differences in the use of 4-word combinations amongst learners of different L1s. These differences concern, in particular, the types and functions of sequences. As the functional analysis revealed, Greek learners use the highest frequency and diversity of discourse devices, which could be attributable to L1-induced stylistic interference (Koutsantoni 2005, Sidiropoulou 2012,). Functional differences could also be detected at the two proficiency levels and across the three learner groups. As Table 2 demonstrates, at the lower proficiency level, all learners rely heavily on referential sequences. With growing proficiency (C1), they employ more discourse-structuring expressions with Greek learners using the highest proportion of such devices.

Table 2: Functional distribution of B2 vs. C1 levels



The structural analysis did not point to any significant differences, which is perhaps attributed to the fact that the three L1s under investigation belong to the same language family.

In the present study, we intend to corroborate our preliminary findings by analysing 4-word-sequences in a larger data set including three CEFR levels (B2, C1 and C2) and responses from learners whose L1 is not of Indo-European origin (Arabic, Chinese and Korean). Our findings will be discussed and linked together with the criterial-features research (Hawkins & Filipović 2012).

References

Biber, Douglas, Conrad, Susan. & Cortes, Viviana (2004) If you look at …: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25(3): 371-405.

Chen, Yu-Hua. & Baker, Paul (2010) Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2): 30-49.

De Cock, Sylvie (1998) A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3(1): 59-80.

Granger, Sylviane (1998) Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P. Cowie, A. P. (ed.) *Phraseology* (pp. 145-160). Oxford University Press, Oxford: 145-160.

Hawkins, John A & Filipović, Luna (2012) *Criterial Features in L2 English*. Cambridge, UCLES/Cambridge University Press.

Jarvis, Scott (2000) Methodological Rigour in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. *Language Learning* 50(2): 245-309.

Koutsantoni, Dimitra (2005) Greek Cultural Characteristics and Academic Writing. *Journal of Modern Greek Studies* 32(1): 97-138.

Laufer, Batia. & Waldman, Tina (2011) Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. Language Learning 61(2): 647-672.

Sidiropoulou, Maria (2012) Greek and English linguistic identities in the EU: A translation perspective. *Pragmatics and Society* 3(1): 89-119.

Vidaković, Ivana. & Barker, Fiona (2010). Use of words and multi-word units in Skills for Life Writing examinations. *Cambridge ESOL: Research Notes* 41: 7-14.

“A chi-square test showed that...” – or did it really? Some remarks on the use of statistical tests in corpus-based research

Bård Uri Jensen
Hedmark University College, Norway
http://privat.hihm.no/buj/
bard.jensen@hihm.no

There has been a steadily growing interest in statistical methods in linguistics for some time now. Statistical methods are an important supplement to other methods in linguistics, both as hypothesis generators and as hypothesis tests.

Hypothesis testing methods express the probability that a tendency found in a sample reflects properties in a population. When we choose a level of significance for a hypothesis test, often $\alpha = 0.05$ in the humanities, we accept a 5 % risk of falsely concluding about a positive result. In a typical humanities context, this is often an acceptable risk.

All hypothesis testing methods, both parametric and non-parametric ones, have certain conditions of use. If the conditions are not fulfilled, the risk of erroneous conclusions will shift either upwards or downwards. If the conditions are not fulfilled, we consequently don't know what risk of false conclusions we are running. That is not acceptable and leaves the use of the tests virtually worthless.

In this speech I will present and discuss conditions for some commonly used statistical tests. I will also present conclusions from recently published research within corpus-based linguistics and discuss whether these conclusions are valid or not, due to possible misapplication of the commonly used statistical tests that the conclusions are based on.

I will concentrate on two tests which are often regarded as ‘simple’ or ‘unadvanced’, the chi-square test and the t-test. These may be simple and unadvanced, but they nevertheless have conditions attached to them. And these conditions can still be violated and are indeed violated quite frequently in internationally published research. This implies that peer-reviewed, internationally published research in many cases makes conclusions which are presented as mathematical truths, but in reality are not much more than guesswork.

A Longitudinal Corpus Study of Teenage EFL Learners' Vocabulary

Jiménez Catalán, Rosa Mª
University of La Rioja, Spain
rosa.jimenez@unirioja.es

Learner corpora as well as Second Language Acquisition (SLA) researchers have looked at interlanguage by means of synchronic research rather than longitudinal research (Granger, 2002). A brief overview with regard to thematic issues related to language learners' lexical acquisition and development yielded the following results: verb tense (Faunier and Littre, 2013), intensifying adverbs (Pérez-Paredes and Díez-Bedmar, 2012), relative pronouns (Byrnes and Sinicrope, 2008), lexical networks **(**Scott, Crossley, and Salsbury, 2009), lexical phrases (Myles, Hooper and Mitchel 1998; Li and Schmitt, 2009), lexical richness (Housen, Bulté, Pierrard and van Daele, 2008), word order (Grümpel, 2009), and articles (Liu and Gleason, 2002). As far as gender and vocabulary are concerned, we find longitudinal studies on learners' lexical errors (Agustín 2010, 2009), word associations (Moreno, 2009) and vocabulary patterns (Ojeda and Jiménez Catalán, 2010).

A content analysis of the literature reveals that most learner corpus studies have paid more attention to language learners in university rather than to language learners in school contexts. This study seeks to contribute to narrow the gap in learner corpus research. We provide a longitudinal corpus study of EFL learners' vocabulary growth throughout 4th, 5th and 6th years of primary education in two types of instruction (English as vehicular language versus English as curricular subject) and three years of secondary education (7th, 8th and 9th grade). The specific research questions are as follows:

(1) Is there a gradual increase in the number of words used in compositions by EFL learners over six school grades?
(2) If such increase is observed, do male and female EFL learners behave in the same way?

The participants wrote a letter in English in six points of time in order to measure their vocabulary gains. After each collection time, the compositions were typed into the computer and submitted to Wordsmith Tools. We classified the words according to alphabetical order and word frequency. We completed this analysis by calculating the means of tokens and types and by running statistical tests.

Our preliminary findings point to: (1) an increase in the number of word types and tokens as the school grade increases; (3) however, the rate of increase is uneven throughout school years; and (2) girls achieve a significantly higher means in tokens and types in all time points.

References

Agustín Llach, MP. (2009) *Gender differences in vocabulary acquisition in foreign language learning in primary education*. Universidad de La Rioja.

Agustín Llach, MP. (2010) Exploring the role of gender in lexical creations. In Jiménez Catalán (ed.) *Gender Perspectives on Vocabulary in Foreign and Second Languages* (pp: 74-92). Palgrave Macmillan.

Byrnes, H. and Sinicrope, C. (2008) Advancedness and the development of relativatizacion in L2 German: A curriculum-based longitudinal study. In Lourdes Ortega and Heidi Byrnes (eds.) *The Longitudinal Study of Advanced L2 Capacities* (pp. 109-139). New York: Routledge.

Faunier, F. and Littre, D. (2013) Tracing Learners' Progress: Adopting a Dual 'Corpus cum Experimental Data' Approach. *The Modern Language Journal*, vol. 97 (1) 61-76.

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

Grümpel, C. (2009) Interlanguage Studies in a Cross Cultural Context: The Interlanguage of Spanish Speakers (L1) in an Approach English (L2), German (L3). *Revista Alicantina de Estudios Ingleses* 22, 315-326

Housen, A., Bulté, B., Pierrard, M. and van Daele, S. (2008) Analysing lexical richness in French learner language. In. Treffers-Daller et M4 (eds.) *Journal of French Language Studies* 18 (3), 277 – 298.

Li, J. and Schmitt, N. (2009) The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing* 18, 85-102.

Liu, D., & Gleason, J. L. (2002) Acquisition of the article the by nonnative speakers of English. *Studies in Second Language Acquisition* 24, 1-26.

Moreno, S. (2009) Boys' and Girls' L2 Word Associations. In Jiménez Catalán, R. (ed.) *Gender Perspectives on Vocabulary in Foreign and Second Languages* (pp. 139-166). Palgrave Macmillan.

Myles, F., Hooper, J., and Mitchel, R. (1998) Rote or Rule? Exploring the Role of Formulaic Language in Classroom Foreign Language Learning. *Language Learning*, vol. 48 (3) 323-364.

Ojeda, J. and Jiménez Catalán, R. (2010) Vocabulary Gender Patterns in EFL Compositions. *Porta Linguarum* 13, 9-28.

Lexical bundles in written learner English: the case of Lithuanian learners

Juknevičienė, Rita
Vilnius University
rita.jukneviciene@takas.lt

Corpus studies of recurrent word sequences have outlined new directions in learner language research. The fact that naturally produced English consists of prefabricated multi-word units gave rise to the question of chunkiness of learner language. This study was designed to investigate lexical bundles in written language produced by learners of two different levels of proficiency. The definition and interpretation of lexical bundles draws on corpus studies of English (Altenberg 1998; Biber et al. 1999; Biber et al. 2004; Hyland 2008). Apart from the term *lexical bundle*, these multi-word expressions have also been referred to as *recurrent sequences* (De Cock 2004), *chunks* (O'Keeffe et al. 2007), *clusters* (Scott 2008), or *n-grams* (Römer 2009). The major focus of this study is on the clause structure and clause segments which tend to cluster and thus form recurrent sequences in learner corpora that represent learners of different proficiency. Undertaken as a corpus-driven analysis, the study also re-addresses the question of what becomes a recurrent sequence in learner language and argues for a more cautious methodological approach to learner corpus material.

The data for this study comes from two corpora of learner English, representing Lithuanian EFL learners. The first corpus (the NEC corpus) consists of secondary school leavers' examination scripts written during the national English examination. This is a new corpus project in Lithuania developed as a resource for test development and analysis of learner language. At its present pilot version, it consists of ca. 150,000 words. The second corpus used here (LICLE) was compiled as an ICLE component (Granger et al. 2009). It represents written English of English Philology students in the senior years of study. Its subcorpus used for this study consists of ca. 190,000 words. The research method involves a contrastive analysis of automatically retrieved sequences of 4-7 words from the two corpora. The sequences were analysed in terms of the clause segments that they span.

The results of the analysis revealed several tendencies. Firstly, written language produced by less proficient learners contains a larger proportion of repetitive lexical strings, which is interpreted as an indication of their limited lexical repertoire and thus more frequent use of certain strings of words. Secondly, the structural analysis showed that learners of different proficiency levels tend to repeat, or cluster as a bundle, different segments of the clause. The language of less proficient learners contains more recurrent sequences that incorporate full sentence stems and predicates. In general, recurrent sequences in LICLE are units that indicate the subsequent complementation pattern. In contrast, sequences in the corpus of intermediate learners predominantly end in a lexical word and contain no indication that the learners have acquired the complementation pattern of the last word in the sequence.

References

Altenberg, Bengt (1998) On the phraseology of spoken English. The evidence of recurrent word-combinations. In: Cowie, Anthony Paul (ed.) *Phraseology. Theory, Analysis, and Applications* (pp. 101-122). Oxford: Clarendon Press.

Biber, Douglas; Conrad, Susan & Cortes, Viviana (2004) If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371-405.

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan & Finegan, Edward (1999) *The Longman Grammar of Spoken and Written English*. London: Longman.

De Cock, Sylvie (2004) Preferred Sequences of Words in NS and NNS Speech. *BELL (Belgian journal of English language and literature)*: 225-246.

Granger, Sylviane; Dagneaux, Estelle; Meunier, Fanny & Paquot, Magali (2009) *International Corpus of Learner English (v2). Handbook, CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

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

O'Keeffe, Anne; McCarthy, Michael & Carter, Ronald (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.

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

Scott, Mike (2008) *Wordsmith Tools. Version 5*. Oxford: Oxford University Press.

Cross-linguistic influence and second language acquisition: corpus-based research

Kaivapalu, Annekatrin
Tallinn University
kaivapa@tlu.ee

The aim of the poster is to introduce the research project *Cross-linguistic influence and second language acquisition: corpus-based research* of Tallinn University (2010-2013, founded by the Estonian Science Foundation). The project addresses the fundamental question of how cross-linguistic influence, especially first language (L1) influence, determines second and foreign language (L2) acquisition and learning (SLA) and is based on parellel use of two Finno-Ugric learner language corpora, *The Estonian Interlanguage Corpus* (EIC) in Tallinn University (http://evkk.tlu.ee) and *the International Corpus of Learner Finnish* (ICLFI) in Oulu University (http://www.oulu. fi/oppijankieli).
The project has following goals:

1) to examine morphological, morphosyntactical and lexical cross-linguistic influence and to find out differences between closely related and unrelated L1 influence on SLA.
2) to specify the role of the first and the second language influence in the acquisition of a third language, while one of the source languages is related to the target language and the other one is not.
3) to investigate relationships between the first language influence and second or foreign language proficiency;
4) to find out factors interacting and competing with L1 influence in the acquisition and processing of L2;
5) to investigate the processing strategies of the learners in closely related and unrelated L1.

The project applies usage-based approach to language acquisition (Bybee 2010). The Construction Grammar -type approach to language serves as a loose overall theoretical framework which connects the individual studies. For tracing the development of second language proficiency, described in terms of fluency, accuracy and complexity (Housen et al 2012), the DEMfad model (Martin et al 2010) is used. In each Domain (D), such as a given linguistic structure or area of vocabulary, the Emergence (E), the first appearance, and the Mastery (M, at the 80 % level of accuracy) is determined. The three parameters used are frequency (f) of occurrence (per 1000 tokens), accuracy (a), as the percentage of target like occurrences, and distribution (d).
The main question of the project is how the emergence and expansion are influenced by the learners' first or other formerly acquired language. According to the Competition Model (MacWhinney 2005), second and foreign language learning is seen as a creation of a new set of grammatical categories, language learning is a process which consists of several factors leading in different directions. The L1 influence is one of the factors in SLA which interacts with others, such as the structures of the languages involved. A comparison of the structures of L1 and L2 makes it possible to find the interfaces between the two languages which are a prerequisite for L1 influence.
The first languages involved in the study are Estonian, Finnish and Russian. The methodological starting point of the project is the framework for investigating the first language influence envisioned by Jarvis (2000: 249–261) and successfully used in former studies (e.g. Kaivapalu & Martin 2007). Writing samples of both corpora are rated by three experienced raters on a scale of *Common European Framework of Reference for Languages: Learning, teaching and assessment* (CEFR).

The poster introduces project participants' studies on the following topics:

- Symmetry of the cross-linguistic influence in the acquisition of closely related languages
- L1 influence and its psycholinguistic reality in closely related languages
- One-to-many mapping between closely related languages and its influence on second language acquisition
- The influence of non-related L1 Russian and closely related L2 Estonian on the acquisition of grammatical tenses of L3 Finnish
- Writing process of L2 Estonian by Russian and Finnish learners
- The expression of grammatical aspect in learner Estonian and Finnish by Finnish and Estonian learners

References

Bybee, Joan (2010) *Language, usage and cognition*. Cambridge: CUP.

Housen, Alex; Kuiken, Folkert & Vedder, Ineke (eds.) 2012. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: Benjamins.

Jarvis, Scott (2000) Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. *Language Learning* 50 (2): 245-309.

Jarvis, Scott (2010) Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook* 10: 169-192.

Kaivapalu, Annekatrin & Martin, Maisa (2007) Morphology in transition: The plural inflection of Finnish nouns by Estonian and Russian learners. *Acta Linguistica Hungarica* 54 (2): 129-156.

MacWhinney, Brian (2005) A unified model of language aquisition. In: Kroll, Judith F. & de Groot, Annette (eds.) *Handbook of bilingualism: Psycholinguistic approaches* (pp.49-67). Oxford: Oxford University Press.

Martin, Maisa, Mustonen, Sanna, Reiman, Nina & Seilonen, Marja (2010) On becoming an independent user. In: Bartning, Inge, Martin, Maisa & Vedder, Ineke (eds.) *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp.57-80). Eurosla Monographs Series 1.

Ringbom, Håkan (2007) *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters LTD.

Use of *of* by Japanese English learners

Kaneko, Tomoko
Showa Women's University
' kaneko@swu.ac.jp

Although a high proportion of prepositions are associated with written rather than spoken language, prepositions are significantly underused in non-native written language corpora compared to the native corpora (Granger and Rayson 1998:126). Japanese EFL Learner (JEFLL) Corpus compiled from data collected from high school English learners shows that the frequency of *of* is less than half of that used by the native speakers in BNC (Uchida 2007). Uchida also shows that *in* is the most frequently used preposition and *of* is the second in JEFLL, while the frequency order of these two prepositions are reversed in BNC. The International Corpus of Learner English (ICLE) consists of sub-corpora of argumentative essay writings by advanced level English learners with various language backgrounds. Japanese ICLE sub-corpus still shows that *of* is left behind in its frequency of use compared to the other prepositions. For example, in the ICLE Japanese sub-corpus, its frequency is in 5th place among the whole vocabulary used, even though it is in second place in the Louvain Corpus of Native English Essays (LOCNESS) as well as the ICLE French and Italian sub-corpora.

Based on this observation, this paper compares the use of *of* in ICLE Japanese sub-corpus to French, German, and Italian sub-corpora as well as LOCNESS to find the gap among them and contemplate why Japanese English learners underuse *of*. First, each corpus was analyzed using WordSmith tools and then the concordance lines centering on *of* were listed. The Oxford Advanced Learner's Dictionary distinguished nine different meanings of the preposition *of*. For example, 1) belonging or relating to somebody, 2) belonging to or being part of something, 3) concerning or showing somebody or something, 4) used with measurements and expressions of time, age, etc. 5) used to show somebody or something belongs to a group, often after *some, a few,* etc., and so on. Since Japanese learners mainly use 'Noun (N) *of* Noun (N)' structures in the corpus, other strings of phrases centering on *of* were not included in the present study. Each use of *of* in 'N *of* N' structures was categorized into 9 meanings and frequencies of each use are compared.

The result shows that Japanese learners overuse *of* in the meaning No. 3, while they underuse *of* in the meaning No.4. The French, German, and Italian learners of English, on the other hand, have a tendency to overuse *of* in the meaning No. 4. They also overuse *of* in the meaning No. 5. It is also found that, contrary to the learners with European language background, Japanese learners use *of* in the meaning No. 2 more often than in the meaning No. 1.

Biber et al. (1994) found that although prepositional postnominal modifiers are much more common than relative or participle clauses, they received the least attention in grammars. This is absolutely true in English textbooks used in Japanese schools. However, there seems to be another reason why Japanese learners underuse the first meaning of *of* (belonging or relating to somebody) and overuse *of* in the second meaning (belonging to or being part of something). It can be explained by comparing the concepts of these two *of*s. Tyler and Evans (2003) explain that even in the basic meaning of *of*, there are two different relationships between the trajector and the landmark. The first relationship explains the meaning No. 1, where the trajector and the landmark are somewhat loosely connected (ex. *the role of the teacher*). The second relationship explains the meaning No. 2, where the trajector is actually a part of the landmark (ex. *the lid of the box*). The fact that the meaning No.2

contains a cognitively less complex concept than No.1 leads Japanese learners to use *of* in the meaning No. 1 more frequently.

Since the author does not speak European languages, how the degree of cognitive complexity works in the use of *of* is still disputable.. In addition, it is necessary to refine the designation of the meanings of *of* in each 'N *of* N' structure. Nevertheless, it is clear from the present study that the background language, educational contexts, and the cognitive complexity of the concept of the target word somewhat affects how well the preposition *of* is learned.

References

Biber, Douglas, Susan Conrad & Reppen, Randi. (1994) Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15(2): 169-89.

Granger, Sylviane & Rayson, Paul. (1998) Automatic profiling of learner texts. In: Granger, Sylviane (ed.) *Learner English on Computer* (pp.119-31). London & New York: Addison Wesley Longman.

Tyler, Andrea & Evans, Vyvyan. (2003) *The Semantics of English Prepositions: Spatial, Scenes, Embodied Meaning and Cognition.* Cambridge: Cambridge University Press.

Uchida, Tomio. (2007) Zenchi-shi no hattatsu (Development of prepositions). In Tono, Yukio (ed.) *Nihonjin Chu-kou-sei Ichiman-nin no Eigo Ko-pasu (English Language Corpus Compiled from 10,000 Junior/Senior High School Students)* (pp.109-16). Tokyo: Shogakukan.

Missing prepositions — a report from an explorative corpus-based study

Kinn, Torodd;
University of Bergen;
Torodd.kinn@lle.uib.no

Tenfjord, Kari
University of Bergen
kari.tenfjord@lle.uib.no

This presentation reports a small explorative study of the use of prepositions in Norwegian learner languages. The study is based on the ASK corpus developed at the University of Bergen and accessible through the newly developed web-based platform Corpuscle (Meurer 2012).

We will present the ASK corpus as well as some results from the preposition study. We will give a very brief presentation of what kind of data and information the corpus contains and we will illustrate some search possibilities, using the preposition study as an example. We will identify patterns emerging from the search results and our categorization of the structural contexts of the prepositions. We will show that across the structural categories there are conceptually based semantic patterns. Furthermore, these patterns are closely connected to some highly frequent collocations.

Prepositions are frequent, grammatically multifunctional and polysemous, straddling the divide between grammar and lexicon. The use of prepositions is perceived as difficult when learning and teaching L2 Norwegian. The learners may choose a wrong preposition, use a redundant preposition, or the preposition may be missing – all this of course in contrast to conventional Norwegian language use.

What kind of study a corpus is suitable for depends on the kinds of textual data and analytic information added to it. ASK contains 1700 written texts collected from two different language tests for adults in Norwegian L2. The tests are intended to measure whether the learner languages are at or above the B1 and B2 levels of proficiency. 1100 of the texts have been reevaluated according to the CEFR description (Carlsen 2012). The texts are written by learners with German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali as their L1.The corpus also contains 200 Norwegian L1 texts written by adults with the same kind of assignments. It is easy to choose sub-corpora by making a selection, for instance based on test level, L1 group, age group, residence time etc.

The annotations in ASK are suitable for a study of learners' use of prepositions. The corpus texts have been manually tagged for ”error”. For each error tag, a ”corr” (correction) tag has been added. The error annotation is done in accordance with the principles described in Tenfjord, Hagen & Johansen 2006). Furthermore, the corpus has been automatically tagged for parts of speech and certain other syntactic and morphological information.

In our preposition study, we first searched for prepositions, resulting in a frequency list from which we chose the three most frequent ones: i ('in'), på ('on') and til ('to'). We then searched for these prepositions when tagged for the error codes W (wrong word), R (redundant) and M (missing). In addition, we searched for prepositions corrected to i, på or til. På was the most frequently missing preposition, and it was also the most frequent preposition used as a correction when a wrong preposition was used. Therefore, på seemed to us to be a suitable object for an explorative study.

When the data had been collected, we categorized the structural contexts in which the preposition was missing: What kind of syntactic construction is the (target language) prepositional phrase part of, and what kind of complement does the preposition have? Thus, we started out with a structurally based error analysis.

Some frequent constructions where på is missing are (more or less) idiomatic expressions (like håpe (på) noe 'hope for sth.'), constructions where the preposition is used to subordinate a clause under an adjective (like sikker (på) at … 'sure (about) that ...'), and relative clauses with a "stranded" preposition (like samme skole som jeg går (på) 'the same school that I go to').

Several of the constructions are both typologically rare and structurally complex – and it is reasonable to assume that they are difficult to acquire. It is also of interest that across the structurally identified constructions many of the idiomatic expressions stand for a cognitive "direction" (attention directed towards something), offering a possibility for a semantically based approach to the teaching of these complex structures.

The results of this small explorative corpus-based study provide a basis for generating new hypotheses about the connection between frequency, collocations and semantics in the context of language learning.

References

Carlsen, Cecilie. (2012) Proficiency level - a fuzzy variable in computer learner Corpora. In: *Applied Linguistics;Volum 33 (2).* 161-183

Meurer, Paul. (2012) Corpuscle – a new corpus management platform for annotated corpora. In: Andersen, Gisle (ed.): *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, *Studies in Corpus Linguistics 49*, John Benjamins.

Tenfjord, Kari; Hagen, Jon Erik & Johansen, Hilde. (2006) "The hows and whys of coding categories in a learner corpus (or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy)". In: *Rivista di Psicolinguistica Applicata (RiPLA) VI(3):* 93-108.

Lexicogrammatical profile of Estonian as a second language on the B1 level: Some results of the corpus-driven study

Mare Kitsnik
Tallinn University
marekitsnik@gmail.com

In Estonia as in the whole European Union the CEFR is used as the basis for determining second language proficiency. In the CEFR the language levels are described in the "can do" style and these detailed descriptions can be applied to all languages. However, curriculum writers, the authors of study materials, the assessors of language proficiency, the language teachers and learners also need more language specific information about vocabulary, phrases and constructions which are important to communicate successfully on particular levels of the language proficiency. In Estonia the first step in specification of the language levels has been taken: the language level descriptions are compiled and published (Ilves 2008, 2010; Hausenberg 2008; Kerge 2008). There the preliminary description of vocabulary and grammatical constructions essential for expressing language functions and general concepts in different language levels is presented, but it still based more on intuition and experience than on the scientific researches.

In my doctoral thesis I will examine the lexicogrammatical profile of B1 and B2 writing performances in Estonian as a second language, both quantitatively and qualitatively. The study is carried out on a sub-corpus of Estonian Interlanguage Corpus which contains written B1 and B2 level examination papers of Estonian as a second language.

The lexicogrammatical construction is meant as co-occurence of words with syntactical relationship between them (*ta tahab minna kontserdile – he wants to go to the concert, pilet on kallis – ticket is expensive, kaks raamatut – two books*).

The main research questions are:

1. Which lexicogrammatical constructions are typical in the writing performances at B1 level.
2. Which lexicogrammatical constructions are typical in the writing performances at B2 level.
3. What are the main similarities and the main differences between the lexicogrammatical constructions at B1 and B2 level.

The main hypotheses are:

1. At B1 level the frequency of the less common lexicogrammatical constructions has been increased.
2. At B1 level the accuracy of using the more common lexicogrammatical constructions has been increased.
3. At B2 level the accuracy of using the less common lexicogrammatical constructions has been increased.

The theoretical framework of my research is based on three dimensions of language proficiency: complexity, accuracy and fluency (Housen, Kuiken 2009). In addition the DEMfad model (Franceschina, Alanen, Huhta & Martin, 2006) concepts are used: domain (areas of developing language skill, construction or a set of constructions, a set of vocabulary etc), emergence (the first occurence of some indication of the presence of a domain), mastery (approximately target-like use of the domain from the standpoint of frequency and distribution), frequency (related to the concept of fluency and calculated per 1000 words of running text), accuracy (compared to native speakers at 80% level), distribution (a cover term for several types of phenomena which can be tracked in learner development).

In the process of investigation the program WordSmith Tools 5,0 (Scott, Tribble 2006) and corpus-driven approach will be used. By using the program, three lists will be compiled for both levels (B1 and B2): the vocabulary frequency list, concordances and key words. Then the words will be grouped (objects, actions and processes, qualities and states, amounts and extents, time and space, relations and connections) and descriptions of frequent lexicogrammatical constructions will be compiled (including usage contexts, collocations and typical incorrectnesses). Then the results of B1 and B2 levels will be compared. Finally the lists of lexicogrammatical constructions including descriptions will be compiled.

The results can be used in development of teaching Estonian as a second language. They can be useful for curriculum writers, the authors of study materials, the assessors of language proficiency, the language teachers and learners.

In my paper the preliminary results of the corpus-driven study of written performance on the B1 level will be presented and discussed.

Keywords: second language acquisition, corpus-driven study, lexicogrammatical constructions, writing performance, B1 and B2 levels, Estonian

References

Franceschina, F.; Alanen, R.; Huhta, A.; Martin, M. (2006). *A progress report on the Cefling project.*

Hausenberg, Anu-Reet jt (2008). *Iseseisev keelekasutaja. B1- ja B2-taseme eesti keele oskus*. Tallinn, REKK, Atlex.

Housen, Alex; Kuiken, Folkert (2009). *Complexity, accuracy and fluency in Second Language Acquisition.* – Applied Linguistics (2009).

Ilves, Marju (2008). *Algaja keelekasutaja. A2-taseme eesti keele oskus*. Tallinn, EKSA.

Kerge, Krista (2008). *Vilunud keelekasutaja. C1-taseme eesti keele oskus*. Tallinn, Eesti Keele Sihtasutus.

Scott, Mike; Tribble, Christopher (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Philadelphia, John Benjamins.

The Sketch Engine interface for a learner corpus annotated with errors and corrections

Kosem, Iztok*; Kovar, Vojtech**; Baisa, Vit**; Kilgarriff, Adam**
*Trojina, Institute for Applied Slovene Studies
**Lexical Computing Ltd.
iztok.kosem@trojina.si; xkovar3@gmail.com; vit.baisa@gmail.com; adam@lexmasterclass.com

One of the common features of learner corpora is annotated errors of language learners. Identification and quantification of learner errors enables the design of various resources to address those errors. In addition, language teachers and authors of learning materials can use that information to prioritize language problems of learners, putting more focus on frequent and wide-spread problems. It is also useful if learner corpora also contain corrections of identified errors as this facilitates classification and helps grouping errors with the same correction.

Part of an effective analysis and usage of learner corpora annotated with errors and corrections is a good corpus tool. It is still more common that a tool is tied to a particular learner corpus, i.e. that it is developed for that corpus alone, as is the case with the International Corpus of Learner English (Granger et al. 2009). Such practice poses problems for researchers and teachers who want to use different learner corpora, as they have to learn how to use different tools. Moreover, it needs to be considered that the users of learner corpora possess different levels of proficiency in corpus use – for example, language teachers who could benefit from having access authentic examples of errors relevant for their students, are usually the ones who require most convincing to start using corpora and tend to be put off by, at least from their perspective, complex corpus tools (Kosem 2008). Developers of tools for corpora with annotated errors and corrections face additional challenges with visualising both types of information in a user-friendly manner. Namely, concordances alone can be overwhelming for some users, never mind if they include corrections, error codes and similar additional information.

This paper presents an interface of the Šolar corpus of young learner writing (Kosem, Rozman & Stritar 2011). The Šolar corpus was built within the Communication in Slovene project and contains texts of Slovene elementary and secondary school students, comprising nearly 1 million words. Although this is a corpus of L1 writing, it has been modelled after learner corpora and contains annotated student errors, as identified by their teachers, and teacher corrections. It was always envisaged that the corpus would be made available to teachers and publishers, but one of the main problems was that these target groups have little knowledge in the use of corpora. Thus, our aim was to design an interface that would be user-friendly and easy to use.

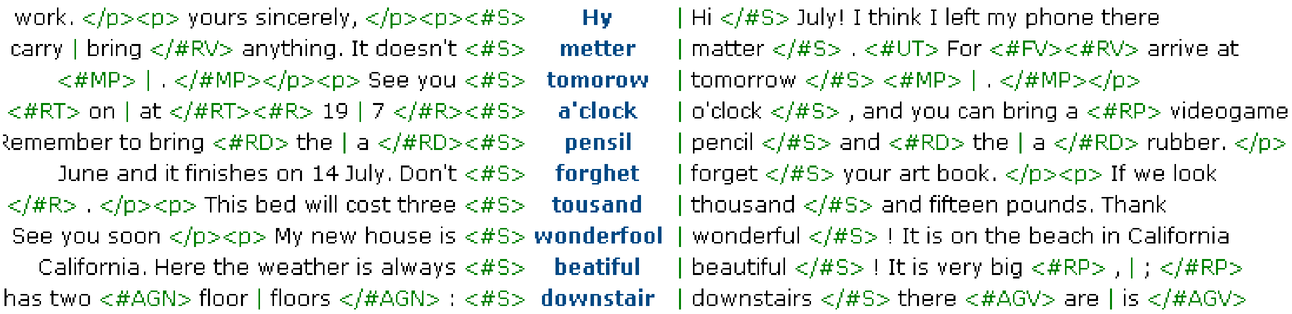
We have decided to use the Sketch Engine (Kilgarriff et al. 2004) as a point of departure for two reasons: it has specialised functionality for an error-annotated learner corpus (developed originally for the Cambridge Learner Corpus) and it is the most widely used corpus tool so it is likely that many users will be familiar with it. We approached the development of the Šolar version of the interface with the following aims:

a) Display of errors and corrections (but not annotation tags) had to be the default option, with both types of information clearly distinguishable from each other, and from non-errors/corrections and the surrounding text. This was achieved by displaying errors in red, corrections in green, and all other text in black. In addition, we replaced the red colour for the node word (default setting) with black (in bold); this meant that the users can easily distinguish between errors and non-errors if conducting a simple search (but not an error search).

b) The users had be able to reveal in the concordance output the annotation tags, which had to be named in an easily understandable manner, i.e. instead of using abbreviations we used one-word names for error categories (e.g. *Besedisce* – meaning Vocabulary). However, we also included this information in the metadata so it could be displayed in a side column, thus reducing the need for viewing concordances with tags to cases where the user wants to see error categories for errors in the context.

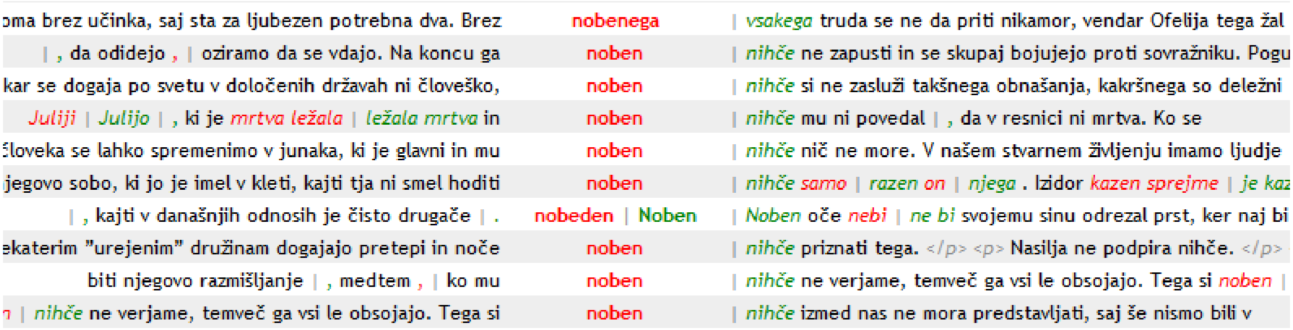
c) The tool had to be localised into Slovene with corpus terminology used only when necessary, and with tips and help provided for each function of the tool. We developed a separate website that included explanations of the main functions of the tool, examples of use, and related screenshots of the interface. The Šolar tool and help pages were linked (question mark icons in the interface).

The final version of the interface (available at http://www.korpus-solar.net) is best demonstrated by showing the difference between the concordance in the original interface, used by Cambridge University Press (Figure 1) and the Šolar corpus interface (Figure 2). In the Šolar corpus interface, errors and corrections clearly distinguished from each other, and from the surrounding context.



```
 work. </p><p> yours sincerely, </p><p><#S>      Hy      | Hi </#S> July! I think I left my phone there
carry | bring </#RV> anything. It doesn't <#S>  metter    | matter </#S> . <#UT> For <#FV><#RV> arrive at
      <#MP> | . </#MP></p><p> See you <#S>    tomorow    | tomorrow </#S> <#MP> | . </#MP></p>
<#RT> on | at </#RT><#R> 19 | 7 </#R><#S>     a'clock    | o'clock </#S> , and you can bring a <#RP> videogame
Remember to bring <#RD> the | a </#RD><#S>     pensil    | pencil </#S> and <#RD> the | a </#RD> rubber. </p>
    June and it finishes on 14 July. Don't <#S>  forghet  | forget </#S> your art book. </p><p> If we look
</#R> . </p><p> This bed will cost three <#S>  tousand   | thousand </#S> and fifteen pounds. Thank
 See you soon </p><p> My new house is <#S> wonderfool | wonderful </#S> ! It is on the beach in California
  California. Here the weather is always <#S>  beatiful  | beautiful </#S> ! It is very big <#RP> , | ; </#RP>
has two <#AGN> floor | floors </#AGN> : <#S>  downstair  | downstairs </#S> there <#AGV> are | is </#AGV>
```

Figure 1. Default view in the Cambridge Learner Corpus interface (erros, corrections and tags displayed).



```
oma brez učinka, saj sta za ljubezen potrebna dva. Brez      nobenega       | vsakega truda se ne da priti nikamor, vendar Ofelija tega žal
      | , da odidejo , | oziramo da se vdajo. Na koncu ga      noben         | nihče ne zapusti in se skupaj bojujejo proti sovražniku. Pogu
kar se dogaja po svetu v določenih državah ni človeško,       noben         | nihče si ne zasluži takšnega obnašanja, kakršnega so deležni
  Juliji | Julijo | , ki je mrtva ležala | ležala mrtva in     noben         | nihče mu ni povedal | , da v resnici ni mrtva. Ko se
človeka se lahko spremenimo v junaka, ki je glavni in mu      noben         | nihče nič ne more. V našem stvarnem življenju imamo ljudje
jegovo sobo, ki jo je imel v kleti, kajti tja ni smel hoditi  noben         | nihče samo | razen on | njega . Izidor kazen sprejme | je kaz
        | , kajti v današnjih odnosih je čisto drugače | .  nobeden | Noben  | Noben oče nebi | ne bi svojemu sinu odrezal prst, ker naj bi
ekaterim "urejenim" družinam dogajajo pretepi in noče        noben         | nihče priznati tega. </p> <p> Nasilja ne podpira nihče. </p>
           biti njegovo razmišljanje | , medtem , | ko mu     noben         | nihče ne verjame, temveč ga vsi le obsojajo. Tega si noben |
n | nihče ne verjame, temveč ga vsi le obsojajo. Tega si      noben         | nihče izmed nas ne mora predstavljati, saj še nismo bili v
```

Figure 2. Default view in the Šolar Corpus interface (erros and corrections displayed).

The Šolar corpus interface was designed for users who are less proficient users of corpora, and we have already received positive feedback from target users, however other users such as researchers have also expressed interest to use such a tool for their own purposes, such as error analysis. As a result, we are now testing the possibility of using the concordance annotation functionality, used for example in the Pattern Dictionary project (Hanks 2009), for annotating errors in the corpus, or annotating already identified errors as part of further classification.

References

Granger, S., E. Dagneaux, F. Meunier and M. Paquot. (2009) *International Corpus of Learner English V2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Hanks, P. (2009) 'The Linguistic Double Helix: Norms and Exploitations'. In: After Half a Century of Slavonic Natural Language Processing (Festschrift for Karel Pala) (pp. 63-80). Brno, Czech Republic: Masaryk University.

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell. (2004) 'The Sketch Engine.' In G. C. Williams and S. Vessier (eds.) *Proceedings of the 11th Euralex International Congress* (pp. 105–116). Lorient, France: Universite de Bretagne-Sud.

Kosem, I. (2008) 'User-friendly Corpus Tools for Language Teaching and Learning.' In A. Frankenberg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference* (pp. 183–192). Lisbon: ISLA.

Kosem, I., T. Rozman and M. Stritar. (2011) 'How Do Slovenian Primary and Secondary School Students Write and What Their Teachers Correct: a Corpus of Student Writing.' In *Proceedings of the Corpus Linguistics Conference 2011*. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-198.pdf.

The Šolar corpus. http://www.korpus-solar.net. Accessed on 26th May 2013.

Word Formation Variation as Features for Native Language Identification

Julia Krivanek, Detmar Meurers
Universität Tübingen
{jkrvnk,dm}@sfs.uni-tuebingen.de

The task of native language identification can be useful for theoretical studies of language transfer (Jarvis 2012) and it can inform applications, e.g., by informing learner models for intelligent language tutoring systems to support different feedback depending on the L1 (Amaral & Meurers 2008). Current L1-classification approaches (e.g., Brooke & Hirst 2012; Bykh & Meurers 2012; Jarvis et al. 2012) achieve high accuracy with surface-based features, such as word and part-of-speech n-grams. However, surface-based approaches make use of large feature sets, which are hard to interpret qualitatively in terms of linguistic insight. In addition, surface features are directly dependent on the genre and topic of the texts being classified, so that results degrade significantly for out-of-domain classification (Brooke & Hirst 2011). Other approaches (Wong & Dras 2009; Bestgen, Granger & Thewissen 2012) make use of error patterns, which capture one conceptually interpretable characteristic of learner language, but typically require manual error annotation.

In this paper, we propose to shift the focus to a new class of features for L1-classification: linguistic variation. In many situations, language offers a range of options for formulating a given message. Indeed, in variationist sociolinguistics, the choices speakers make have successfully been used to identify relevant speaker properties (cf. Tagliamonte 2011). Adapting this perspective, we propose to make use of variation features for native language identification. We make use of the variationist method observing where speakers make choices in the language system, but different from variationist sociolinguistic research we then investigate the impact of the L1 (rather than the social properties focused on in sociolinguistics). Making this general idea concrete, we describe an experiment we carried out on German learner texts using word formation variation as features for L1-identification.

We use the term word formation to refer to the range of processes through which new words are formed. Typically a given language offers several options. Which options get used when and how the options are realized differs across languages. New words can be formed with the help of derivational morphemes or without them, the process can change a word's category or not, and so on. Accordingly, we can define variables such as the ones in Figure 1 and use their variants as features for L1-classification.

Variables	**Variants**	**Examples**
Morpheme alternation	no affix	*Frau<NN>* + *Welt<NN>* → *Frauenwelt<NN>*
	suffix	*Feminist<NN>* + ***in<SUFF>*** → *Femimistin<NN>*
	prefix	***un<PREF>*** + *gerecht<ADJ>* → *ungerecht<ADJ>*
	verb particle	***auf<VPART>*** + *geben<V>* → *aufgeben<V>*
Derived category alternation	noun	*anerkennen<V>* + *ung<SUFF>* → ***Anerkennung<NN>***
	verb	*auf<VPART>* + *geben<V>* → ***aufgeben<V>***
	adjective	*entsprechen<V>* → ***entsprechend<ADJ>***
	adverb	*möglich<ADJ>* + *weise<SUFF>* → ***möglicherweise<ADV>***
Source category alternation	noun	***Feminist<NN>*** + *in<SUFF>* → *Femimistin<NN>*
	verb	***anerkennen<V>*** + *ung<SUFF>* → *Anerkennung<NN>*
	adjective	***möglich<ADJ>*** + *weise<SUFF>* → *möglicherweise<ADV>*

Figure 1: Some word formation variables

For example, the *morpheme alternation* allows us to distinguish word formation without affix from that using suffixes or using prefixes The *derived category alternation* supports distinguishing derived from basic variants. The *source category alternation* supports identifying which source categories undergo a word formation process.

As learner corpus data we used 185 essays from the Falko learner corpus of German (Reznicek et al. 2012), written by learners with five native languages (English, Polish, Russian and Danish, and a native German control group), with an average length of 470 words. The data was annotated using the RFTagger (Schmid & Laws 2008) providing part-of-speech and morphological information.

As features, we took four categories (noun, verb, adjective and adverb) and compiled out all possible variations of the word formation variables described above. These variations were then counted for each text and normalized by the derived category. After removing features which did not occur in the data set, we obtained 29 features.

For classification, we used the WEKA SMO classifier (Witten & Frank 2005) and report the results of leave-one-out evaluation. Using only the 29 word formation features, we obtained a classification accuracy of 55.1%, which is encouraging given the random baseline of 20% for this balanced five class problem. Just as in the current NLI approaches for English, the accuracy can be increased by introducing a combination of different feature types, as we demonstrate in Bykh et al. (2013); we here instead provide an analysis of the word formation variation features as the focus of this paper.

An analysis of the confusion matrix shows that the German control group data is most clearly singled out, whereas many confusions arise within the Slavic group (8 Polish texts are identified as Russian, 12 Russian ones as Polish). We therefore are exploring the use of cascading classification to first distinguish language families (e.g., Slavic vs. others) followed by a second classification trained only on the subdistinctions within a language family (e.g., Polish vs. Russian). We expect that the features which are most effective at these different stages will differ clearly and meaningfully, in line with the findings of Vajjala & Loo (2013), who used a cascading classifier in a proficiency classification task.

One can also anylze the results of our approach in terms of *overuse/underuse* (Lüdeling et al. 2011). In order to detect distinctive features, one compares the frequencies of a variant of a given variable across the L1 groups. Comparing the L1-German control group with the other L1 groups, we, for example, found that the phrasal verb feature "verb particle + verb" (e.g., *auf<VPART>geben<V>*) was underused by all learners, with native Danish learners being the closest to native German usage. Native speakers of Slavic languages, lacking phrasal verbs, and English, where particles follow different distributional patterns than in German, showed the strongest underuse.

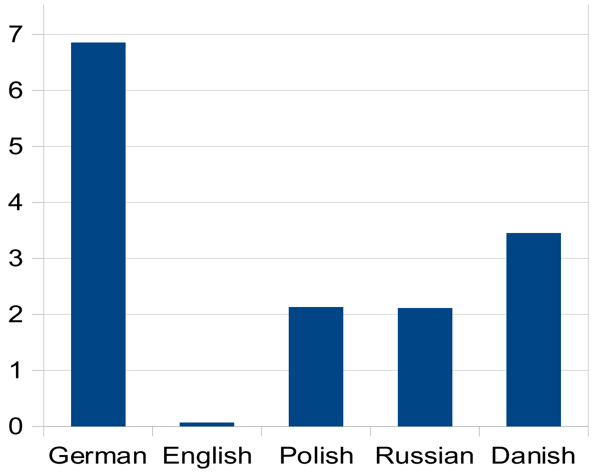


Figure 2: Relative frequency of phrasal verbs in German texts across different L1s

In conclusion, an analysis of variation in word formation provides an effective and insightful perspective for L1-classification. As such it further populates the landscape of data-driven and theory-driven approaches (Meurers et al. 2013) in a way yielding qualitatively interpretable features. At the same time, it can also be integrated into ensemble classifiers combining different sources of information for L1-classification (Bykh et al. 2013) to further improve the quantitative state-of-the-art in terms of classification accuracy.

References

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL http://purl.org/dm/papers/amaral-meurers-call08.html.

Bestgen, Y., S. Granger & J. Thewissen (2012). Error Patterns and Automatic L1 Identification. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 127–153.

Brooke, J. & G. Hirst (2011). Native Language Detection with 'Cheap' Learner Corpora. In *Learner Corpus Research 2011 (LCR 2011)*. Louvain-la-Neuve.

Brooke, J. & G. Hirst (2012). Robust, Lexicalized Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 391–408.

Bykh, S. & D. Meurers (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 425–440. URL http://purl.org/dm/papers/bykh-meurers-12.html.

Bykh, S., S. Vajjala, J. Krivanek & D. Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA. URL http://purl.org/dm/papers/Bykh.Vajjala.ea-13.html.

Jarvis, S. (2012). The Detection-Based Approach: An Overview. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 1–33.

Jarvis, S., G. Castañeda-Jiménez & R. Nielsen (2012). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 34–70.

Lüdeling, A., H. Hirschmann & A. Zeldes (2011). Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German. In Y. Kawaguchi, M. Minegishi & W. Viereck (eds.), *Corpus Analysis and Diachronic Linguistics*, Amsterdam: John Benjamins.

Meurers, D., J. Krivanek & S. Bykh (2013). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Proceedings of 4th International Conference on Corpus Linguistics (CILC 2012)*. To appear.

Reznicek, M., A. Lüdeling, C. Krummes & F. Schwantuschke (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Ver. 2.0.* URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

Schmid, H. & F. Laws (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*. Stroudsburg, PA, vol. 1, pp. 777–784.

Tagliamonte, S. A. (2011). *Variationist Sociolinguistics: Change, Observation, Interpretation*. John Wiley & Sons.

Vajjala, S. & K. Loo (2013). Role of Morpho-syntactic features in Estonian Proficiency Classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*. URL http://aclweb.org/anthology/W13-1708.pdf.

Witten, I. H. & E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam; Boston, MA: Morgan Kaufmann, 2nd ed.

Wong, S.-M. J. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. pp. 53–61.

Investigating acquisition of prosodic focus marking and native perception of learners' English intonation

Kunevich, Maria[1]; Hudson, Toby[1]; Post, Brechtje[1]; Alexopoulou, Dora[1]
[1]University of Cambridge, Department of Theoretical and Applied Linguistics
msk41@cam.ac.uk; toh22@cam.ac.uk; bmbp2@cam.ac.uk; ta259@cam.ac.uk

The work presented here is part of a project investigating the acquisition of information structure categories in L2 English by learners with different first languages. The current study investigates the realisation of prosodic prominence. In English prosodic prominence is one of the main means to convey the information structure of an utterance by highlighting constituents of high informational importance. In our study we are interested, firstly, in whether L2 English learners make appropriate use of nuclear accent placement to convey linguistic focus in speech production, and secondly, how native English speakers perceive the attempted accent placement by the learners. The background to this investigation is the observation that L2 learners often experience major difficulties in the area of prosody (cf., Mennen, 2007); and the hypothesis is that learners do not only make errors in placing nuclear stress but also use different cues to signal it.

Our approach is to construct a purpose-built data set in which the prosodic focus environment is carefully controlled. Our corpus currently consists of semi-spontaneous utterances produced by 36 Russian learners of English at three levels of proficiency (corresponding to A2, B1, B2 levels of the Common European Framework of Reference) as well as 12 native speakers of Standard Southern British English. The utterances are declarative sentences of two types, transitive and intransitive. The transitive sentences are of the type subject - verb - object, with broad focus (i) and narrow focus on the subject (ii) or the object (iii). The intransitive sentences consist of subject and verb, with broad focus (iv) and narrow focus on the subject (v) or the verb (vi). These patterns of prosodic minimal triplets were produced for six lexicalisations, elicited as responses to questions, e.g.:

(1) *What happened while I was out? - Diana danced the flamenco.*

(2) *Who danced the flamenco? – Diana danced the flamenco.*

(3) *What did Diana dance? - Diana danced the flamenco.*

The question–answer pairs were randomised into three sets and presented to the participants together with filler question-answer pairs so that each target pair was separated by a filler pair. This allowed us to compile a corpus with data that are directly comparable, allowing us to analyse the intonational contours of utterances produced by different speakers in identical contexts.

We are interested in which pitch, durational, intensity, spectral (and segmental) features are salient for a) unambiguous communication of prominence, and b) sounding 'authentic', i.e. sufficiently similar to native productions, or considered 'good enough' by the native listener. To address this and to provide a listener-oriented perspective, a subset of the utterances was used as stimuli for the native perception online experiment. The task for the listeners was to identify the prominent word in an utterance, or if none of the words was more prominent or they were all equally stressed to choose a "none of the words/equal prominences" option.

The utterances from the subset are also being analysed acoustically to determine the sources of any errors. The acoustic analysis has two components, phonological and phonetic. First, the location of the accented word(s) is determined by the experimenters (words are labelled as accented or unaccented), using auditory perception and examination of the pitch contour. In the phonetic part of the analysis the following acoustic measurements are extracted: the height and alignment of the pitch peak of the accented word(s) and the duration of the vowels in stressed syllables in broad and narrow focus conditions.

Preliminary results show that learners differ in their realisation of narrow and broad focus, and though they often make an appropriate choice for the *placement* of pitch accent in different focus conditions, its phonetic *realisation* is inappropriate. We also found that the degree of deaccenting correlates with the proficiency level: the higher the proficiency, the greater is the difference between the F0 peak values of first and final word in narrow focus condition. The next step is to compare the results from the perception experiment and to examine which acoustic features have the greatest impact on the perception of prominence.

Finally, all sound files will be passed through automatic prosodic prominence detection algorithms (designed for use on native speech) so we can see to what extent this output matches human judgements and how acoustic parameters like changes in fundamental frequency and duration of syllable nuclei duration contribute to the percept prominence in learner speech.

In sum, integrating prosodic labeling with the contextual (discourse context) and listeners' rating information in the corpus annotation, and comparing both with automatic prosodic prominence detection will provide more insight into the development and internal structure of learner prosody. Our next step will be to incorporate learners with other first language backgrounds, which will be achievable with the EF-Cambridge Open Language Database currently being built.

References:

Bley-Vroman, Robert (1983) The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33, 1–17.

Buhmann, Jeska, Caspers, Johanneke, van Heuven, Vincent, Hoekstra, Helen, Martens, Jean-Pierre & Swerts, Mark (2002) Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus. *Proc. LREC 2002,* 779-785.

Gut, Ulrike (2009) *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German.* Frankfurt: Peter Lang

Mo, Yoonsook, Cole, Jennifer, Lee, Eun-Kyung (2008) Naïve listeners' prominence and boundary perception. *Proc. Speech Prosody 2008,* 735-738.

Mennen, Ineke (2007) Phonological and phonetic influences in non-native intonation. In: Trouvain, Jürgen and Gut, Ulrike (eds.) *Non-native prosody: Phonetic description and teaching practice.* Berlin: Mouton de Gruyter, 53–76.

How to Annotate Morphologically Rich Language? Problems and Solutions.

Lehto, Liisa-Maria; Brunni, Sisko; Jantunen, Jarmo Harri
Affiliation: University of Oulu
liisa-maria.lehto@oulu.fi; sisko.brunni@oulu.fi; jarmo.jantunen@oulu.fi

The International Corpus of Learner Finnish (ICLFI) is a learner language corpus which consists of texts written by learners of Finnish. The data consists of circa 1.1 million tokens, texts were collected from speakers of 22 mother tongues. (Jantunen 2011.) So far (February 2013) about 23% of the ICLFI is morphologically annotated and now annotation will be extended to error annotation. However, there are also problems that need to be solved in the annotation processes. This poster presents solutions to annotating morphologically rich language and sheds light on principles to be decided and to the annotation process itself. Both grammatical and error annotation are discussed from the morphological viewpoint.

Morphology and morphosyntax bring their own challenges to the annotation processes. In general, fully automatic annotation, both grammatical and error annotation, is not possible in learner corpus research (see e.g. Díaz-Negrillo & Fernández-Domínguez 2006; Rooy & Schäfer 2003). That is because when combining lexical and grammatical morphemes, language learners make up forms, which fully automatic analyzer cannot process (see e.g. Dagneaux et al. 1998). Therefore the ICLFI corpus follows semi-automatic tagging procedure: in grammar annotation, the texts are first analyzed using a word processor in order to find and manually correct spelling and inflection mistakes. Then the texts are analyzed using Connexors functional dependency grammar parser (Fi-fdg, Järvinen et al. 2004), which produce morpho-syntactic information and lemmatize the data. Finally spelling mistakes are manually restored to the annotated text files and the automatic analysis is manually checked. (Jantunen 2011.) Checking annotation manually enables the annotator to add up all relevant alternative tags in the annotated file. Thus it is possible to disambiguate the tags and morpho-syntactic forms when there is more than one possible interpretation of the word form produced by a language learner.

Learner corpora are especially useful when all the errors in the corpus have been annotated with the help of standardized system of error tags. An error-tagged learner corpus gives researchers access to error statistics and automated error analysis. (Granger 2003.) An error annotation system should allow consistent and systematic tagging (Fitzpatrick & Seegmiller 2004; Granger 2003) and for the sake of comparing the data and the results of the annotations of different corpora, the annotation systems should be more or less symmetric. However, Díaz-Negrillo and Fernández-Domínguez (2006) point out that researchers often design their own error tagging systems and use different models in tagging. Although this is clearly a disadvantage what comes to the comparability of the annotated data, there is sometimes a need to find corpus-specific solutions. This is often the case when corpora represent languages from typologically different language groups. Furthermore, annotation systems ought to be informative, detailed, reusable and general (see Granger 2003; Díaz-Negrillo & Fernández-Domínguez 2006).

In the error annotation system of ICLFI, tagsets allow retrievals in different levels of errors. Finnish, which is a synthetic language with a rich morphophonological system, allows a large amount of information coded in one morpheme or a word form. This is why, in the error taxonomy of ICLFI, morphological and morpho-syntactical error categories are the most detailed and so the error tagsets are the most complex, which differ from the error tagging systems such as the one in ICLE. When a word is inflected in Finnish, both stem and suffixes may vary greatly. Phonetic, qualitative and quantitative, variation in vowels and consonants also increases complexity. Furthermore, there is also a rich system on both case and number congruence. Our poster represents some concrete solutions that have been made to solve annotation problems that exist in grammatical and error annotation of learner Finnish.

References

Dagneaux, Estelle, Denness, Sharon & Granger, Sylviane (1998) Computer-aided error analysis. *System* 26: 163–187.

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

Fitzpatrick, Eileen & Seegmiller, M. Steve (2004) The Montclair Electronic Language Database project. In: Connor, Ulla & Upton, Thomas (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 223–237). Amsterdam: Rodopi.

Granger, Sylviane (2003) Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal 20*: 465–480.

Jantunen, Jarmo Harri (2011) Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. [International Corpus of Learner Finnish: typology, variants and annotation.] *Lähivõrdlusi. Lähivertailuja 21*: 86–105.

Järvinen, Timo; Laari, Mikko; Lahtinen, Timo; Paajanen, Sirkku; Paljakka, Pirkko; Soininen, Mirkka & Tapanainen, Pasi (2004) Robust Language Analysis Components for Practical Applications. In: Gambäck, Björn & Jokinen, Kristiina. (eds.) *Coling 2004*, *Proceedings of the Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces* (pp. 53–56).

van Rooy, Bertus, & Schäfer, Lande (2003) An evaluation of three POS taggers for the tagging of the Tswana learner English corpus. In: Archer, Dawn, Rayson, Paul, Wilson, Andrew & McEnery, Tony. (eds.) *Proceedings of the Corpus Linguistics 2003 conference, 28-31 March 2003* (pp. 835–844). Lancaster: Lancaster University.

The role of conventionalized language in the acquisition and use of articles by EFL learners – a crosslinguistic perspective

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

The article system is one of the most pervasive features of the English grammar. Yet, in spite of their prominence, articles seem particularly hard for EFL learners to acquire. Even advanced students seem to struggle with the correct use of articles and frequently make errors. Students whose mother tongues do not have article systems find it especially difficult to acquire this feature of the English grammar. However, even students whose L1s have articles find it difficult to use English articles accurately.

Many studies have examined the acquisition of this grammatical feature, and several theories and hypotheses have been proposed to account for this process (e.g. Hakuta, 1976; Tarone, 1985, Parrish 1987; Thomas 1989; Butler 2002; Jarvis, 2002; Ekiert 2004; Li & Yang 2010; Crompton 2011). However, researchers are still far from agreeing on a single model that explains the acquisition of articles.

The difficulty in acquiring articles has been ascribed to problems in mastering a complex system of grammatical, semantic, and pragmatic relations. Earlier studies usually concentrated on the analysis of obligatory contexts for article use and the analysis of learners' accuracy levels in these contexts. Such an approach implies the belief that the use of articles is rule-based, and that learners gain higher and higher levels of mastery of these rules throughout the process of learning.

Yet, it has long been asserted that language processing is not solely rule-based. Sinclair (1990) proposes two complementary principles explaining language use. According to him, in addition to being built from scratch (based on generating grammatical structures), language is processed through the idiom principle, i.e., by selecting ready-made multi-word chunks of language from the phrasicon.

Articles are frequently part of such multi-word expressions. Thus, it can be hypothesized that at least some obligatory contexts for the use of articles are acquired within larger lexical phrases, and are processed as such. So far, few studies considered the acquisition of articles from this perspective, and if they did do so, they tended to treat the idiomatic uses of articles only marginally (Ekiert 2004; Li and Yang 2010).

The study reported in this paper was meant to fill this gap. It was exploratory in nature and it aimed to draw attention to the role of the conventional uses of language in the selection of articles by learners of English with the mother tongues featuring and not featuring article systems. More specifically, it sought to establish which uses of the articles *the* and *a/an* in student writing could be accounted for by the learners' use of conventionalized multi-word phrases rather than by the application of rules relating to grammatical, semantic, and pragmatic relations and whether there were any differences in this respect between learners with different L1s.

The data used in this study were drawn from several corpora. Expository and argumentative essays written by L1 Polish, Spanish and German learners at four different proficiency levels, ranging from beginners to advanced students, were drawn from the International Corpus of Crosslinguistic Interlanguage (ICCI) (Tono et al. 2012) and the International Corpus of Learner English (ICLE) (Granger et al. 2009). Thus, the analyses were performed on 12 learner data sets. The learner production was compared with native data taken from the Freiburg-London-Oslo-Bergen (FLOB) and Freiburg Brown (FROWN) corpora.

As the initial step in data analysis, the frequencies of the definite and indefinite articles in the subcorpora were tabulated. Next, lists of all three-word combinations containing the articles *the* and *a/an* were generated for each learner subcorpus using *Collocate*. Finally, the lists of three-word sequences in the learner data were compared against the native corpora in order to detect those combinations that in fact function as lexical bundles in English.

By analyzing the overall frequencies of articles and the frequencies of articles in lexical bundles, the study demonstrated that students with increasing proficiency became increasingly sensitive to the frequencies of articles and their reoccurring lexical contexts, and the conventional uses of lexical combinations became increasingly responsible for the selection of articles in the interlanguage. The conventional selections achieve and even surpass native-like frequencies, while the rule-based occurrences remain underused even by advanced learners. This tendency was true for learners with and without the article system in their L1s.

The awareness of the existence of the idiomatic uses of articles is certainly not new; however, until now, researchers had assumed that idiomatic uses played a marginal role in the acquisition of the article system by EFL learners. The present study demonstrated that even though a large proportion of the occurrences of articles were motivated by structural, semantic, and pragmatic rules related to the expression of referentiality, specificity, and countability in English, the role of conventional language in the acquisition and use of articles cannot be underestimated.

References

Butler, Yuko Goto (2002) Second language learners’ theories on the use of English articles: an analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition* 24(3): 451-480.

Crompton, Peter (2011) Article Errors in the English Writing of Advanced L1 Arabic Learners: The Role of Transfer. *Asian EFL Journal. Professional Teaching Articles*, 50: 4-34.

Ekiert, Monika (2004) Acquisition of the English article system by speakers of Polish in ESL and EFL settings. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 4(1): 1-23.

Granger, Sylviane; Dagneaux, Estelle; Meunier, Fanny & Paquot, Magali (2009) *International Corpus of Learner English v2. Handbook + CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Hakuta, Kenji (1976) A case study of a Japanese child learning English as a second language. *Language Learning*, 26: 321-351.

Jarvis, Scott (2002) Topic Continuity In L2 English Article Use. *Studies in Second Language Acquisition*, 24: 387-418.

Li, Haiyan & Yang, Lianrui (2010) An Investigation of English Articles' Acquisition by Chinese Learners of English. *Chinese Journal of Applied Linguistics* 33(3): 15-31.

Parrish, Betsy (1987) A new look at methodologies in the study of article acquisition for learners of ESL. *Language Learning* 37(3): 361-383.

Sinclair, John (1990), *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Tarone, E. (1985) Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning* 35: 373-403.

Thomas, Margaret (1989) The acquisition of English articles by first- and second language learners. *Applied Psycholinguistics* 10: 335-355.

Tono, Yukio, Kawaguchi, Yuji & Minegishi Makoto (eds.) (2012) *Developmental and Cross-linguistic Perspectives in Learner Corpus Research.* Amsterdam/Philadelphia: John Benjamins.

Software

Barlow, M. (2004) *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.

Assigning proficiency levels to computer-mediated communication - preliminary results from a learner corpus of Japanese university students' online writing

Marchand, Tim; Akutsu, Sumie
J. F. Oberlin University, Tokyo
marchand@obirin.ac.jp; smakutsu@obirin.ac.jp

For a number of years, scholars have strived to transfer insights gained from Learner Corpus Research (LCR) to the study of second language acquisition (SLA) and the practice of English language teaching (ELT), offering suggestions for future paths to follow along the way (Tono, 1999; Meunier, 2010). Several issues remain pertinent, however, for researchers interested in the overlap between LCR and both SLA and ELT. 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, as L1 interference may manifest itself differently at different levels of proficiency (Jarvis and Pavlenko, 2008), the assigning of learner proficiency within an individual corpus remains an important point of contention to resolve (Carlsen, 2012). Finally, the distinction has been made of using learner corpora for either delayed pedagogical use or immediate pedagogical use (DPU or IPU) in the production of corpus-informed ELT materials, with the latter somewhat underrepresented until recently (Granger, 2009; Meunier, 2010).

This poster presentation aims to broach these three issues by specifically addressing the problem of assigning proficiency levels to a learner corpus compiled from a university course in Japan. The course in question uses computer-mediated communication (CMC) for the dual purpose of providing lesson materials online and collecting student written output to develop the learner corpus. Here, CMC takes the form of a blog where, each week, an article about a recent news item, together with supporting class materials, is posted online for students to access. Students read the news story and, after a classroom session, write their reactions to the story on the class blog. These comments form the basis of the learner corpus and are assessed using a rating scale specifically designed to help with the analysis of the corpus.

Prior to investigation, a CMC-based corpus was assumed to be worth researching for three reasons. Firstly, research into a corpus of native English-speaker CMC suggests that its grammatical and lexical features differ significantly from both written and spoken registers of English, although overall its features may be considered as an intermediary register between the two (Yates, 1996; Murray, 2000; Marchand, 2013). This suggests that CMC could provide an interesting bridge between the more typical types of learner corpora, while at the same time being facilitative of processes beneficial to SLA from an interactionist perspective (Smith, 2004). Secondly, digital technology is often considered to be an effective way of connecting with the current population of students, as blogging and social networking are modes of communicating that many language learners use in their daily lives (Alm, 2006; Erbaggio et al., 2010). Therefore, a CMC learner corpus is not only readily comparable to native-speaker online communication, but it also replicates real-world behaviour in the learners' L1. Finally, other studies suggest that learner corpora of CMC make ideal pedagogical sources, especially of the IPU kind that have until recently been lacking (Belz and Thorne, 2006; Belz and Vyatkina, 2008).

While the assumptions listed above relate to the advantages of a CMC-based learner corpus and ELT, in order to make the research more relevant for the field of SLA, the designation of proficiency levels was seen to be of critical importance. Therefore, the poster presentation will seek to address the following research question: What is the best way to assign proficiency levels to individual corpus texts within a CMC-based corpus?

In answering the research question, the results of a learner-centred method and a text-centred method of level assignation (Carlsen, 2012) will be compared. The learner-centred approach uses a methodology similar to the one outlined in Pendar and Chapelle (2008), where identifiable traits from individual learner profiles are used to automatically assign levels of proficiency. The text-centred approach uses a rating scale derived from a Performance Decision Tree (PDT) that has been justified as effective in rating learner performances in other contexts (Fulcher et al., 2011).

The poster will highlight the preliminary findings of the research. These are that CMC is indeed a suitable medium for learner corpus construction; that a text-centred approach to level assignation is a better predictor of certain linguistic features within the corpus; and that a simplified version of the rating measurement tool could be used by learners themselves to analyse their own productions, thereby fulfilling recent demands for more IPU materials from learner corpora.

References

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

Belz, Julie, & Thorne, Steven (eds). (2006) *Computer-mediated intercultural foreign language education.* Boston, MA: Heinle & Heinle.

Belz, Julie, & Vyatkina, Nina. (2008) The pedagogical mediation of a developmental corpus for classroom-based language instruction. *Language Learning & Technology* 12(3): 33-52.

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

Erbaggio, Pierluigi, Gopalakrishnan, Sangeetha, Hobbs, Sandra, & Liu, Haiyong. (2010) Enhancing student engagement through online authentic materials. *International Association for Language Learning Technology* 42(2).

Fulcher, Glenn, Davidson, Fred, and Kemp, Jenny. (2011) Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28(1): 5 - 29.

Granger, Sylviane (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.

Jarvis, Scott, & Pavlenko, Aneta. (2008) *Crosslinguistic Influence in Language and Cognition.* New York: Routledge.

Marchand, Tim (to appear 2013) Speech in written form? A corpus analysis of computer-mediated communication. *Linguistic Research* 30 (2).

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

Murray, Denise (2000) Protean communication: The language of computer-mediated communication.

TESOL Quarterly 34(3): 397-421.

Pendar, Nick, & Chapelle, Carol. (2008) Investigating the promise of learner corpora: Methodological issues *CALICO Journal* 25: 189-206.

Smith, Bryan (2004) Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition* 26(3): 365-398.

Tono, Yukio (1999) Using Learner Corpora in ELT and SLA Research. Paper presented at the Symposium on the Roles of Corpora in Language Teaching and Language Engineering of the 12th World Congress of Applied Linguistics (AILA), Tokyo.

Yates, Simeon (1996) Oral and written linguistic aspects of computer conferencing: A corpus-based study. In Herring, Susan (ed) *Computer-mediated communication: Linguistic, social, and cross cultural perspectives* (pp. 29-46). Amsterdam & Philadelphia: John Benjamins.

Criterial features of pragmatic competence in a spoken corpus of Japanese learners of English to profile different levels of proficiency

Miura, Aika
Tokyo Keizai University
dawn1110am@gmail.com

This paper aims to explore what kind of *criterial features* of pragmatic competence can be identified to profile different levels of proficiency of Japanese learners of English, by investigating the NICT JLE (Japanese Learner English) Corpus. The Corpus contains more than 1-million-word interview transcripts of approximately 1,200 Japanese EFL learners taking a speaking proficiency test called the Standard Speaking Test (SST) (Izumi, Uchimoto, & Isahara 2004). The SST has five stages; (1) answering warm-up questions, (2) describing a single picture, (3) having a role-play with the examiner, (4) narrating picture sequences, and (5) answering wind-down questions. The subjects who took the test were assessed holistically into nine proficiency levels, which are novice (Level 1, 2 and 3), lower intermediate (Level 4 and 5), mid-intermediate (Level 6 and 7), upper intermediate (Level 8) and advanced levels (Level 9).

The present study aims to identify *criterial features* which specify what learners know and can do in English at each level of proficiency, especially in the domain of pragmatics (Hawkins and Filipović, 2012). The notion of *criterial features* originates in the English Profile, which is "to produce Reference Level Descriptors for English linked to the general principles and approaches of CEFR (Common European Framework of Reference)" (English Profile 2011, p.2). The CEFR provides can-do descriptors (i.e., Reference Level Descriptors), according to six proficiency levels including "Basic User" (A1 and A2), "Independent User" (B1and B2) and "Proficient User" (C1 and C2). Thus, according to the CEFR, "pragmatic competences are concerned with the functional use of linguistic resources (production of language functions, speech acts), drawing on scenarios or scripts of interactional exchanges" (Council of Europe 2001, p.13).

In the area of interlanguage pragmatics, discourse completion questionnaire (DCT) is one of major data collection. However, it has been criticized for not representing the features found in naturally occurring interactions (Chang 2010). Role-plays are possible alternatives as "they are useful tools for probing learners' ability to instantiate sociopragmatic and pragmalinguistic knowledge in interaction", "in a highly automatized fashion" (Kasper and Blum-Kulka 1993 p.61).

The current research deals with the learner data of role-pay tasks at Stage 3 in the NICT JLE Corpus. The following research questions will be addressed.
(1) What kind of language functions and speech acts can be found as *criterial features* to specify different levels of proficiency in the NICT JLE Corpus, referring to the Reference Level Descriptors in the CEFR?
(2) How are the SST levels in the NICT JLE Corpus correspondent with the CEFR Level, in terms of pragmatic competence?

The preliminary study dealt with a *shopping* role-pay which gives a situation where the examiner (interlocutor) plays a role of shop assistant and the subject (learner) plays a customer. Five types of speech acts were identified in the learners' production: (i) expressing the initial intention to purchase an item, (ii) asking for trying-out, (iii) asking for an alternative item, (iv) showing a negative reaction, and (v) negotiating for discount. The first one was found to be correspondent with A1, the next three with B1, and the last one with B2 level, based on illustrative scales for "transactions to obtain goods and services" in the area of spoken interaction provided in the CEFR (Council of Europe 2001 p.80). Thus, language features in each speech act were categorized into different degrees of politeness. The occurrence of less polite forms such as *I want* in (i) and (ii) drops as the proficiency increases. The use of mitigation such as *I prefer*, *a little bit* and *maybe* appears at Level 4 onwards. The higher proficiency learners show a tendency of using more polite forms with some hedges in the speech act of requesting. Other tasks such as *train*, which instantiate a situation where the same degree of social distance between a role of subject and interlocutor as *shopping*, will also be examined.

References

Chang, Yuh-Fang (2010) 'I no say you say is boring': the development of pragmatic competence in L2 apology. *Language Sciences* 32: 408-424.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

English Profile: Introducing the CEFR for English Version 1.1. (August 2011). (Retrieved 5 February, 2013 from http://www.englishprofile.org/images/pdf/theenglishprofilebooklet.pdf)

Hawkins, J. A., & Filipović, L. (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework.* Cambridge: Cambridge University Press.

Izumi, E., Uchimoto, K., & Isahara, H. (eds.) (2004). *Nihonjin 1200 nin no eigo speaking corpus.* [*L2 Spoken Corpus of 1200 Japanese Learners of English*]. Tokyo: ALC.

Kasper, G., & Blum-Kulka, S. (1993) Speech Act Realization. In G. Kasper & S. Blum-Kulka. (eds.), *Interlanguage Pragmatics* (pp. 59-63). Oxford: Oxford University Press.

Individual variation and the roles of L1 and proficiency in the longitudinal L2 development of English grammatical morphemes

Murakami, Akira
University of Cambridge
am933@cam.ac.uk

1. Background

Despite strong interests of SLA researchers in learner corpora (e.g., Housen, 2002), they have not fully entered mainstream SLA research yet. Possible reasons include (a) learner corpora have focused on the cross-sectional data revealing how L2 is used, rather than the longitudinal process of L2 acquisition, and (b) learner corpus research has been primarily descriptive, and had little to contribute to theory-building (Hasko, 2013). The present study fills the gaps by interpreting the data from a longitudinal learner corpus under the framework of dynamic systems theory (DST). More specifically, the study investigates the longitudinal development of L2 English grammatical morphemes.

The specific research questions addressed are the following;

i. Do learners' L1 and proficiency affect the developmental trajectories of accuracy of morphemes? Is the developmental pattern different across morphemes?

ii. To what extent is the intra- and inter-learner variability observed in the development?

2. Corpus

EF-Cambridge Open Language Database (EFCamDat) is a learner corpus containing learners' essays written in Englishtown, which is the online school of Education First. A course at Englishtown consists of 16 Lessons with eight Units each. At the end of each Unit is a free composition in which learners are asked to write on a certain topic. Thus, there can be up to 128 essays per learner. Learners in Englishtown receive feedback from native-speaker teachers on each essay. The feedback includes identification and correction of grammatical morphemes, among other things. The present study utilizes the feedback to calculate accuracy of morphemes.

3. Target Morpheme, L1 Groups, and Proficiency

The study targets three morphemes; articles, past tense *-ed*, and plural *-s*. Ten typologically diverse L1 groups were targeted; Portuguese, Chinese, German, French, Italian, Japanese, Korean, Russian, Spanish, and Turkish. A course at Englishtown covers A1 to C2 in the Common European Framework of Reference (CEFR) levels. The subcorpus of EFCamDat used in the study consists of 139,735 essays by 46,702 learners totalling 10 million words.

4. Data Analysis

For each essay and for each morpheme, the number of obligatory contexts, that of errors, and the instances of overgeneralization errors were identified with the aid of the feedback to the essays.

I built mixed-effects and generalized additive models predicts accuracy on the basis of morpheme, L1, proficiency, essay number, and their two-way interactions. Essay number refers to the number of essays the learner has written at that point and explains longitudinal within-learner development. In the mixed-effects models, learner ID was entered as a random-intercept, and morpheme and essay number as random-slope.

5. Results and Discussion

The mixed-effects models disclosed large individual variation in overall accuracy, accuracy difference between articles and the other morphemes, and the rate of development. At the same time, proficiency explains some part of the variance of the rate of development. That is, the developmental patterns are different depending on learners' proficiency. The accuracy transition tends to be flatter for the learners of higher proficiency. L1, however, is not likely to affect the developmental trajectory of morphemes. The developmental shape varies across morphemes as well. Past tense *–ed*, for instance, tends to exhibit flatter development than articles. Overall, however, individual variation is larger than the effect of the predictors. To the best of my knowledge, this is the first study that empirically quantified individual variation.

The findings were interpreted with DST (Verspoor, & de Bot, & Lowie, 2011). One's linguistic system is a dynamic system and any dynamic system interacts with numerous other variables, which invites constant change with chaotic variation. In other words, variability is the norm rather than the exception. This is what resulted in large intra- and inter-learner variation in the present study. At the same time, however, factors such as proficiency can influence development due to the characteristics of dynamic systems called "complete interconnectedness" and "attractor state". Complete interconnectedness refers to the fact that all the variables within a system (e.g., learner's L1 and L2) are connected to and influence each other. Attractor states are the states where a dynamic system prefers to (temporarily) settle down, and can affect the entire system as a "basin of attractors".

References

Hasko, V. (2013). Capturing the dynamics of second language development via learner corpus research: A very long engagement. *Modern Language Journal, 93*(S1), 1-10.

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., (Eds.), *Computer learner corpora, second language acquisition and foreign language learning* (pp. 77–116). Amsterdam: John Benjamins.

Verspoor, M., De Bot, K., & Lowie, W. (2011). *A dynamic approach to second language development: Methods and techniques*. Amsterdam: John Benjamins.

How can an error-annotated corpus tell us what to teach, and when?

Murcia-Bielsa, Susana; MacDonald, Penny
Universidad Autónoma de Madrid; Universitat Politècnica de València
susana.murcia@uam.es; penny@idm.upv.es

This paper presents the work carried out on error analysis under the TREACLE project. TREACLE aims at profiling the lexico-grammatical skills of Spanish learners of English at various proficiency levels in order to inform English language curriculum design. To this end, we have used a corpus of texts produced by Spanish learners of English at University level, to which we have applied both error analysis – to see what the learners are doing wrong – and automatic syntactic analysis – to show what structures students are actually using. This paper, however, focuses only on the error annotation part of our project.

Our study of errors takes an approach similar to that taken by others (e.g. Dagneaux et al. 1998), exploring the grammatical competence of learners by looking at the errors they make at each proficiency level. We do this so as to answer the following research questions: what are the most frequent errors made by Spanish university learners of English at each proficiency level? How can we map these errors onto a program of studies? In order to answer these questions, our study error-annotated learner texts taken from two corpora produced by Spanish university learners of English: the WriCLE corpus (Rollinson & Mendikoetxea 2010), containing texts produced by English Studies students, and the MiLC Corpus (Andreu et al. 2010), produced by students studying English for Specific Purposes. Of these texts, we annotated 307 texts, or 113,000 words, using UAM CorpusTool (O'Donnell 2008), identifying all errors in each text, producing a total of 16,200 errors. The error taxonomy identifies 113 distinct error categories at the most delicate level, distributed over lexical, grammatical, punctuation, pragmatic and phrasing errors. Furthermore, each essay in the corpus is associated with a score in the Oxford Quick Placement Test (UCLES 2001), so as to establish the grammatical proficiency level of each learner.

This error annotated corpus is a resource that can be used in various ways to inform teachers in a Spanish context as to what students need to learn (or be taught), with what degree of emphasis, and in what order. For instance, it is clear that the more frequently students make a particular error, the more emphasis is required in teaching that phenomenon. For this reason, we will present the most common errors in our corpus, focusing on lexical and grammatical errors. For instance, among the lexical errors, our findings indicate that language transfer is not the major source of errors in our corpus; on the contrary, intralanguage errors dominate even at lower levels of proficiency.

However, we will focus on grammatical errors, detailing the ten most frequent errors within this type and explaining what rules they break. In agreement with prior studies (Díez-Bedmar 2010, among others), errors related to the presence or absence of the article are most common. Other frequent errors include wrong choice of preposition, subject-finite agreement and absence of an obligatory subject.

It is well established that development of linguistic ability is staged, in that certain concepts are more easily acquired if other concepts have been acquired beforehand. Some grammatical concepts may not be effectively taught if prior concepts are not already acquired. Consequently, we show how the error-annotated corpus can be used to order the error types in terms of their "level of difficulty", with errors made more often by lower-level learners occurring earlier in the list, and those made by higher level students occurring later in the list. This difficulty-ordered list can be used to ensure that content related to the critical error types is taught to learners who are ready to receive this content.

References

Andreu, M., Astor, A., Boquera, M., MacDonald, P., Montero, B. & Pérez, C. (2010) Analysing EFL learner output in the MiLC project: An error *it's, but which tag? In: M.C. Campoy, B. Belles-Fortuno & M.L. Gea-Valor (eds.), *Corpus-Based Approaches to English Language Teaching* (pp. 167-179). London: Continuum.

Dagneaux, E., Denness, S. & Granger, S. (1998) Computer-aided error analysis. *System* 26: 163-174.

Díez-Bedmar, M.B. (2010) From secondary school to university: The use of the English article system by Spanish learners. In: Bellés-Fortuño, B., Campoy-Cubillo, M.C. & Gea-Valor, M.L. (eds.), *Exploring Corpus-based Research in English Language Teaching* (pp. 45-55). Castelló de la Plana: Publicacions de la Universitat Jaume I.

O'Donnell, M. (2008) The UAM CorpusTool: Software for corpus annotation and exploration. In: Bretones Callejas, Carmen M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 1433-1447). Almería: Universidad de Almería.

Rollinson, P. & Mendikoetxea, A. (2010) Learner corpora and second language acquisition: Introducing WriCLE. In: Bueno Alonso, J. L., González Álvarez, D., Kirsten Torrado, U., Martínez Insua, A.E., Pérez-Guerra, J., Rama Martínez, E. & Rodríguez Vázquez, R. (eds.), *Analizar datos > Describir variación / Analysing data > Describing variation* (pp. 1-12). Vigo: Universidade de Vigo (Servizo de Publicacións).

UCLES (2001) *Quick Placement Test (Paper and pencil version)*. Oxford: Oxford University Press.

How Do Educational Settings at the Secondary Level Impact on Learners' Use of the English Passive? – Evidence from the Secondary-Level Corpus of Learner English (SCooLE)

Möller, Verena
Université catholique de Louvain; Universität Hildesheim
verena.moeller@uni-hildesheim.de

In the south-west of Germany, the educational system at the secondary level offers a variety of settings for language learning and acquisition. In more and more schools, EFL (English as a Foreign Language) education is being supplemented with CLIL (Content and Language Integrated Learning) programmes, in which subjects such as Geography, History and Biology are taught in English during specific years (cf. MKJS 2004). Hence, the question arises whether CLIL is indeed as beneficial as it is assumed to be, or if, as has been argued by Bruton (2011), the success of CLIL programmes is simply based on the selectivity involved in many of them.

First and foremost, CLIL programmes offer an increase in exposure to the English language. However, CLIL materials, being scientifically oriented, also constitute a genre that is virtually absent from EFL settings. As research suggests that the passive is one of the characteristic features of scientific text (cf. Svartvik 1966, Holtz 2011), it has been chosen as a diagnostic criterion to investigate the impact of CLIL programmes on written learner language. The fact that the passive often alternates with a synonymous active structure, thus enabling learners to avoid it, adds to its importance in differentiating between more advanced learners and less advanced ones. Moreover, it is hoped that insights will be gained with respect to the lexis-grammar interface in language learning as passive forms of certain verbs are treated as lexical chunks by EFL materials.

To find out whether CLIL materials are indeed similar to scientific text, a corpus of teaching materials (Teaching Materials Corpus, TeaMC, ~1,000,000 words) was compiled. It comprises the following subcorpora:

(1) EFL materials Year 7-10;
(2) CLIL materials Year 7-10;
(3) EFL materials Year 11-12.

In a preliminary analysis, the passive was indeed found to occur almost three times more frequently in subcorpus 2 than in subcorpus 1, and even with a considerably higher frequency than in subcorpus 3, which acts as a reference norm that learners are supposed to aspire to.

To investigate differences in the written interlanguage of learners from EFL and CLIL programmes, the Secondary-Level Corpus of Learner English (SCooLE) was compiled. Data was elicited from Year 11 learners in mere EFL as well as EFL+CLIL settings at various secondary schools across the south-west of Germany. Participants were presented with two sets of essay topics, one of which was formulated in the passive. Learners subsequently typed two short argumentative essays in class. All in all, the SCooLE comprises about 850 essays, amounting to a total of around 250,000 words.

Due to the fact that the elicited text data was found to be highly deviant, the corpus had to be preprocessed in order to normalise especially those forms which have a serious impact on the automatic retrieval of passive constructions. This was, on the one hand, effected on the basis of VARD output (Variant Detector, cf. Rayson & Baron 2011), on the other hand by manually annotating typical misspellings. For annotation of part-of-speech, various tools were tested for their performance on interlanguage at this level. This resulted in the decision for concurrent use of the TreeTagger (cf. Schmid 1994) and CLAWS (Garside & Smith 1997), which, taken together, were shown to offer a recall rate (cf. Granger 1997) of 94 %. However, a number of erroneous passive constructions, which seem of particular relevance for the purpose of this study, remained irretrievable. Hence, manual annotation of all passives was effected.

To avoid results being influenced by intervening variables that might affect the performance of the two groups of learners (e. g. cognitive capacities, aspects of motivation or language learning/acquisition history in the individual learner), a questionnaire as well as two psychometric tests were administered. The information obtained from this procedure was included into the SCooLE in a rich set of metadata on learner variables.

A preliminary analysis shows that CLIL learners indeed use the passive more frequently than their non-CLIL counterparts. However, discrepancies were found with respect to cognitive capacities and other variables as well. It is thus one of the future aims of this study to determine whether or not the differences found in the interlanguages of the two groups are due to educational settings or a result of CLIL programmes being selective.

This paper describes the procedures involved in the compilation of the SCooLE in as far as they are relevant to the investigation of the passive. Furthermore, a comparison between the SCooLE and the TeaMC is effected, providing a quantitative analysis of passive constructions by using measures such as passive ratio (cf. Granger 2013). A preliminary qualitative analysis is carried out in order to describe the challenges involved in the investigation of the English passive in learners that often do not yet entirely master the lexical, morphological and syntactic processes involved in the use of this structure.

References

Bruton, Anthony (2011) Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System* 39: 523-531.

Garside, Roger & Smith, Nicholas (1997) A hybrid grammatical tagger: CLAWS4. In: Garside, Roger; Leech, Geoffrey & McEnery, Anthony (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102-121). London: Longman.

Granger, Sylviane (2013) The passive in learner English: Corpus insights and implications for pedagogical grammar. In Ishikawa, Shin'ichiro (ed.). *Learner Corpus Studies in Asia and the World. Vol. 1. Papers from LCSAW2013* (pp. 5-15). Kobe: School of Languages and Communication, Kobe University.

Granger, Sylviane (1997) Automated Retrieval of Passives from Native and Learner Corpora. Precision and Recall. *Journal of English Linguistics* 25(4): 365-374.

Holtz, Mônica (2011) *Lexico-grammatical properties of abstracts and research articles. A corpus-based study of scientific discourse from multiple disciplines.* Darmstadt: Technische Universität, PhD Thesis.

MKJS (Ministerium für Kultus, Jugend und Sport Baden-Württemberg) (2004) *Struktur des Unterrichts in den deutsch-englischen Abteilungen der Gymnasien.* www.schule-bw.de/unterricht/faecher/englisch/bilingual/organisation/merkblatt_o4o5.pdf (retrieved 10/07/2010).

Rayson, Paul & Baron, Alistair (2011) Automatic error tagging of spelling mistakes in learner corpora. In Meunier, Fanny; De Cock, Sylvie; Gilquin, Gaëtanelle & Paquot, Magali (eds.) *A Taste for Corpora. In honour of Sylviane Granger, Studies in Corpus Linguistics, 45.* Amsterdam: John Benjamins.

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing.* Manchester.

Svartvik, Jan (1966) *On Voice in the English Verb.* The Hague & Paris: Mouton.

What can oral learner corpora reveal about preposition use?

Susan Nacey; Anne-Line Graedler
Hedmark University College, Norway
susan.nacey@hihm.no; anneline.graedler@hihm.no

Mastering English prepositions is generally acknowledged as difficult, "… a traditional and recurring nightmare for all learners of English" (Littlemore & Low 2006: 285). In determining the appropriate preposition, learners face multiple challenges, including e.g. the polysemous nature of English prepositions as well as a lack of complete correspondence between English preposition use and preposition use in the learner's L1. Such potential problems are compounded by the manner in which prepositions may be presented in grammar books, where their various meaning extensions are frequently portrayed as arbitrary, leaving learners with few options other than to memorize prepositions "narrow context by narrow context" (Lindstromberg 1998: 227) and/or develop good dictionary-using habits (see e.g. Parrott 2010).

This paper adds empirical evidence concerning the real magnitude of the challenge that preposition use presents, through investigating the use of English prepositions in oral language produced by advanced learners. This investigation answers the following questions:

1) How often do these learners produce an inappropriate preposition?
2) Is there a correlation between inappropriate use and L1 influence?
3) Is there a significant difference between Norwegian learners' preposition use in oral and written language?

The data for the investigation is the Norwegian subcorpus of the Louvain International Database of Spoken English Interlanguage (Gilquin et al. 2010). The subcorpus contains 50 interviews of advanced English L2 students, amounting to approximately 13 hours of recorded and transcribed conversation.

All contextually inappropriate prepositions in the material have first been identified, indicating the frequency with which these learners produce inappropriate prepositions as well as showing which prepositions prove most difficult. One particular focus in this regard is whether challenges increase as the contextual meaning shifts away from a core, concrete meaning to a more peripheral, metaphorical meaning. All prepositions have thus also been classified according to metaphorical status (i.e. metaphorical or non-metaphorical), using the Metaphor Identification Procedure (Steen et al. 2010).

Further, the contextually inappropriate prepositions have subsequently been categorized in terms of their congruence between the L1 and L2 by virtue of two factors: 1) the syntactic structures required by the two languages in the particular context, and 2) the correspondence between the basic meanings in congruent cases. In congruent cases both languages require prepositions in context (factor 1). Application of factor 2 shows that there are three congruency patterns: basic congruence, where the basic meaning of the L2 preposition corresponds to the basic meaning of the L1 equivalent (example 1); divergent congruence, where the basic meanings of the L2 and L1 prepositions do not correspond (example 2), and zero congruence, where neither of the languages require a preposition (example 3). Example 2 is thus an indicator of L1 influence on the student's learner language, while also illustrating effects of real- time processing on preposition use in spoken language: hesitation, reformulation and self-correction (which in this case resulted in an inappropriate preposition).

Inappropriate use, L2	**Corresponding L1 prep**	**Appropriate L2 prep**	**Congruence**	**Metaphor status**
1. saying . why can't it just work and arguing <u>*at*</u> each other (NO037)	med	with	congruent: basic	metaphor
2. I realized that . living *in* a <u>*on*</u> a country . with few people around you is fantastic (NO008)	på	in	congruent: divergent	non-metaphor
3. it's basically: . f= fantasy game [...] and you . do . tasks and . move up <u>*in*</u> a level... (NO026)	Ø	Ø	congruent: Ø	metaphor

In addition, a cross-study comparison is carried out to provide a broader perspective on preposition use in learner language. Our empirical results from the analysis of oral preposition use in LINDSEI are thus compared with those from a previous investigation of preposition use in the written language of advanced English L2 students. This earlier study was based on 20,000 words from argumentative essays collected in the Norwegian component of the International Corpus of Learner English (Granger et al. 2009). Here it was shown that only 4.5% of the prepositions are contextually inappropriate, that L1 transfer does play an important role in production of the relatively few inappropriate prepositions produced, but that there is no correlation between core and peripheral meanings and inappropriate prepositions (Nacey 2010; forthcoming). In short, prepositions seem much less problematic than is generally believed.

References

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

Gilquin, G., S. De Cock, & S. Granger. (2010) *Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Louvain-la-Neuve: Presses universitaires de Louvain.

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

Lindstromberg, Seth (1998) *English prepositions explained*. Amsterdam: John Benjamins.

Littlemore, Jeannette & Low, Graham. (2006) *Figurative thinking and foreign language learning*. Basingstoke: Palgrave Macmillan.

Nacey, Susan. (2010) *Comparing linguistic metaphors in L1 and L2 English*. Unpublished doctoral dissertation, University of Oslo, Oslo.

Nacey, Susan. (forthcoming) *Metaphors in Learner English*. Amsterdam & Philadelphia: Benjamins.

Parrott, Martin. (2010) *Grammar for English language teachers*. Cambridge: Cambridge University Press.

Steen, Gerard., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, & Tryntje Pasma. (2010) *A method for linguistic metaphor identification: from MIP to MIPVU*. Amsterdam & Philadelphia: Benjamins.

Inter-rater reliability testing of article error tags in a small learner corpus: an argument for framework simplicity

Nickalls, Richard
English for International Students Unit, University of Birmingham.
r.nickalls@birmingham.ac.uk

1. Introduction

The assumption is often made that a single human annotator can reliably tag a corpus not only in terms of language correctness, but also in terms of complex explanatory classification frameworks. This study investigates this assumption by applying inter-rater reliability tests to a small corpus error-tagged for English article use.

2. Background

Han, Chodorow, & Leacock (2006) report a Kappa coefficient of just 0.48 in choice of article agreement between a computer and human classifiers. It was therefore decided to manually tag the original corpus of Mandarin L1 learners' English article interlanguage for correctness and then for context according to an adaptation of Bickerton's semantic space framework (1981). As shown in Figure 1, this much-used framework (Heubner 1983; Fen Chuan 2001; Humphrey 2007; Diez-Bedmar & Papp 2008) asks researchers to construe whether a referent is specific [± SR] and known [± HK] to the hearer.

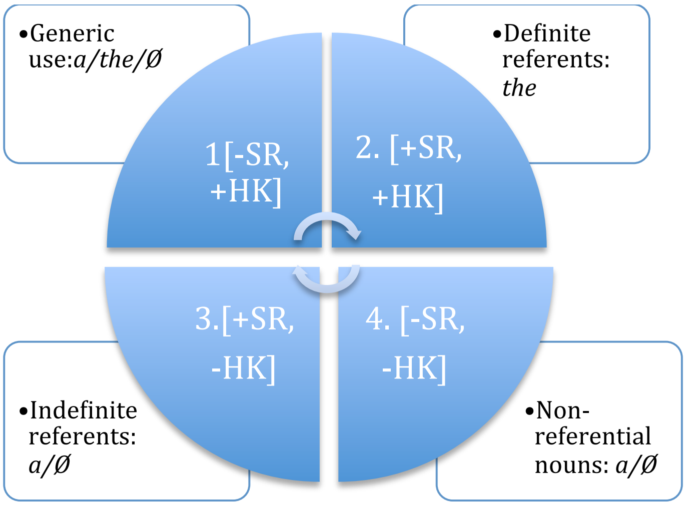

•Generic use:*a/the/Ø*
•Definite referents: *the*
1[-SR, +HK]
2. [+SR, +HK]
3.[+SR, -HK]
4. [-SR, -HK]
•Indefinite referents: *a/Ø*
•Non-referential nouns: *a/Ø*

Figure 1. Bickerton/Heubner framework of article use

Three research questions were investigated:
1) To what extent would raters reliably classify article use as 'correct' or 'incorrect'?
2) Would correctness be consistently classified over time?
3) How reliably would the complex classification framework be applied?

3. Method

A stratified random sample (n=112) was taken. The three new classifiers were all L1 speakers of English with over 10 years of English teaching experience, holding British Council TEFL-Q level teaching qualifications with either a Masters or Doctoral level of education. All researchers received identical training and reference materials and each was shown four short tutorial videos via a virtual learning environment (VLE) interface. Each also underwent standardisation on the VLE through classifying 25 noun phrases and receiving immediate online feedback. The three researchers then classified the sampled noun phrases using the online interface, which showed the noun phrases within their immediate sentence and a hyperlink to view the whole essay context of each noun phrase.

4. Findings

In the first session, the only dichotomous choice was of 'correctness' of the 112 noun phrases with the researchers instructed to 'classify as incorrect only if stylistically or grammatically impossible within an academic writing context without changing the form of the head noun'. As shown in table 1, the three raters showed a high degree of agreement with the initial annotator. As outlined in table 2, a Fleiss' Kappa of 0.74 again shows consistent agreement about correctness among all raters overall.

Table 1: Coefficient comparisons between first rater and others

	Percent Agreement	Agreements	Disagreements	Cases	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
rater 1 and 2	86.61	97	15	112	0.73	0.73	0.73
rater 1 and 3	90.18	101	11	112	0.80	0.80	0.80
rater 1 and 4	87.50	98	14	112	0.75	0.75	0.75

Table 2: Overall correctness ratings in first session of all raters

Coders	Cases	Average pairwise percent agreement	Fleiss' Kappa	FK observed agreement	FK expected agreement	Krippendorff's Alpha
4	112	86.9	0.74	0.87	0.50	0.74

In the second session (three weeks later) the researchers were asked to review 28 of the same noun phrases and classify them again in terms of their correctness/the explanatory framework. As shown in table 3, the first rater made identical correctness classifications to the ones made two years previously. There were nevertheless minor inconsistencies (Cohen Kappa between 0.79 and 0.93) between the other raters' first and second classifications.

Table 3: Consistency of rating over time.

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
Rater 1	100	100	1.00	1.00
Rater 2	89.29	0.79	0.79	0.79
Rater 3	96.43	0.93	0.93	0.93
Rater 4	89.29	0.79	0.79	0.79

However, the Bickerton-based framework showed much lower inter-rater agreement, as can be seen in Table 4. A kappa of 0.26 reflects the 0.25 probability of random chance agreement. Furthermore, even if the probability of chance agreement is not offset, the average percentage agreement (44.64%) is low given the training measures taken to ensure standardization.

Table 4: Reliability of the application of the explanatory framework

Average pairwise percent agreement	Fleiss' kappa	Krippendorff's Alpha
44.64%	0.26	0.26

5. Discussion

These findings confirm expectations that human raters would be more reliable and consistent than computers at decisions of correctness. However, classifications with the complex explanatory framework were not reliable. It might naturally be argued that further rater training could result in higher Kappa agreement. Yet extended standardisation would probably result in only marginal improvements among one team of researchers, while the implied benefit of such frameworks is that studies can be replicated and compared. These findings must certainly be treated with caution given the study's small size. Indeed, future research using more raters and an extended sample has already began with the purpose of further testing the Bickerton/Heubner framework and developing an alternative framework. Nevertheless, the findings of this study do present a clear case for questioning assumptions of human reliability, and for regarding simple tagging frameworks as more robust and reliable than complex ones.

References

Bickerton, D. (1981) Roots of language. Repr. 1985. Ann Arbor, Mich.: Karoma.

Diez-Bedmar, M.B. and Papp, S. (2008) 'The use of the English Article System by Chinese and Spanish learners'. In Gilquin, G., Papp, S. & Diez-Bedmar, M.B. (Eds.) *Linking up Contrastive and Learner Corpus Research.* Amsterdam, Atlanta, Rodipi 147-175.

Fen Chuan, C. (2001). 'The acquisition of English articles by Chinese learners'. *Second Language Studies*: 20(1): 43-78.

Han, N., Chodorow, M., Leacock, C. (2006). 'Detecting errors in English article usage by non--native speakers'. *Natural Language Engineering*: 12, pp 115--129

Huebner, T. (1983) Longitudinal analysis of the acquisition of English. Ann Arbor: Karoma Pubr.

Humphrey, S.J. (2007). 'Acquisition of the English Article System: Some Preliminary Findings'. *Journal of School of Foreign Languages* [Online] available from http://library.nakanishi.ac.jp/ [accessed on 1/6/2011].

Data commentary in science writing: a corpus-based comparison of research articles and master's theses in technical fields for formative self-assessment practices

Nordrum, Lene; Eriksson, Andreas
Chalmers University of Technology/Lund University; Chalmers University of Technology
lene.nordrum@englund.lu.se; andreas.eriksson@chalmers.se

In the wide context of English for Academic Purposes (EAP) and English for Specific Purposes (ESP), increasing attention is devoted to an integrated approach of a linguistically oriented bottom-up corpus analysis and discourse-oriented top-down analysis of macro-structural units, or moves, such as Swales' (1990) model *Creating A Research Space* (CARS) of introduction sections in research papers (Ädel & Reppen 2008; Biber et al. 2007;). One prerequisite for this approach is a genre-specific corpus annotated for discourse moves (Flowerdew 2004). Although such corpora work at the expense of general language description, they enable a thicker analysis and description of disciplinary discourse than a more general corpus. Importantly, these corpora have the applied benefit of providing students with a wider repertoire of disciplinary-sanctioned linguistic resources that can form the foundation for corpora-supported learning activities. Some promising work has been carried out in this vein, notably Kanoksilapatham's (2005) analysis of the linguistic manifestation of moves in biochemistry articles, and Stoller & Robinson's (2012) pedagogically oriented study of moves in chemistry articles.

In this study, we use a discipline-specific corpus of research articles and master's theses to consider the commenting on figures, images and tables for result presentation in science writing. The multi-modal nature of science writing has been pointed out as an 'important problem' in linguistic approaches to disciplinary discourse (Shaw 2007), but remains under-investigated. Two issues are particularly important in this context*: modal affordance*, i.e., knowledge of the strengths and weaknesses of particular modes, and *transduction*, i.e., translation of the visual mode to the written mode (Kress, 2010). From an applied perspective, research has shown that integrating written and visual modes represent a complex task for students (Wharton 2012), and that transduction practices can vary even between closely related fields (Stoller & Robinson 2013).

The study aims to discuss how the discipline-specific corpus can be used in formative self-assessment practices in ESP courses at technical universities. As a first step, this aim involves comparisons of expert (research articles) and novice (master theses) writing. The corpus comprises data commentaries from 10 research articles and 10 master's theses from applied and theoretical fields in three hard-science disciplines and is annotated for discourse moves by means of the UAM corpus tool, developed by Michael O'Donnell. Preliminary results from applied chemical engineering indicate that novice writers rely more on 'global comments' of visual modes than more experienced writers, and point less to particular parts of a visual representation. This might be due to master students' putting too much trust in their audience's capacity for making sense of visual material. There is also indication that novice writers less often comment on the reason for choosing a certain visual representation compared to experienced writers, which we believe could be due to novices' limited experience with alternative visual modes and their particular modal affordances. A final preliminary observation ties in with the tendency for 'global comment', namely that students sometimes present more or less 'naked' results with little or no discussion of implication.

We argue that our corpus has important applied potential in that students can use it to explore differences in authentic novice and expert data commentaries in their own discipline, which facilitates awareness of disciplinary conventions. This awareness can then be used to develop formative self-assessment practices, which we believe is imperative to universities for two related reasons. First, pedagogical and curricular developments in science education emphasize students' awareness of communicative practices and variance. Second, as audits of higher education become increasingly informed by student work, students' self-assessment skills can provide institutions in higher education with important competitive advantages.

References

Ädel, A. & R. Reppen (2008) *Corpora and Discourse: The Challenges of Different Settings.* Amsterdam: Benjamins.

Biber, D., U. Connor, J. Jones & T. Upton (2007) *Discourse on the Move.* Amsterdam: Benjamins.

Flowerdew, L. (2004) The argument for using English specialized corpora. In: U. Connor & T. .A. Upton (Eds.). *Discourse in the Professions: Perspectives from Corpus Linguistics* (pp. 11-33). Amsterdam: John Benjamins.

Kanoksilapatham, B. (2005) Rhetorical structure of biochemistry research articles. *English for Specific Purposes* 24: 269-292.

Kress, G. (2010) *Multimodality: a Social Semiotic Approach to Contemporary Communication.* London: Routledge.

Shaw. P. (2007) Introductory Remarks. In: Fløttum, K. (Ed). *Language and Discipline Perspectives on Academic Discourse* (pp. 2-13). Newcastle: Cambridge Scholars Publishing

Stoller, F-L. & M.S. Robinson (2013) Chemistry journal articles: An interdisciplinary approach to move analysis with pedagogical aims. *English for Specific Purposes* 32: 45-57.

Swales, J.M. (1990) *Genre Analysis. English in Academic and Research Settings.* Cambridge: Cambridge university press.

Wharton, S. (2012) Epistemological and interpersonal stance in a data description task: Findings from a discipline-specific learner corpus. *English for Specific Purposes* 31: 261-270.

Demonstration of UAM CorpusTool 3.0 for learner corpus annotation and exploration

O'Donnell, Mick
Universidad Autonoma de Madrid
Michael.odonnell@uam.es

This session will demonstrate UAM CorpusTool 3.0, software for the manual and automatic annotation of texts (O'Donnell 2008). The software is available free for Windows and Macintosh from http://www.wagsoft.com/CorpusTool/.

The software offers an easy-to-use interface for annotating multiple texts on a number of linguistics levels. The user can add as many layers as they like, choose between either using the predefined coding schemes (e.g., the Louvain Error Scheme, an Appraisal scheme, etc.), or building their own scheme from scratch. Schemes can be modified even after coding has started, and the software ensures that the codings are consistent with the scheme even as it is modified.

Most commonly, the software is used for manual annotation: the user selects a segment of text, and then assigns features to that segment. A special option is available for error coding: when the user indicates the layer involves coding errors, a space is provided on the coding interface to type in the correction of the error. Error-coded texts can then be saved in an HTML format for return to students, who can then view their essays, with errors highlighted in red, and place the cursor over an error to see its category, the correction, and any comments provided by the coder/teacher.

Where the texts being annotated are in English, the user can add layers for automatic annotation, including basic syntactic structure (e.g., Subject/Predicate/Direct Object, etc.), Theme/Rheme (following Halliday & Matthiessen 2004) and Transitivity (also following Halliday, e.g., Actor/Process/ Goal/Circumastance, etc.). Each element of structure is also assigned syntactic features. For instance, in the syntactic analysis, each clause is assigned a feature of Voice (active vs. passive), Mood (declarative, interrogative, imperative), Tense-Aspect (simple-present, present-perfect, past-continuous, modal-perfect, etc.), clause-type (relative, that-clause, wh-nominal, infinitive, etc.), and so on. These features can be used in segment search, and in statistical studies, both built into the tool.

The demonstration will focus on the use of these automatic layers with a small corpus of essays by learners of English. We will start by showing how to load texts into the system. Then, layers of analysis will be defined: one layer to define writer characteristics (e.g., proficiency level, gender, academic year), and other layers for syntactic, thematic and transitivity analyses.

The demonstration will then show how this automatic annotation can be used to study the learner language. Firstly, a comparative study will contrast syntactic, thematic and transitivity options between low and high proficiency learners, showing which of the linguistic features are significantly different between the two groups.

We will then move on to various ways to visualize patterns in the corpus. This will include:

- Bar charts presenting the changing pattern of usage of various syntactic features as proficiency increases.
- Use of Principle Components Analysis to show how sets of texts relate to each other.
- Tag-clouds showing the syntactic features that are most key to a particular proficiency level.

References

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

O'Donnell, Mick (2008) Demonstration of the UAM CorpusTool for text and image annotation. *Proceedings of the ACL-08:HLT Demo Session* (CompanionVolume), Columbus, Ohio, June 2008 (pp. 13–16). Association for Computational Linguistics. pages 6.

An English collocations E-workbook designed to Brazilian Portuguese speakers

Orenha-Ottaiano, Adriane
Universidade Estadual Paulista “Júlio de Mesquita Filho”
adriane@ibilce.unesp.br

Recent studies have revealed the relevance of collocations in the current sphere of second language learning and teaching (Thomas, forthcoming; Meunier & Granger 2008; Nesselhauf, 2005; Orenha-Ottaiano 2012; Sinclair 2004; Conzett 2000 etc.). Due to that, 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 are related to current world news such as *Financial crises in Europe*; *Unemployment*; *Elections in the US*; *Bullying*; *Marijuana Legalization* etc.

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. Based on the 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. For instance, as learners seemed to be influenced by their mother tongue, they translated the collocation *realizar primárias* into *realize primaries*, instead of *hold primaries*. Some students also had problems translating *derrubar a resistência* (= *to break down the resistance*). Sixty per cent of the translation options are recurrent in English, such as *break down the resistance* and *take out the resistance*. However, some combinations employed by the students are not frequently used in English (*overthrew the resistance*, *knock down the resistance* and *topple the resistance*) or not used at all (*upend the resistance*).

As the collocations E-workbook is being designed for Brazilian Portuguese speakers, the exercises are being 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 have a 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 are also given the collocational activities built for the proposed E-workbook. Teachers are encouraged to do and evaluate them, so that we can choose the ones which are more suitable for their learning and for the teaching of their own students.

We hope 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. We believe that 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. Additionally, students will count on a new electronic material specially designed for Brazilian Portuguese learners of English.

References

Davies, Michael (1990-2012) *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available: <http://corpus.byu.edu/coca/.>. Acessed: April 20th, 2012.

Granger, Sylviane (eds.) (2008) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia:John Benjamins.

Mackin, Ronald. (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.

Nesselhauf, Nadja (2005) *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: John Benjamins.

Orenha-Ottaiano, Adriane (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, Mike (2008) *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.

Sinclair, John McHardy (2004) *How to use corpora in Language Teaching*. Amsterdam: John Benjamins.

Thomas, James (forthcoming) Stealing a march on collocation. *TALC 10 Proceedings*.

Comparisons are odorous: native-speaker data in learner corpus research

Osborne, John
Université de Savoie, France
John.Osborne@univ-savoie.fr

Analyses of learner corpora generally involve making comparisons between different groups of speakers: between learners with different language backgrounds or at different levels of proficiency, between the productions of the same learners at different points in time, or more rarely, between productions in different target languages. It is also frequent for native-speaker data to enter into the comparison, either explicitly, when comparable corpora of native-speaker productions are used to identify divergences between native productions and learner productions, or more implicitly, for instance in annotating errors in a learner corpus. However, the status of the native speaker, both in linguistics and in language acquisition research, has been the subject of considerable discussion. Reservations about the use of native-speaker data as a standard against which to compare learner productions can be situated at several levels:

- The concept itself is problematic. It can be argued that there is really no such thing as a native speaker, who could serve as a yardstick of acceptability (Paikeday 1985), that the native speaker can only be defined negatively (Davies 1993, 2001; Han 2004), or that native-speakerhood is better regarded as a gradient term with core and peripheral features (Escudero & Sharwood Smith 2001).
- Even if we have criteria for identifying a native speaker, the learner's interlanguage should be studied in its own right, not as a degenerate form of the target system (Bley-Vroman 1983).
- For L2 learners, native-like performance is an unattainable goal (Cook 1993); conversely, some native-speakers may be unable to perform satisfactorily in tasks that are commonly demanded of learners at more advanced levels (Hulstijn 2011).
- Languages are not the property of any particular group of speakers; they are "open source" (Seidlhofer 2012). Consequently, there is no reason why second language users should be expected to conform to the patterns observed in native-speaker usage.
- By learning a second (or subsequent) language, learners become different from monolinguals (Cook 2003), in ways that may have an impact on their their own L1, so that the native speaker standard could be viewed not as a static benchmark, but as a "moving target" (Brown & Gullberg 2008).

In the light of these reservations, what role is there for comparisons with native-speaker data in learner corpus research? I will argue that, if appropriate precautions are taken, the variability of native-speaker date can in fact be an advantage in analysing learner corpora, particularly with respect to what the Common European Framework of Reference identifies as the "generic qualitative factors which determine the functional success of the learner/user" (Council of Europe 2001: 128), namely fluency and propositional precision.

References

Bley-Vroman, Robert. (1983) The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1): 1–17.

Brown, Amanda & Gullberg, Marianne (2008) Bidirectional crosslinguistic influence in L1-L2 encoding of manner in speech and gesture: A study of Japanese speakers of English. *Studies in Second Language Acquisition* 30(2): 225-251.

Cook, Vivian (1999) Going beyond the native speaker in language teaching. *TESOL Quarterly* 33 (2): 185-209.

Cook, Vivian (2003) Introduction: The changing L1 in the L2 user's mind. In: Cook, Vivian (ed.), *Effects of the Second Language on the First* (pp. 1-18). Clevedon: Multilingual Matters,

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Davies, Alan (1991) *The Native Speaker in Applied Linguistics*. Edinburgh University Press.

Davies, Alan (2003) *The Native Speaker: Myth and Reality*. Clevedon: Multilingual Matters.

Han, Zhaohong (2004) To be a native speaker means not to be a nonnative speaker. *Second Language Research* 20(2): 166-187.

Hulstijn, Jan (2011) Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly* 8(3): 229-249.

Paikeday, Thomas (1985) *The Native Speaker is Dead!* Toronto and New York: Paikeday Publishing. Online, http://www.paikeday.net/speaker.pdf

Paola Escudero and Michael Sharwood Smith (2001) Reinventing the native speaker or 'What you never wanted to know about the native speaker so never dared to ask.' In: Foster-Cohen, Susan. & Nizegorodcew, Anna (eds), *Eurosla Yearbook: Volume 1* (pp. 275–286). Amsterdam & Philadelphia: John Benjamins.

Seidlhofer, Barbara (2012) Anglophone-centric attitudes and the globalization of English. *Journal of English as a Lingua Franca* 1(2): 393 – 407.

Cross-linguistic influence and formulaic language: French EFL learners' use of recurrent word sequences under scrutiny

Paquot Magali
FNRS, Université catholique de Louvain, Centre for English Corpus Linguistics
magali.paquot@uclouvain.be

The last few years have witnessed a remarkable boom in the number of phraseological studies that examine learners' use of lexical bundles, i.e. "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al. 1999: 990). Some of the studies have put specific patterns of misuse, overuse and underuse of lexical bundles down to the learners' mother tongue (see Paquot & Granger, 2012 for a review). Allen (2011: 111), for example, partly ascribed Japanese students' infelicitous use of the singular noun *result* in bundles such as *result of this experiment* (e.g. in the Japanese learner's sentence: *The result of this experiment was expressed by following graphs*.) to the fact that the corresponding form in Japanese can denote both single and multiple findings and similarly attributed Japanese learners' overuse of *it can be said that* to the L1, as its translational equivalent is repeatedly used in Japanese academic writing. Rica (2010) adopted a lexical bundle approach to the study of linking adverbials in EFL learner writing and noted that a large proportion of the multi-word connectors that non-native writers overused were very similar to those word sequences which learners use in their L1 to express similar meanings (e.g. English and Spanish *I think ~ Creo que* and *for example ~ por ejemplo*). However, no study has targeted transfer effects on EFL learners' production of recurrent word sequences as their primary object of investigation.

One of the main objectives of the author's current research is to fill in this existing gap by conducting careful transfer studies of recurrent word combinations in EFL learner writing (Paquot, in press; Paquot, 2013). The present work deals with lexical bundles of various lengths (from 2 to 5 words) and focuses more particularly on:

1. Correct and incorrect multi-word sequences that fulfil organizational or rhetorical functions, e.g. contrasting (*on the contrary*), exemplifying (*let us take the example*), concluding (*in conclusion, we can say*),
2. Preferred co-occurrences that exemplify collocations (e.g. *deeply rooted*) colligations (e.g. *considered as*) and syntactic structures (e.g. *role to play*).

The study makes use of Jarvis's (2000) methodological framework to test transfer effects on recurrent word sequences in the French component of the *International Corpus of Learner English* (ICLE) (Granger et al. 2009) as compared to nine other ICLE learner sub-corpora. Intra-L1-group homogeneity (Effect 1) is most evident when directly compared with inter-L1-group heterogeneity (Effect 2) (Jarvis 2000), and I, therefore, rely on comparison of means tests and post hoc tests to examine Effects 1 and 2. While the first two effects readily lend themselves to automatic and quantitative evaluation, intra-L1-group congruity between French learners' L1 and IL performance does not. Assessing this third effect requires a more qualitative approach. First, the use of each lexical bundle was carefully analysed in ICLE-FR. The next steps consisted in identifying the French potential 'equivalent' of each lexical bundle in context, describing its use in French L1 corpora and comparing learners' L1 and IL patterns of use.

Applying Jarvis's (2000) unified framework on learner corpus data brings to light interesting findings relating to L1 influence on word use. It helps to identify a number of transfer effects that remain largely undocumented in the SLA literature: transfer of function, transfer of the phraseological environment, transfer of style and register, and transfer of L1 frequency. These transfer effects make up what, following Hoey (2005), I refer to as transfer of 'lexical priming'. EFL learners' knowledge of words and word combinations in their mother tongue includes a whole range of information about their preferred co-occurrences and sentence position, stylistic or register features, discourse functions and frequency. Primings for collocational and contextual use of (at least a restricted set of frequent or core) L1 lexical devices are particularly strong in the mental lexicon of adult EFL learners. They are the result of many encounters with these lexical items in L1 speech and writing. Mental primings in the L1 lexicon probably influence EFL learners' knowledge of English words and word sequences by priming the lexico-grammatical preferences of an L1 lexical item to its English counterpart. These results support Kellerman's claim that the 'hoary old chestnut' according to which transfer does not afflict the more advanced learner "should finally be squashed underfoot as an unwarranted overgeneralization based on very limited evidence" (Kellerman, 1984: 121). However, they also suggest that the main effect of the first language on higher-intermediate to advanced EFL learners' use of recurrent word combinations is not errors (compare with transfer effects on learners' use of collocations as reported in Nesselhauf, 2005 and Laufer & Waldman, 2011). Rather, findings provide more subtle evidence of L1 influence in the form of patterns of overuse and underuse (see also Neff van Aertselaer, 2008; Paquot, 2010).

References

Allen, D. (2011). Lexical bundles in learner writing: an analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education* 1.

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

Granger S., Dagneaux E., Meunier F. & Paquot M. (2009) *The International Corpus of Learner English*. (second edition). CD-Rom and Handbook. Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com

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

Jarvis S. (2000) Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning* 50(2): 245-309.

Kellerman E. (1984) The empirical evidence for the influence of the L1 in interlanguage. In Davies A., C. Criper and A. Howatt (eds) *Interlanguage* (pp. 98-122). Edinburgh: Edinburgh University Press.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2): 647–672.

Neff van Aertselaer, J. (2008). Contrasting English-Spanish interpersonal discourse phrases: A corpus study. In F. Meunier, & Granger, S. (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 85–100). Amsterdam & Philadelphia: Benjamins.

Nesselhauf N. (2005) *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: Benjamins.

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London & New-York: Continuum.

Paquot, M. (2013). Transfer effects on French EFL learners' use of textual phrasemes. Paper presented at EUROSLA, University of Amsterdam, Amsterdam, The Netherlands, 28-31 August 2013.

Paquot, M. (in press). Transfer effects and lexical bundles. *International Journal of Corpus Linguistics* 18(3).

Paquot, M. and Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics* 32: 130-149.

Rica, J.P. (2010). Corpus analysis and phraseology: transfer of multi-word units. *Linguistics and the Human Sciences* 6: 321–343.

Compiling a Learner Corpus of Singaporean English: Design, Tools, and Challenges

Park, Kwanghyun; Alsagoff, Lubna
National Institute of Education, Singapore
kwanghyun.park@nie.edu.sg; lubna.alsagoff@nie.edu.sg

In Singapore's multilingual context, English has occupied an important position in the economic prosperity and the multicultural diversity of the nation (Alsagoff 2010; Chew 2006). Efforts to ensure the intelligibility of Singapore English are evident in the educational policy as evident in the new language syllabus launched in 2010. However, a majority of Singaporean children grow up speaking their heritage language, or the home language, as a primary means of communication, while using English in more formal settings. Therefore, the contact between English and one or more other languages spoken for informal communication is a common linguistic phenomenon in Singapore and it may also influence the English used by Singaporean children. This variety of English is often considered as a deviation from 'standard English', or non-standard English.

From an educational perspective, there seems to be an obvious benefit in describing, understanding, and focusing on the differences between the version of English used by Singaporean young learners and the variety of English taught in schools as 'standard' English, i.e. British English. The major challenge to the language educators in Singapore, however, is a lack of a large corpus of learner English that will allow a systematic and comprehensive examination of English used by the young learners. A very recent effort is Guo and Hong's (2009) study of the development and use of grammatical metaphor in a corpus of 21 Primary 5 and 12 Secondary 3 student recount essays. This study compares different phases but has not yet provided a comprehensive database. In our presentation, we will describe an emerging effort to design and compile such a large-scale learner corpus of Singaporean learners' English.

In this preliminary phase, we aim to collect the data over the span of six-year primary school education. The data will largely consist of approximately 2,000 written works of primary school children. Using the data, two kinds of analyses will be performed. First, drawing on the methodological framework suggested in learner corpus research (Granger 2003, 2004), the data will be analyzed to reveal the instances of misuse (or non-standard use), underuse, and overuse of lexical items and grammatical patterns. Second, using the statistical method suggested in corpus-based register studies (Biber 1998), a multi-feature/multi-dimension analysis will be performed to identify the overall patterns and their change across different stage of the children's linguistic development.

The proposed analyses, i.e., computer-assisted error analysis and multi-feature/dimension analysis, are only possible when the data is fully annotated for their non-standard use and lexico-grammatical features (features suggested in Biber's work). For annotating non-standard uses, two experienced annotators will identify and tag each instance of them using an XML-based software program. A specialized tag set, or annotation scheme, will be developed for the purpose. For multi-feature dimension analysis, a computer program has been written to process some features, while other features that the program cannot annotate will be manually tagged. This multi-feature annotation will be, again, done in XML format to be integrated with the error annotation in order to reveal the correlation between the lexico-grammatical features and the non-standard uses (Alsagoff & Ho 1998).

One useful outcome of the project will be a web-based system that will allow searching and browsing the non-standard uses of English of the Singaporean learners. The system will serve as a convenience tool for teachers to look up the examples of the common errors and non-standard uses in Singaporean learners' English. Further, the system will be a useful resource for teacher education by allowing the pre-service teachers to understand the non-standard features of Singaporean English. Finally, the system will be a valuable resource for researchers of Singaporean English by providing a comprehensive collection of learner language data as well as a computerized tool to access the data. In presenting our work in progress, we will discuss our design and the current state of the project and also share with the audience the challenges we have encountered.

References

Alsagoff, Lubna (2010) English in Singapore: culture, capital and identity in linguistic variation. *World Englishes* 29(3): 336–348.

Alsagoff, Lubna & Ho, Chee Lick (1998) The grammar of Singapore English. In: Foley, T. Kandiah et al. (eds.) *English in New Cultural Contexts: Reflections from Singapore* (pp.127-151). Singapore: Oxford University Press.

Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Chew, Phyllis (2006) Remaking Singapore: Language, culture and identity in a globalized world In: Tsui, Amy & James W. Tollefson (eds.) *Language policy culture and identity in Asian contexts* (pp. 73-93). Mahwah: Lawrence Erlbaum.

Granger, Sylviane (2003) Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3): 465-480.

Granger, Sylviane (2004) Computer learner corpus research: Current status and future prospects. In: Ulla Connor & Thomas T. Upton (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 123-145). Amsterdam: Rodopi.

Guo, Libo & Hong, Huaqing (2009) Metaphorization in Singaporean student writing: A corpus-based analysis. In: Silver, Rita, Goh, Christine & Alsagoff, Lubna (eds.) *Language acquisition and development in new English contexts* (pp. 112-131). London: Continuum.

Using data mining to discover evidence of cross-linguistic influence in learner corpora

Pepper, Steve
University of Oslo
pepper.steve@gmail.com

This presentation describes an investigation into cross-linguistic influence ('language transfer') in Norwegian interlanguage using the statistical methods of data mining (Pepper 2012).

Data mining is the process of discovering patterns in large data sets and involves methods at the intersection of artificial intelligence, machine learning, statistics and database systems. Such methods are widely used with financial, meteorological and medical data, and have also been applied in stylometry, text classification and natural language identification.

The idea of using these methods in transfer research was first suggested by Jarvis (2010) and dubbed the 'detection-based approach' (to distinguish it from the more traditional 'comparison-based approach'). Its first systematic application with learner corpora was a series of studies employing data from the International Corpus of Learner English. The goal of these studies was to predict the L1 backgrounds of L2 English learners on the basis of lexical, stylistic and other features of their written language – a goal that was achieved with remarkable success (Jarvis & Crossley 2012).

In the present study, the same techniques were applied using data from the Norwegian Second Language Corpus in order to address the following research questions:

1. Can the methods of data mining be used to identify the L1 background of learners of L2 Norwegian on the basis of their use of lexical features of the target language?
2. If so, what are the best predictors of L1 background?
3. And can those predictors be traced to cross-linguistic influence?

The source data consisted of Norwegian interlanguage texts written by 1,000 second language learners from ten different L1 backgrounds (German, English, Dutch, Spanish, Polish, Russian, Serbo-Croat, Albanian, Somali, Vietnamese). There was also a control corpus of 100 texts written by native speakers. Word frequencies computed from this data were analysed using multivariate statistical methods, including analysis of variance (ANOVA) and discriminant analysis.

Discriminant analysis is defined by Klecka (1980) as "a statistical technique which allows the researcher to study the differences between two or more groups of objects with respect to several variables simultaneously." In the present study, the 'objects' under study were learner texts; the 'groups' were defined according to the authors' L1 backgrounds; and the 'variables' were the relative frequencies of the 50-60 most commonly occurring words in the texts. Based on this input, mathematical models were computed that proved capable of predicting the authors' L1 backgrounds with a statistically significant degree of accuracy.

For example, in a test involving the five L1 groups German, English, Polish, Russian and Somali, the author's L1 was correctly predicted for 288 of the 500 texts, giving an overall success rate of 57.6%, compared to the 20% success rate that would be expected if the prediction was completely random. This shows that the discriminant analysis had found a significant amount of L1 grouping structure in the data and the first research question was thus answered in the affirmative. Further analysis using the method of feature selection made it possible to determine exactly which lexical features contributed most to the model and thus constituted the best predictors of L1 background (research question 2).

Those predictors were subjected to additional tests in order to determine which L1 groups the various lexical features served to separate. The results both confirmed existing knowledge and revealed a number of hitherto unsuspected patterns. For example, the well-known tendency for Russian and Polish speakers to omit indefinite articles was confirmed (and shown to apply to Somali learners as well). More surprisingly, Dutch speakers were shown to overuse the modal verb *skal* ('shall') to a significantly greater extent than other learners. Germans were found to overuse the masculine form of the indefinite article (*en*) and underuse the neuter form (*et*) – whereas the situation was the reverse for English speakers. Another unexpected finding was that Russians tend to avoid the word *eller* ('or'), despite the fact that the corresponding Russian word (или) bears a strong formal resemblance to its Norwegian counterpart. These and other findings were subjected to (diagnostic) contrastive analysis (Gast 2012) in order to answer the third research question, and in most cases it proved possible to attribute the tendency in question to language transfer in one form or another.

This presentation focuses on the use of discriminant analysis and its applicability to a wide range of problems in learner corpus research.

References

Gast, Volker. 2012. Contrastive Analysis. In Michael Byram & Adelheid Hu (eds.) *The Routledge Encyclopedia of Language Teaching and Learning*, 2nd Edition. London: Routledge.

Jarvis, Scott. 2000. Methodological rigour in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning* 50:2.

—— 2010. Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook* 10.

Jarvis, Scott & Scott A. Crossley (eds.) 2012. Approaching Language Transfer through Text Classification. Explorations in the detection-based approach. Bristol: Multilingual Matters.

Klecka, William R. 1980. *Discriminant Analysis*. Quantitative Applications in the Social Sciences 19. London: Sage Publications.

Pepper. Steve. 2012 *Lexical transfer in Norwegian interlanguage – A detection-based approach.* Master's thesis in Linguistics, University of Oslo.

A corpus-based study on the distribution of complements and adjuncts in learner language: will we reveal <major findings><in this study> or will we reveal <in this study><minor findings>?

Pérez-Guerra, Javier; Martínez-Insua, Ana Elina
Universidade de Vigo
jperez@uvigo.es; minsua@uvigo.es

This study deals with the integration of syntactic complements and adjuncts in predicate projections in learner English in an attempt to determine whether the production of such constituents in English by non-native speakers is influenced or not by their first language. The analysis is based on sequences of complements and adjuncts in predicates (VPs) such as (1) and (2):

(1) *deal* [*with the construction*] [*in a somewhat strange way which will lead to odd results*]
(2) *deal* [*in a somewhat strange way which will lead to odd results*] [*with the construction*]

The study is couched in frameworks which analyse student's interlanguage (see Eubank et al. 1997 for recent definitions) and approximative linguistic systems (Nemser 1971) in general, and follows Granger's (1996) comprehensive Integrated Contrastive Model (ICM; see Granger 2002, 2009; Gilquin 2008: 6–8) to interlinguistic analysis.

Two forces are claimed to determine the order of complements and adjuncts in a given category: the 'syntactic' force of 'complements-first' (Quirk et al. 1985: 49-50; Hawkins 2007) and the 'processing' force of end-weight (Quirk et al. 1985: 1398; Hawkins' 2004 'Minimize Domains'). We will explore the effects of both forces on the distribution of adjuncts and complements in VPs in the following corpora:

– the 100,000-word learner spoken corpus of English VICOLSE, produced by Spanish University students of English (Tizón-Couto 2012),

– the native corpus LOCNEC (Centre for English Corpus Linguistics, Université catholique de Louvain), used as the English native control corpus, and

– examples retrieved from ADESSE (University of Vigo, http://adesse.uvigo.es), a 1.5-million-word database of (native) Spanish, which acts as the Spanish native control corpus. Our research here is based on a sample of 207,948 words of spoken Spanish (in Spain) from ADESSE.

VICOLSE and LOCNEC are comparable corpora since the compilation of the former has followed the design of tasks, topics and transcription conventions used in LOCNEC, inherited from the LINDSEI project. In order to control for modality issues, we will compare the results with those drawn from a comprehensive corpus of written Modern English (PPCMBE).

The findings reveal that the syntactic principle of complements-first is strong in native spoken English, in native written English and in non-native (Spanish-L1) spoken English, while it is not significant in native spoken Spanish. Regarding end-weight, we measure the number of times the second constituent is longer than the first one in the corpora, and conclude that the processing principle of end-weight is strong in complement-first and complement-last constructions in native spoken English, in non-native spoken English and in native spoken Spanish, being specially strong in complement-last constructions in native written English.

The empirical analysis of the data lead to the following concluding remarks: (i) both English and Spanish syntax significantly influence the learners' productions as far as compliance with the syntactic principle of complements-first is concerned; and (ii) observance of the processing principle of end-weight is conditioned by modality (spoken versus written) and not by the learners' interlanguage.

References

Eubank, Lynn, Larry Selinker & Michael Sharwood Smith (1995) *The current state of Interlanguage*. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle (2008) "Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation". In: Gilquin, Gaëtanelle, Szilvia Papp & María Belén Díez-Bedmard (eds.) *Linking up contrastive and learner corpus research* (pp. 3-33). Amsterdam: Rodopi.

Granger, Sylviane (1996) "From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora". In: Aijmer, Karin, Bengt Altenberg & Mats Johansson (eds.) *Languages in contrast. Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.

Granger, Sylviane (2002) "A bird's-eye view of computer learner corpus research". In: Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.) *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins.

Granger, Sylviane (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. 13-33). Amsterdam: John Benjamins.

Hawkins, John A. (2004) *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. (2007) "Performance and grammatical varation in the ordering of verb, direct objects and obliques". Delivered at DGfS, Siegen.

Nemser, William (1971) "Approximative systems of foreign language learners". *International Review of Applied Linguistics* 9/2: 115–123.

PPCMBE = Kroch, Anthony, Beatrice Santorini & Ariel Diertani (2010) *Penn Parsed Corpus of Modern British English*.

Tizón Couto, Beatriz (2013 forthcoming) *Clausal complements in native and learner spoken English. A corpus-based study with VICOLSE*. Bern: Peter Lang.

One-to-many mapping between closely related languages and its influence on second language acquisition: a corpus-based study of native and learner Finnish

Pällin Kristi; Kaivapalu Annekatrin
Tallinn University
kristi.pallin@tlu.ee; annekatrin.kaivapalu@tlu.ee

It is widely accepted that learners exhibit systematic differences based on different L1 and L2 configurations. In closely related languages, due to extensive formal and semantic similarity, the L1 influence is mainly positive. Nevertheless, different patterns in L1 and L2 may cause negative L1 influence as well. (Ringbom 2007). According to the structural-contrastive theory, cases in which one L1 phenomenon corresponds to one L2 phenomenon are the easiest to acquire. The most difficult are the L2 phenomena not found in L1, and those in which the L1 phenomenon corresponds to many different L2 counterparts (Latomaa 1993).

Although the Estonian and Finnish languages are closely related, the Estonian verb *minema* 'go' and Finnish verb *mennä* 'go' have developed in different directions. In Finnish two separate and parallel paradigms remain: *mennä: menen* 'to go: I go' and *lähteä: lähden* 'to start to go, to leave: I start to go, I leave'. In Estonian the two verbs collapse into one paradigm with complementary distribution: *minema: lähen* 'to go: I go'. These typologically interesting paradigms are challenging for Estonian learners of Finnish, causing problems both in learning and in use of the second language. Verbs *mennä* and *lähteä* thus present a unique opportunity to study the influence of a one-to-many mapping on second language acquisition.

The research questions addressed in this study are the following: 1) how and in what contexts do the Estonian learners of Finnish use *mennä*- and *lähteä*-verbs, 2) whether and how does the use of *lähteä* and *mennä* vary on proficiency levels A2-C1, 3) whether and how does the learner use of *lähteä*- and *mennä*-verbs differ from their use in the native corpus. The theoretical framework of the research is based on three dimensions of language proficiency: complexity, accuracy and fluency (Housen, Kuiken 2009) which can be analyzed by DEMfad model (Martin et al 2010). In this study, as in the DEMfad model, the number of *lähteä*- and *mennä*-verbs per 1000 tokens of running text is used as an overall measure of fluency. Accuracy is seen as the number of target-like expressions in comparison with native Finnish; *distribution* is used as a cover term for *complexity* in traditional sense as well as for *variability*. To explore the L1 influence, the unified methodological framework of Jarvis (2000: 249-261, 2010) and the Three-Phase Comparative Analysis earlier tested in translation studies (Jantunen 2004) are used. The data is selected from the Estonian subcorpus of the International Corpus of Learner Finnish (ICLFI). Learner language is compared on the native Finnish corpus and on the multi-source-language subcorpus of ICLFI where none of the source languages is dominant, so it can be seen as a source of representative data for learner Finnish in general.

This presentation focuses on comparing the use of *lähteä*- and *mennä*-verbs in the Estonian subcorpus and the native Finnish corpus. The preliminary results of the study show that in the learner subcorpus on the proficiency level A2 only the spatial meaning of the verbs is used. In terms of accuracy there is some confusion concerning morphosyntactic features at the proficiency levels A2 and B1, as well as mixing of *lähteä*- and *mennä*-verbs in those contexts where native speakers use only one of the verbs.

On the level B2, the verbs *lähteä* and *mennä* as verbs of movement are mostly replaced with other verbs describing movement, and the verbs *mennä* and *lähteä* are used in more abstract contexts. Compared to the native corpus, in the learner corpus *lähteä* is underused at all proficiency levels.

References

Housen, A. , Kuiken, F. 2009. Complexity, accuracy and fluency in Second Language Acquisition. − Applied Linguistics 30(4), 461−473.

Jantunen, J. H. 2004. Untypical patterns in translations. Issues on corpus methodology and synonymity. – A. Mauranen & P. Kujamäki (eds.) Translation Universals. Do they Exists? Amsterdam: John Benjamins, 101−126.

Jarvis, S. 2000. Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. − Language Learning 50 (2), 245−309

Jarvis, S. 2010. Comparison-based and detection-based approaches to transfer research. − L. Roberts, M. Howard, M. Ó Laoire, & D. Singleton (eds.). EUROSLA Yearbook 10 Amsterdam: Benjamins, 169−192.

Martin, M., Mustonen, S., Reiman, N. Seilonen, M. 2010. On becoming an independent user. – I. Bartning, M. Martin & I. Vedder (eds.) Communicative proficiency and linguistic development: intersections between SLA and language testing research. EUROSLA Monograph Series 1, 57 −80

Latomaa, S. 1993. Mitä hyötyä on oppijoiden kielitaustan tuntemisesta? − Aalto, E. & Suni, M. (eds.) Kohdekielena suomi. Näkökulmia opetukseen. Korkeakoulujen Kielikeskuksen selosteita 1. Jyväskylä: University of Jyväskylä, 9–31.

Ringbom, H. 2007. Cross-linguistic similarity in Foreign Language Learning. Clevedon: Multilingual Matters.

Text and annotation mining tools in the PLEC learner corpus

Pęzik, Piotr
University of Łódź
pezik@uni.lodz.pl

Introduction

The PELCRA Learner English Corpus (PLEC) is a research project funded by a grant from the Polish Ministry of Science and Higher Education (Pęzik 2012). Launched in late 2010 the project is aimed at compiling an annotated corpus with a spoken component for quantitative and qualitative analyses of Polish learner English. The corpus contains samples of learner English, such as essays, in-class and in-exam assignments, letters, MA theses and many other types written compositions authored by Poles using English as a foreign language (2.8 million words in total). It also features a time-aligned, error- annotated spoken sub-corpus of learner English (200 000 word segments, which roughly corresponds to 25 hours of continuous recordings). The spoken component comprises informal interviews conducted with learners of English representing a variety of proficiency levels and social backgrounds. It has been recently made available under a Creative Commons license.

The aim of this paper is to describe the main text and annotation mining tools developed within the project, including 1) a scalable, multi-modal learner corpus search engine, 2) a syntactic and error annotation browser, 3) a formulaic sequence extractor and 3) a phraseology analysis tool.

Availability

The tools and resources described in this paper are available online at *pelcra.pl/plec*.

Online search engine

The orthographic and linguistic tiers of the corpus annotation can be explored through a dedicated search engine supporting complex part-of-speech queries with slop factors and metadata “faceted browsing” views. The search engine was implemented using a customized version of the Apache Lucene library and it can be used in other corpus projects as it scales well with the size of the collection up to billions of segments. The engine provides multimodal access to the written and spoken data components; users can stream audio snippets for utterances matching their queries. Learner and text profiles such as proficiency levels, domains, genres and register can be used as search criteria.

Syntax and error annotation browsers

There are two types of error annotation in PLEC. Firstly, a general, learner error taxonomy was adopted for the manual annotation of errors in a selection of the corpus. Secondly, the entire spoken component of the corpus has been annotated for word mispronunciations and used to compile an index of words commonly mispronounced by Polish learners of English (Zając and Pęzik 2012). Additionally, the corpus has been syntactically parsed using the Stanford Dependency Parser (De Marneffe, MacCartney, and Manning 2006). All of these tiers can be explored through dedicated online annotation browsers presented in this paper.

Formulaic sequence extraction

The project also explores the possibility of applying a special n-skip-gram algorithm for extracting key formulaic sequences from the spoken component of the corpus and comparing them with n-skip-gram lists extracted from the spoken component of the British National Corpus. It is argued that this method of identifying formulaic sequences is an improvement over the widely-used n-gram analysis methods in that it factors in the nesting of shorter phrases within longer n-grams. Many such n-grams (with n>=1) with distributional characteristic specific to Polish learner English are identified with this method, including the clause-breaking use of *I don't know* discussed in this paper. (See http://pelcra.pl/PLEC/phrases.do for sample results).

Phraseology indexing

Large-scale analysis of native English corpora indicates that more than 50 percent of simple noun phrase usage are largely fixed combinations reproduced from memory, rather than spontaneous, compositional syntagms. To assess this aspect of use of English as a foreign language method of analyzing learners' phraseological competence is proposed which relies on the automatic identification of recurrent word combinations in the learner corpus. We explain how a BNC-based collocation tagger based on the HASK online dictionaries (see pelcra.pl/hask_en) is used to identify and annotate instances of selected types of phraseological units in the learner data. This in turn makes it possible to estimate the so-called phraseological index of learner English samples, which is a rough measure of native-like idiomaticity in non-native texts. One of the surprising outcomes of this analysis is the general observation that learner English can have both a significantly higher as well as significantly *lower* phraseological index with learners over-using a small selections of multiword linking adverbials in school and academic texts.

Acknowledgments

Research funded in the years 2010-2013 by a grant from the National Science Centre (N N104 205039).

References

De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D Manning. 2006. "Generating Typed Dependency Parses from Phrase Structure Parses." In Proceedings of LREC, 6:449–454.

Pęzik, Piotr. 2012. "Towards the PELCRA Learner English Corpus." In Corpus Data Across Languages and Disciplines, ed. Piotr Pęzik, 28:33–42. Łódź Studies in Language. Peter Lang.

Zając, Magda, and Piotr Pęzik. 2012. "Annotating pronunciation errors in the PLEC spoken learner corpus." In Proceedings ot TALC 10 Conference. Warsaw.

A corpus-based investigation of gender agreement and L1 transfer in Norwegian as a second language

Ragnhildstveit, Silje
University of Bergen, Department of Linguistic, Literary and Aesthetic studies
silje.ragnhildstveit@lle.uib.no

This poster presents some preliminary results from my corpus-based investigation of gender *agreement* and *L1 transfer* in Norwegian as a second language. The main hypothesis is that learners with a native language (L1) *without* agreement (Vietnamese) will have greater problems with gender agreement than learners with an L1 *with* agreement (German, Spanish and English). To uncover transfer, I use Jarvis' (2000) method and compare the performance of learners with L1s that differ with respect to the category of grammatical gender and the presence of agreement, see Table 1.

	Target language	**Native language**			
Type of language	**Synthetic**			**Synthetic/Analytic**	**Analytic**
Language	**Norwegian**	**German**	**Spanish**	**English**	**Vietnamese**
Classification of nouns	X	X	X	–	X
Gender	X	X	X	–	–
Agreement	X	X	X	X	–
M – F – N	X	X		–	–

Table 1 Target language and native languages

According to Jarvis (2000:253), transfer is at play if the following criteria can be shown to be satisfied: "inter-L1-group heterogeneity in learners' IL performance", "intra-L1-group homogeneity in learners' IL performance" and "intra-L1-group congruity between learners' L1 and IL performance".

My investigation is corpus-based, using *ASK – Norsk andrespråkskorpus* (Norwegian Second Language Corpus). The corpus consists of texts written by adult second language learners of Norwegian sitting for two different language tests, one *intermediate level* test (approximately at B1 according to the *Common European Framework of Reference for Languages* (CEFR)) and one *advanced level* test (approximately at B2 according to the CEFR). My investigation is based on texts from the intermediate level test, where there are 100 texts from each L1 group.

According to Hockett (1958:231), grammatical gender is "classes of nouns reflected in the behavior of associated words". There are two aspects of gender in this definition. First, it refers to gender as one type of *nominal classification*. Second, it refers to *agreement*. In a linguistic perspective, gender assignment, which reflects the aspect of nominal classification, is viewed as more or less predictable in gender languages across the world, and governed by semantic and formal gender assignment rules (Corbett 1991). The assignment of gender to nouns in Norwegian is *mostly unpredictable*, and hence viewed as difficult for L2 learners. However, there are *some* semantic and formal rules for gender assignment. There is a connection between biological sex and masculine and feminine gender assignment. For instance, *mann* ('man'), *gutt* ('boy'), *bror* ('brother') are assigned masculine gender and *kvinne* ('women'), *jente* ('girl'), *søster* ('sister') are assigned feminine gender. The formal rules for gender assignment are mainly connected to the derivational morphology of the noun. For instance, verbal nouns with the ending *-sjon*, such as *gratulasjon* ('congratulation'), are assigned masculine gender. Grammatical gender is reflected in adjectives and determiners that agree in gender, for instance in noun phrases like in example (1) and (2):

(1) *en* *fin* *bil*
a.MASC *nice.MASC* *car(MASC)*

(2) *et* *fint* *hus*
a.NEUT *nice.NEUT* *house(NEUT)*

Second language learners of a gender language have to assign gender to each noun in the target language, and they further have to display the gender agreement in syntactic contexts that require it. From a theoretical point of view there are two factors that make certain grammatical categories difficult to acquire, lack of transparency between form and meaning and redundancy (DeKeyser 2005:3). Grammatical gender fulfils both of these factors, because of little transparency between the gender and the semantics of the noun, and the lack of an obvious communicative function. Hence, the acquisition of grammatical gender is considered as notoriously difficult for second language learners.

In a previous corpus-based investigation (Ragnhildstveit 2009), using the same data material and Jarvis' method as described earlier, I investigated L1 transfer and gender *assignment*. I did this by comparing correct and incorrect gender assignment between groups with L1s that are different with respect to their gender systems, see Table 2.

Target language	Native language				
Norwegian	German	Spanish	Dutch	Vietnamese	English
Masculine Feminine Neuter	Masculine Feminine Neuter	Masculine Feminine	Common gender Neuter	- Classifier language	-

Table 2 Target language and native languages

Despite the presumed difficulty when it comes to acquisition of gender in a second language, all language groups had a high degree of correct gender assignment. Further, I found that the learners seem to use some of the assignment rules for gender. A very interesting finding was that the Vietnamese group had more correct gender assignment than the other groups, in spite of the fact that Vietnamese does not have grammatical gender. I proposed several possible explanations for this, for instance L1 transfer and different learning strategies.

This result led me to hypothesize that even if the Vietnamese learners have more correct gender assignment than the other L1 groups in Ragnhildstveit (2009), they may not manage to get the gender agreement correct, since their L1 does not have agreement. Maybe Sabourin et al. are right when they say that:

> It is possible that gender assignment can be done based on more general cognitive skills (simply learning or memorizing what gender goes with what item) and thus can, with enough experience, be learned by any L2 learner. In contrast, gender agreement, relying on more linguistic strategies, can only be learned in the L2 if the same strategies are present in the L1. (Sabourin et al. 2006: 27)

This poster will show results from my exploration of this hypothesis, based on a limited investigation of the type of noun phrase in (1) and (2) above. The agreement in example (2) can in principle be displayed in four different ways, (A), (B), (C) and (D), by the learners.

(A) *et* *fint* *hus*
a.NEUT *nice.NEUT* *house(NEUT)*

(B) *et* *fin* *hus*
a.NEUT *nice.MASC* *house(NEUT)*

(C) *en* *fint* *hus*
a.MASC *nice.NEUT* *house(NEUT)*

(D) *en* *fin* *hus*
a.MASC *nice.MASC* *house(NEUT)*

The first question is whether the learners display agreement between the indefinite article and the adjective, like in (A) and (D), or if they display “disagreement”, like in (B) and (C). The second question is whether the Vietnamese learners exhibit more disagreement than the other learners.

References

Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.

DeKeyser, Robert. 2005. What Makes Learning Second ‐ Language Grammar Difficult? A Review of Issues. *Languag Learning* 55 (0):1-25.

Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: The Macmillan Company.

Jarvis, Scott. 2000. Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. *Language Learning* 50 (2):245-309.

Ragnhildstveit, Silje. 2009. Genustildeling og morsmålstransfer i norsk mellomspråk. En korpusbasert studie, Institutt for lingvistiske, litterære og estetiske studier, Universitetet i Bergen, Bergen.

A Contrastive Interlanguage Analysis of motion events

Reshöft, Nina

University of Bremen and University of Paderborn (Germany)

n.reshoeft@uni-bremen.de; reshoeft@campus.upb.de

This study looks at the expression of motion events in typologically different languages and at transfer effects resulting from typologically similar and different L1s and L2s. It shows how German and Spanish learners of English express motion events in their English L2.

Research on motion events has shown that speakers of Germanic and Romance languages differ in the ways in which path and manner of motion are expressed (cf. Talmy 1991). Talmy distinguishes two language types, *verb-framed* (e.g., Romance) and *satellite-framed* (e.g., Germanic) languages on the basis of different lexicalization patterns used for the expression of semantic components associated with motion events. According to Talmy's typology, path and manner of motion are expressed in different syntactic elements (*verb* and *satellite*) in these two language types.

Written narratives were collected from adult native speakers of German, English and Spanish in their respective L1s. This data served as a reference corpus. In addition, two learner corpora containing narratives written by German and Spanish learners of English were compiled. Similar to Cadierno (2004), the task was based on a wordless picture story book that elicited different types of spatial descriptions.

Based on the results presented in previous research on motion events, the expectation was to find the dominant patterns of the learners' L1s (Spanish and German) to be reflected in their (English) interlanguage. More specifically, it was hypothesized that speakers of German and Spanish pay different attention to spatial and non-spatial dimensions of motion in their written L2 production.

The analysis compared the semantic categories that were expressed by learners of English from typologically different L1s, German and Spanish. Both types of *Contrastive Interlanguage Analysis* were involved, i.e., a comparison of native and non-native varieties of English and a comparison of different interlanguages of English (Granger 1996). As for the first type, the English L1 data were compared to the German L2 data and to the Spanish L2 data. Second, the German L2 data were compared to the Spanish L2 data.

The semantic elements of all motion expressions in the narratives were coded for a range of concepts and relations that are relevant for the ways in which motion events are lexicalized in the two language types. Different aspects of spatial relations were analyzed, as well as the expression of non-spatial elements, many of which have been traditionally analyzed as manner of motion. The annotation of spatial concepts and relations was based on the linguistically-motivated ontology of the Generalized Upper Model spatial extension (*GUM-Space*, Bateman et al. 2010). GUM-Space describes the semantics of spatial terms and the relation between the concepts underlying linguistic expressions of space. The corpus data was analyzed with regard to semantic elements and the syntactic elements by which they are expressed.

Results from the reference and learner corpora support the expectation that the L2 patterns are influenced by the respective L1s. The results further indicate interesting differences between certain subtypes of what have been traditionally subsumed under the notions of manner and path. The results are discussed with regard to *conceptual transfer* (Jarvis & Pavlenko 2008).

References

Bateman, John., Hois, Joana, Ross, Robert & Tenbrink, Thora (2010) A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14): 1027-1071.

Cadierno, Teresa (2004) Expressing motion events in a second language: A cognitive typological approach. In: Achard, Michel & Niemeier, Susanne (eds.) *Cognitive linguistics, second language acquisition and foreign language pedagogy* (pp. 13-49). Berlin: Mouton de Gruyter.

Granger, Sylviane (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner. In: Aijmer, Karin, Altenberg, Bengt & Johansson, Mats (eds.) *Languages in Contrast. Text-based cross-linguistic studies* (pp. 37–51). Lund Studies in English 88. Lund: Lund University Press.

Jarvis, Scott & Pavlenko, Aneta (2008) *Crosslinguistic influence in language and cognition*. New York: Routledge.

Slobin, Dan Isaac (2004) The many ways to search for a frog: linguistic typology & the expression of motion events. In Strömqvist, Sven & Verhoeven, Ludo (eds.) *Relating Events in Narrative. Vol. 2* (pp. 219-257). Mahwah, NJ: Lawrence Erlbaum Associates.

Talmy, Leonard (1991) Path to Realization: A Typology of Event Conflation. *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society* (pp. 480-519). Berkeley Linguistics Society, University of California, Berkeley.

Zlatev, Jordan & Yangklang, Peerapat (2004) A third way to travel: The place of Thai in motion event typology. In Strömqvist, Sven & Verhoeven, Ludo (eds.) *Relating Events in Narrative. Vol. 2* (pp. 159-190). Mahwah, NJ: Lawrence Erlbaum Associates.

Compilation, Analysis and Findings of Yonsei English Learner Corpus (YELC)

Rhee, Seok-Chae; Jung, Chae Kwan
Yonsei University; Korea Institute for Curriculum and Evaluation
scrhee@yonsei.ac.kr; ckjung@kice.re.kr

In recent years, researchers have become increasingly interested in the creation and pedagogical use of English learner corpora. Many studies (e.g. Tono et al., 2001; Granger, 1994, 2003; Biber, 2006) have shown that learner corpora not only create significant contributions to second language acquisition research, but also contribute to the construction and evaluation of language tests by advancing our understanding of English learners.

However, little attention has been paid to the Korean EFL (English as a Foreign Language) learners' corpus so far. The Yonsei English Learner Corpus (YELC), funded by the Brain-Korea (BK) 21, is a Korean EFL learner corpus compiled by Yonsei University from 2011 to 2012. Over, 3000 Korean male and female high school graduates who were accepted to Yonsei University for further studies participated in this project.

YELC consists of 6,572 authentic student writing examples at nine different English proficiency levels (A1, A1+, A2, B1, B1+, B2, B2+, C1, C2) which we have refined and extended the original six proficiency levels (A1, A2, B1, B2, C1, C2) of the Common European Framework of Reference for Languages (CEFR), providing a total of 1,099,427 words.

(1) *A level*
  a. *A1: 82 texts*
  b. *A1+: 370 texts*
  c. *A2: 1,368 texts*
(2) *B level*
  a. *B1: 2,346 texts*
  b. *B1+: 1,410 texts*
  c. *B2: 756 texts*
  d. *B2: 162 texts*
(3) *C level*
  a. *C1: 74 texts*
  b. *C2: 4 texts*

In this paper, we will describe the compilation process of how we have 'corpusized' from the text data to a sound corpus. Once the process of 'corpusization' is introduced, we will report interesting linguistic features that different proficiency levels of Korean learners of English show. This study will also discuss the potential use of the YELC which is not freely available for research purposes (http://www.uclouvain.be/en-cecl-lcworld.html).

References

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

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today*, 39: 25-29.

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: 538-546.

Tono, Y., Kaneko, T., Isahara, H., Izumi, E., Saiga, T., Kaneko, E. (2001). Developing a one million word spoken EFL learner corpus. *Proceedings of the JALT*, Kitakyushu, Japan, pp. 871-880.

Yonsei English Corpus Lab
http://web.yonsei.ac.kr/yonseicorpuslab

How to get YELC:
http://web.yonsei.ac.kr/yonseicorpuslab/신청서양식 46.htm

Spanish EFL university discourse competence: A longitudinal study of EAP development

Juan Pedro Rica-Peromingo and JoAnne Neff-van Aertselaer
Universidad Complutense de Madrid
juanpe@filol.ucm.es / neffjoanne@hotmail.com

The purpose of this longitudinal study is to compare the use of certain devices for discourse competence by two different one-semester second year classes of Spanish EFL students of English Studies (2012-2013) at the Complutense University, Madrid (CEFR levels B1, B2 and C1).The corpora analyzed are:

1) written data from initial and final essays collected from 10 students at differing levels, as measured by the *Oxford Quick Placement test* administered at the beginning of the academic year; and
2) written data collected, initially and finally, in the 2nd semester from the same students, currently enrolled in B2+ (although linguistic competence is not necessarily at this level).

The broader research purpose, completed in four different steps, was to test the effectiveness of adaptations made to the descriptors outlined in the *Common European Framework of Reference for Languages* (CEFR, Council of Europe 2001) for curricular guidelines. The selection of students at different levels was meant to ascertain if a certain threshold level exists below which EFL students are unable to integrate the devices described below into their writing.

As a result of both the publication of the CEFR guidelines for linguistic competence and the Bolognia process for the reform of higher education (Keeling, 2006) –advocating curricular studies that focus on productive activities that are purposeful, lead towards some explicit objectives, and are measurable– specific guidelines for academic writing needed to be developed and explicit training for structural and rhetorical features provided for Complutense EFL students, now obliged to write an end-of-studies paper as a graduation requirement.

The CEFR descriptors are under-defined (Hawkins & Buttery 2010) since their use may be applied to various different groups of learners (immigrant adult learners, secondary school learners, etc.). These descriptors are helpful, but are too general to provide for task-specific competencies on which students could be assessed. Some development of CEFR writing guidelines has emerged in Green (2012) who briefly lists some lexico-grammatical exponents as “text characteristics”; yet, still missing in his account is a middle category linking discourse cues (especially lexical phrases) to the construction of an argumentation schema (Andrews 2010). In addition, as Fleming (2009) has pointed out, broad descriptors are not necessarily the answer to specific teaching objectives, in this case, identifying specific structural and rhetorical features.

As a first step in curricular change, the Complutense Writing Group created an initial framework (Neff 2013), as an intermediate scaffold for student use in academic reading and writing. It included two groupings of features: 1) structural (aspects like encapsulation and prospection, claims and supporting data); and, 2) rhetorical (aspects such as reporting verb use for alignment or non-alignment with the sources used, the use of modal verbs, etc).

The second phase identified specific devices used by expert academic writers for inclusion into teaching units. Following taxonomies used by Hyland (2000), Biber et al. (1999). Paquot (2010) and Rica (2007), discourse devices included in the adapted CEFR descriptors were searched for: verbs of mental processes and of communication(e.g., reporting verbs), linking phraseological units, items expressing doubt and certainty (e.g., epistemic and deontic verbs and adverbials), and person markers (evidence of impersonalization or the lack of). Hyland (2000), Springer (2012) and others have shown that academic competence depends on their use or absence and thus, they form part of the specific CEFR-adapted descriptors.

A third step involved a wide variety of pedagogical tasks targeting specific areas for substantial improvement in EFL texts. Since course time is short and over-all language acquisition is rather imperceptible, rather than relying on commercial "academic textbooks", pedagogical tasks were developed: 1) for reading, developing discourse analysis skills; 2) the construction of writing tasks specific to Spanish EFL university students.

The fourth step is to detect the extent to which the Complutense students have actually been able to use these features, in other words, to what extent was the curricular design successful. The 1st-semester findings show that students have incorporated features such as a variety of reporting verbs, more sophisticated discourse markers, fewer personal pronouns ("I"/"we") and more generic "you" as subjects. However, the final 1st-semester compositions did not seem to show improvement in the following: the use of epistemic verbal and adverbial items and an overuse of deontic lexical phrases. These results are presently being fed into the 2nd-semester pedagogical activities in an attempt to revise the curricular design accordingly. The structural and rhetorical features included in our framework, plus the specific pedagogical activities, suggest that students can only learn more complex discourse features if they are at a certain threshold level.

References

Andrews, R. (2010). *Argumentation in Higher Education.* London: Routledge.

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

Council of Europe. (2001). *Common European Framework of Reference for Languages*: *Learning, Teaching and Assessment*. Cambridge: Cambridge University Press. (also accessable at: http://culture.coe.int/portfolio >Documentation > Common European>Framework of Reference for Languages).

Fleming, M. (2009). The use of descriptors in learning, teaching and assessment. Council of Europe, <www.coe.int/lang>, (accessed January 2013).

Green, A. (2012). *Language Functions Revisited.* Cambridge: Cambridge University Press.

Hawkins, J. and P. Buttery. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, Vol 1: 1-23.

Hyland, K. (2000). *Disciplinary discourses. Social interaction in academic writing.* London/New York: Longman.

Keeling, R. 2006. The Bologna Process and the Lisbon research agenda: the European Commission's expanding role in higher education discourse. *European Journal of Education,* Vol. 41, (2): 203-223.

Neff, J. (in press). Contextualizing EFL argumentation writing practices within the *Common European Framework* descriptors. *Journal of Second Language Writing.*

Paquot,M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.

Rica, J.P. (2007). *Estudio fraseológico del uso de colocaciones gramaticales y grupos léxicos en textos argumentativos nativos y no nativos: análisis de corpus de estudiantes.* Unpublished PhD Thesis. English Department I, Universidad Complutense de Madrid.

Springer, P. (2012). *Advanced Learner Writing.* A corpus-based study of the discourse competence of Dutch writers of English in the light of the C1/C2 levels of the CEFR. Oisterwijk: Philip Ernest Springer.

The use of learner corpus as an aid to teach and learn Italian collocations

Sabino, Marilei Amadeu
UNESP – Universidade Estadual Paulista – 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), 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.*

Financial support: FAPESP (Proc. 2013/14233-7).

Using an automatic parser as a language learner model

Schneider, Gerold; Gintarė Grigonytė
University of Zurich; University of Stockholm
gschneid@es.uzh.ch; gintare@ling.su.se

Native speakers rarely completely analyze sentences into their individual components, instead we perceive and produce semi-preconstructed phrases (Sinclair 1991, Stefanowitsch and Gries 2003). Lexical expectations (Hoey 2005) guide our interpretation, creative and analytic use of language is very restricted. When native speakers construct sentences they employ argument structures, alternations (Levin 1993), choice of synonyms and register as subtle operations (Pawley and Syder 1983). Although grammatical variation seems abundant (e.g. Rohdenburg & Mondorf 2003) it is severely restricted by complex, and interacting factors up to being nearly deterministic (Bresnan et al. 2007). Sentences are rendered in the way that they are due to many complex and interacting factors, and even subtle failures increase both the human and the automatic processing load.

Language learners sometimes produce syntactically incorrect sentences, but more often they fail to use these subtle factors as successfully as native speakers do. For example (1) and shows a lexical preference error.

(1a) Original: *Usually , I go to the library , and I rent these books.*
(1b) Corrected: *Usually , I go to the library , and I borrow these books.*

These failures can lead to increased processing times for the human listener, and possibly even ambiguities or misunderstandings, as in (2) and (3).

(2a) Original: *I am going to the present for my family.*
(2b) Corrected: *I am going to buy presents for my family.*
(3a) Original: *Kindly and gently computer game I bought for them.*
(3b) Corrected: *I bought a harmless computer game for them.*

We use an automatic robust probabilistic parser (Schneider 2008) as psycholinguistic model of syntactic and idiomatic expectation. A broad-coverage parser can serve as a psycholinguistic language model, because it predicts attachment decisions based on grammar rules and lexical preferences, because its statistical model can be extended by semantic and discourse-level factors, and it learns form real-word data. Entrenched structures get higher scores, as they are expected.

Keller (2010) suggests the use of broad-coverage robust parsers as cognitively plausible models. We apply such an approach to Learner English and show that parser performance (Figure 1) and parser scores are significantly lower for Learner data than for corrected. We have manually annotated 100 sentence pairs from the NICT Japanese Learner English (JLE) Corpus[1]. It contains 120,000 sentence pairs of consisting of an original language learner sentence and a corrected sentence (see (1)-(3)).

[1] http://alaginrc.nict.go.jp/nict_jle/index_E.html

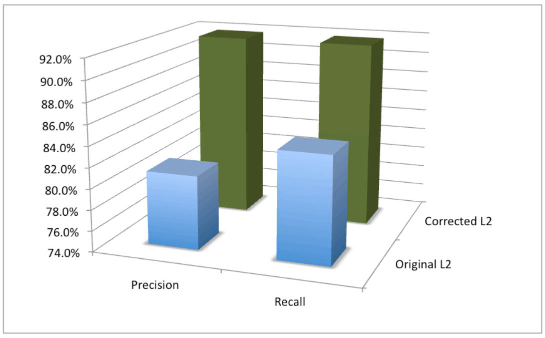


Figure 1. Parser Performance

Our hypothesis is that L2 utterances do not fit the model very well – equally the human listener and the computational parser model – and thus lead to lower parser scores, in correlation to increased processing times for human listeners. We compared parser scores between original and corrected sentences (Figure 2), but also the fragmentation of the parser output (Figure 3).

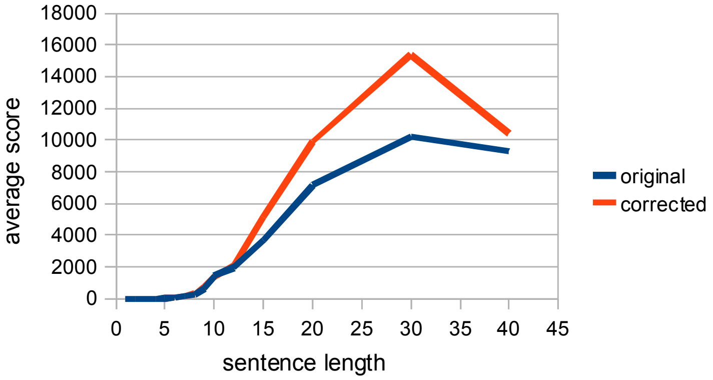


Figure 2. Parser score by sentence length, comparing original and corrected utterances

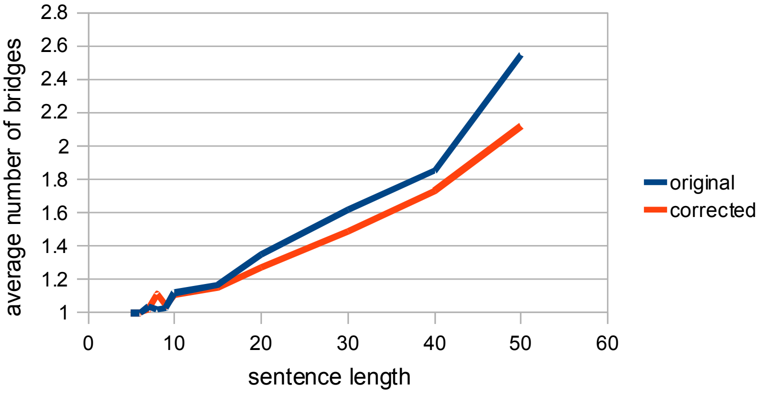


Figure 3. Parser fragmentation levels

There also is a correlation between the competence level of language learners and parser scores. Figure 4 plots parser scores for our parses of the CEEAUS written corpus (Ishikawa 2009).

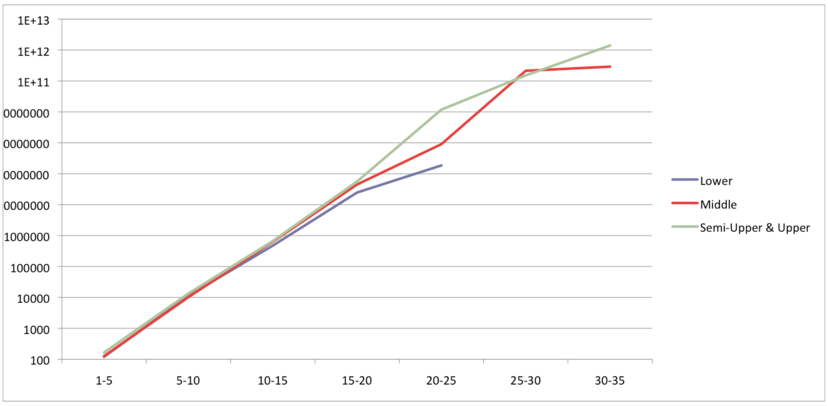


Figure 4. Parser scores compared to language learner competence level

We suggest the use of syntactic parsers as psycholinguistic tools for the analysis of Learner data and step towards a model-based approach to entrenchment.

References

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, Cognitive Foundations of Interpretation. Royal Netherlands Academy of Science, Amsterdam, pages 69–94.

Hoey, Michael. (2005). Lexical priming: A New Theory of Words and Language. Routledge.

Ishikawa, Shin. (2009). Vocabulary in interlanguage: A study on corpus of English essays written by Asian university students (CEEAUS). In K. Yagi and T. Kanzaki, (eds): *Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan*, pages 87–100, Nishinomiya, Japan. Kwansei Gakuin University Press.

Keller, Frank. (2010). Cognitively Plausible Models of Human Language Processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, 60-67. Uppsala.

Levin, Beth, (1993). English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago Press.

Pawley, Andrew and Syder, Frances Hodgetts. (1983). Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency. In Richards, J. C. & Schmidt, R. W. (Eds.), Language and Communication. London: Longman. 191–226.

Rohdenburg, Guenter & Mondorf, Britta, eds. (2003). Determinants of Grammatical Variation in English, Mouton de Gruyter, Topics in English Linguistics 43.

Schneider, Gerold. (2008). Hybrid Long-Distance Functional Dependency Parsing. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Sinclair, John, (1991). Corpus, concordance, collocation: Describing English language. Oxford: OUP.

Stefanowitsch, Anatol & Gries, Stefan Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 209–43.

Two new sub-corpora of Estonian Interlanguage Corpus (EIC)

Shmeleva Alisa
Tallinn University
lisochka@tlu.ee

The Estonian Interlanguage Corpus (EIC) of Tallinn University is collection of written texts produced by learners of Estonian as second language and learners of Russian as a third language and native speakers of Russian. EIC consisted of seven different sub-corpora: 1. Competition works in Estonian (as L2); 2. Academic learner language (Estonian as L2); 3. Estonian academic language (Estonian as mother tongue); 4. Estonian Interlanguage core corpus (as L2); 5. The collection of the texts of State Examination and Qualification Center (Estonian as L2), 6. The corpus of Russian language as the third language and 7. The corpus of Russian language as mother tongue.

There are two new sub-corpora of Estonian Interlanguage Corpus (EIC) under consideration in this paper: *The corpus of Russian language as the mother tongue* (L1) and *The corpus of Russian language as the third language* (L3). The paper focuses on the compilation of these sub-corpora.

The main purpose which is linked to the compilation of the corpus of Russian language as mother tongue (L1) and the corpus of Russian language as L3 is to compare the use of Russian as L1 with the acquisition of Russian as L3 in Estonia and to compare it with standard language using Russian National Corpus.

Also on the basis of these sub-corpora it is possible to compare the use and acquisition of Russian language as the mother tongue (L1) by students in secondary education with the acquisition of Estonian as the second language (L2) in order to determine to which extent the usage and acquisition of L1 and L2 by students with Russian as their mother tongue are different or similar (e.g. lexical richness and morphosyntactic complexity).

Students in secondary education with Russian as their mother tongue acquire Estonian as the second language, because it is official language in Estonia and English or any other language as third (L3). Students with Estonian as their mother tongue acquire English as second (L2) language and Russian as the third language.

The general aim of research is to determine what kind of common features and differences are appearing in the use of language and its acquisition.

The sub-corpus of Russian language as the mother tongue and *the sub-corpus of Russian language as the third language* are monitor, synchronic corpora, which size is constantly growing. At this time the *sub-corpus of Russian as L1* consists of examination papers, home and class essays by students in secondary education and contains 371 texts (approximately 200 000 words). The *sub-corpus of Russian as L3* includes examination papers and letters by students in secondary education and contains 279 texts (approximately 60 000 words).

Learners' texts (Russian as L1) were written by hand in the classroom, at home, and on the examination and texts (Russian as L3) were written on the examination. All of them were digitized (typed) before any further processing. L1 texts were collected from schools in Estonia (e.g. Narva city in North-East region of Estonia) and from the State Examination and Qualification Center, L3 texts were collected from the State Examination and Qualification Center, where are kept all examination papers by students. There is metadata about students, whom texts were collected from schools in Narva city, for instance age, sex, mother tongue, language and social background, etc. and there is no metadata about students, who wrote the papers on the examination, because it is confidential data (every paper has its own code).

In this paper will be presented principles of the compilation of these sub-corpora (design and nature, data collection, corpora size and their representativeness, preparation of the texts for inclusion them into the corpus) and the different opportunities to use them according to the research question or purpose and also given an overview of done work in connection with these sub-corpora: their current state.

References:

Biber, Douglas (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.

Kennedy, Graeme (1998) *An introduction to corpus linguistics*. New York: Longman.

Aijmer, Karin & Altenberg, Bengt (1991) (Eds.) *English corpus linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

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

Atkins, Sue, Clear, Jeremy & Ostler, Nicholas. (1992) *Corpus Design Criteria*. Literary and Linguistic Computing 7(1): 1 – 16.

Stefanowitsch, Anatol (2003) *Corpus Compilation*. Corpus linguistics. http://www-user.uni-bremen.de/~anatol/docs/corp_compilation.pdf, 07.02.2013.

L1 influence in the development of learner Finnish: Comparing two learner corpora

Siivelt, Keaty; Mustonen, Sanna
Tallinn University; University of Jyväskylä
ksiivelt@tlu.ee; sanna.s.mustonen@jyu.fi

The traditional paradigm of transfer studies has been limited primarily to the identification and quantification of the phenomenon, while the current research investigates also the causes and limitations of the phenomenon and seeks for theoretical explanations. International empirical transfer research focuses increasingly on the positive transfer based on crosslinguisticsimilarity between the languages one already knows (L1, L3) and the language one is studying (L2) (Jarvis 2010).

In language learning and particularly in learning a closely related language, similarity across the languages is an important factor. Learners try first to find something similar, something they already know (Ringbom 2007: 1). In closely related languages the potential for similarities is higher than in of non-related languages. This corpus-based study investigates the development of the use of local cases in learner Finnish, focusing on Estonian learners of Finnish and especially on identifying the effect of positive transfer.

In Finnish and Estonian there are 14 - 15 cases, eight of which form a subsystem called local cases. The local case system is based on oppositions of directionality and quality and act like prepositions in Indo-European languages: TO in Finnish and Estonian the illative, allative and translative case, IN/ON/AT the inessive, adessive and essive case and FROM the elative and ablative case (Martin et al 2010: 64).

Written data is collected from the L1 Estonian subcorpus of International Corpus of Learner Finnish - ICLFI, 86 178 tokens in total. To identify the effect of positive transfer ICLFI data is compared to the results of Sanna Mustonen's study (Martin et al 2010: 64–67) on the development of the use of Finnish local cases in the Cefling Project. Comparison data consists of writing samples from adults taking the Finnish National Proficiency Certificate exams including more than 20 different L1 backgrounds, 53 019 tokens in total. All texts from both corpora have been independently rated to be at a given CEFR level. To track the development of local cases DEMfad Model is used (Franceschina et al 2006). Language proficiency is commonly described as developing in three dimensions: complexity, accuracy, and fluency (the CAF-triad). In this study three dimensions, similar to CAF-triad, are being analysed: frequency, accuracy and distribution (*fad* in DEMfad). Distribution is described both as the usage of concrete phrases compared to abstract ones, and also as lexical diversity. Because of the small size of corpora, for statistical analysis the log-likelihood test is used. WordSmith Tools 5.0 is used for qualitative analysis.

Presentation will focus on following research questions:

How do the frequency, accuracy and distribution develop on different levels
of language proficiency?

Does the effect of transfer change qualitatively and quantitatively on different levels
of language proficiency? How?

Does the positive transfer lead to faster development of Finnish local cases used by L1 Estonian learners of , measured as frequency, accuracy and distribution, compared to learners of other L1 backgrounds?

Preliminary results show that local cases emerge in learner language from early on regardless of their L1 background. The use of concrete expressions decline and the use of abstract expressions rise when the proficiency level rises. Text types could be a reason for this: in ICLFI corpus fictive texts change to argumentative at the level B1, as described in CEFR (2003). It could be assumed that argumentative texts lead learners to use local cases more in abstract meanings than fictive.

Accuracy develops as expected in both corpora: as proficiency level rises the learners become more accurate. Data shows that concrete phrases are harder to produce by the means of accuracy than those with abstract meanings.

Results show that despite the differences in the use of local cases in Estonian and Finnish, learners with closely related L1 acquire native-like level of frequency and accuracy at earlier stages than those of other language backgrounds. The development of accuracy shows that L1 Estonian learners achieve at least 89% accuracy in their use of local cases already at the level A2. The underlying objective similarities make it possible for learners to assume and perceive similarity across these languages. This is especially important in the early stages of learning.

References

Franceschina, Florencia, Riikka Alanen, Ari Huhta & Maisa Martin. (2006) A progress report on the Cefling project. Paper presented at SLATE Workshop in Amsterdam. https://www.jyu.fi/hum/laitokset/kielet/cefling/en/pub/alanenetal06 (08.06.2011).

Jarvis, Scott. (2010) Comparison-based and detection-based approaches to transfer research. In: Roberts, Leah, Martin Howard, Muiris Ó Laoire & David Singleton (eds.) *EUROSLA Yearbook 10* (pp. 169−192). Amsterdam: Benjamins.

Martin, Maisa, Sanna Mustonen, Nina Reiman & Marja Seilonen. (2010) On Becoming an Independent User. In: Barting, Inge, Maisa Matin & Ineke Vedder (eds.) *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp. 57−81). European Second Language Association. Eurosla Monographs Series 1.

Ringbom, Håkan. (2007) *Cross-linguistic Similarity in Foreign Language Learning.* Clevedon: Multilingual Matters LTD.

ICLFI = http://www.oulu.fi/hutk/sutvi/oppijankieli/

Cefling = Cefling - Linguistic Basis of the Common European Framework for L2 English and L2 Finnish, project pages https://www.jyu.fi/hum/laitokset/kielet/cefling/en

Learner corpora as a pedagogical resource in specialized translator training

Ingrid Simonnæs and Sunniva Whittaker
NHH Norwegian School of Economics
Ingrid.Simonnas@nhh.no; Sunniva.Whittaker@nhh.no

Our paper aims to describe how TK-NHH, a multilingual corpus of texts translated by candidates sitting for the Norwegian national translator accreditation exam, can be used as a pedagogical tool for courses in specialized translation. TK-NHH contains texts produced by both successful and unsuccessful candidates and can as such be considered as a learner corpus. The use of corpus based studies is well established in translator training and translation research (e.g. Baker 2004; Monzo 2008; Biel 2010; Laviosa 2010). However, the TK-NHH is unique in its kind in terms of language combinations as well as the specialized domains covered, viz. economics, legal and technical domains and the production context. TK-NHH is a dynamic corpus and contains target texts in four languages (English, French, German and Spanish) of the same Norwegian source texts used for the accreditation exam from 2007 to 2012 (as of February 2013). The number of target texts varies according to the number of candidates sitting for the exam into each of the four languages in a given year, up to 33 target texts of the same source text. The texts are aligned (Hofland & Johansson 1998) and can be queried using the Corpus Work Bench (CWB) developed by IMS (Institut für Maschinelle Sprachverarbeitung.

Our hypothesis is that variation in the translation of specialized terms may be a good indicator of the degree of difficulty posed by the term in question, contrary to what is the case in texts produced by domain experts (cf. Freixa 2006). Terms for which there are multiple translations in the corpus can be easily identified and classified using the CWB. Given the rapid increase in specialized terminology, both novice and expert translators are likely to come across unfamiliar terms that have not yet found their way into bilingual dictionaries. We are interested in exploring the coping strategies used by translators in an exam context where they have no internet access. We will focus on both culture bound terms for which there are no ready equivalents in the target languages, and culture neutral terms that may have a corresponding term in the target languages, but where expert knowledge is required to identify it. A typical example of the former is the Norwegian legal term *medmor* (female partner of a woman who has become pregnant through assisted reproductive technology (ART), the former assuming parental responsibility in lieu of the biological father) translated by the candidates into *joint mother, co-mother, joint status as mother*; *mère associée, co-mère* and *Teilmutter, Mitmutter* respectively in the languages under investigation. A typical example of the latter is *kjernekapital* (*core capital* or *tier 1 capital*). This term occurs in the Norwegian transposition of an EU directive. Corresponding terms therefore exist in the official languages of all EU and EEA member states.

Both in the case of culture bound and culture neutral terms, in depth knowledge of the conceptual system the terms belong to is the key to a felicitous translation. The various renderings found in the corpus not only give us insight into the cognitive processes used by the candidates, but can also be used to enhance learners' translation quality assessment skills.

References

Baker, Mona (2004). The treatment of variation in corpus-based translation studies. In: Aijmer, Karin & Hasselgård, Hilde (eds). *Translation and Corpora. Selected Papers from the Göteborg-Oslo Symposium 18-19 October 2003* (pp. 7-15). Göteborg: Acta Universitatis Gothoburgensis.

Biel, Lucja (2010). Corpus-based Studies of Legal Language for Translation Purposes: Methodological and Practical Potential. In: Heine, Carmen & Engberg, Jan (eds). *Online proceedings from the XVII European LSP Symposium 2009* (pp. 1-15).
URL: http://www.asb.dk/fileadmin/www.asb.dk/isek/biel.pdf.

Freixa, Judit (2006). Causes of denominative variation in terminology. A typological proposal. In: *Terminology* 12 (1): 51-77.

Hofland, Knut & Johansson, Stig (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In: Johansson, Stig & Oksefjell, Signe (eds*). Corpora and Cross-linguistic Research. Theory, Method, and Case Studies* (pp. 87-100). Amsterdam: Rodopi.

Laviosa, Sara (2010): Corpus-based translation studies 15 years on: Theory, findings, applications. In: *SYNAPS* 24/2010: 3-12.

Monzo, Esther Nebot (2008). Corpus-based Activities in Legal Translator Training. In: *The Interpreter and Translator Trainer* 2 (2): 221-252.

From stumbling blocks to stepping stones: The use of the Finnish partitive case in learners from related and non-related L1 backgrounds

Spoelman, Marianne
University of Oulu
marianne.spoelman@oulu.fi

Finnish belongs to the Finno-Ugric language family and is particularly well known for its rich and complex morphology. Consisting of fifteen cases, the Finnish case system also comprises a partitive case, which is a typical case characterizing Finnic languages. In Finnish, the partitive is primarily used as one of the cases of the object, the existential subject and the predicative. More specifically, partitive case-marked noun phrases represent that side of the object, subject and predicative case alternations expressing aspectual unboundedness, negative polarity and quantitative unboundedness. Since these factors all affect Finnish object case-marking, partitive objects are licensed in aspectually unbounded and/or negated sentences, and in the case of objects denoting quantitatively unbounded entities (cf. for an example sentence in which all of these three factors are present). In contrast, the case-marking of the existential subject is merely affected by polarity and quantitative boundedness and predicative case-marking by quantitative boundedness only (cf. (2)-(5) for examples illustrating the nominative-partitive subject and predicative case alternations).

(1) *Hän ei rakasta lunta-Part.Sg ~ koiria-Part.Pl*
'(S)he does not like snow ~ dogs'

(2) *Pöydällä on ~ ei ole leipää-Part.Sg*
'There is (some) bread ~ no bread on the table'
(3) *Pöydällä on leipä-Nom.Sg*
'There is a loaf of bread on the table'

(4) *Kahvi on hyvää-Part.Sg*
'Coffee is nice'
(5) *Auto on uusi-Nom.Sg*
'The car is new'.

Probably as a consequence of the fact that the case alternations differ in certain respects and clear-cut grammar rules cannot always be formulated, the use of the partitive case remains a constant struggle for learners of Finnish. However, as the use of the partitive case is essentially similar in the closely related Estonian language (opposed to Finnish, Estonian can be characterized by slight differences with respect to the aspectual object case alternation, the presence of an additional existential sentence type and the lack of an equivalent to the Finnish nominative-partitive predicative case alternation), this study investigates the use of partitive objects, subjects and predicatives in writings produced by Estonian, German and Dutch learners of Finnish. By comparing the use of different partitive functions in groups of learners from a closely related L1 background (i.e. Estonian) and non-related L1 backgrounds (German/Dutch) at different levels of L2 proficiency, the current study aims to explore and explain how the presence versus absence of relevant prior linguistic knowledge and inter-/intralingual similarities upon which learners can draw is reflected in their writings. The study hereby primarily builds on Jarvis' (2000; 2010) rigid framework for identifying L1 influence. Because the study also aims to identify learners' major stumbling blocks in the use of the Finnish partitive case (either shared by all groups of learners or encountered by specific groups of learners), the purpose of the study is basically two-fold in that it aims to provide valuable insights into the phenomena of inter- and intralingual influence, on the basis of which pedagogical implications can subsequently be drawn.

Research materials (Estonian learner corpus 82,749 words; German LC 60,490; Dutch LC 47,753) were selected from the *International Corpus of Learner Finnish* (ICLFI - a written corpus containing homework assignments from university students of Finnish as a foreign language), aligned to the CEFR levels of proficiency (A1-C2) and analyzed on the basis of sets of combined error-frequency analyses, involving relative frequencies of occurrence, partitive-requiring contexts and partitive over- and underuse errors.

As will be discussed, several intriguing instances of positive and negative L1 influence were detected in the Estonian learner corpus. On the whole, the Estonian learner corpus did not only generally show significantly fewer partitive errors than the remaining learner corpora but also specific error patterns (particularly at the lower proficiency levels) that were due to L1-L2 differences. In contrast, overgeneralization of L2 grammar rules was virtually absent from the Estonian learner corpus but particularly prevalent in the lower proficiency components of the remaining learner corpora.

The conspicuous differences between the learners from related and non-related L1 backgrounds not only indicate that -and how- prior linguistic knowledge matters but also suggest that in order to potentially turn stumbling blocks into stepping stones, the teaching of the use of the partitive case to Estonian learners of Finnish should emphasize (subtle) L1-L2 differences, while learners from non-related L1 backgrounds would particularly benefit from being assisted to gain additional insight into the similarities and differences between the case alternations.

References

Jarvis, Scott (2000) Methodological rigor in the study of transfer: Identifying L1 influence in the Interlanguage Lexicon. *Language Learning* 50: 245-309.

Jarvis, Scott (2010) Comparison-based and detection-based approaches to transfer research. *EURASLA 2010 Yearbook.*

English/Japanese loanword cognates in Japanese English learner writing

Struc, Nicolai; Wood, Nicholas
Reitaku University, Japan
astruc@reitaku-u.ac.jp; nickwood.utu@polka.ocn.ne.jp

One of the characteristics of the modern Japanese language is the abundance of predominately European languages-based loanwords, or *gairaigo,* which comprise an increasing proportion of its lexicon. While a strict definition of loanwords might arguably include all Sino-Japanese words, we adopt the definition proposed by Irwin (2011), which specifies *gairaigo* as foreign words borrowed "…after the mid-16th century and whose meaning is (…) intelligible to the general speech community" (p. 10). Of interest among these, are those that originate from English – over 80% by the mid-1950s and still increasing (Irwin, 2011). A survey of a major Japanese dictionary in 1989 found 10% of words were borrowed (Masui, 1999 as cited in Tomoda, 2005). While many of these words are highly specialized, usage of more common words in television, advertising, newspapers, and conversation is well documented (e.g., Ohshima, 1994; Tomoda, 1999; 2005), and they serve various functions from filling lexical gaps to adding 'special effect' to Japanese words with similar meanings.

Daulton (2008) has demonstrated that a large proportion of the BNC 3000 most frequent word families are present in Japanese as loanwords. In the second language acquisition process, these loanwords may function as a form of cognate, which invites the question of their role in English language acquisition among Japanese learners. In terms of receptive skills, loanwords have been shown to facilitate aural comprehension (Brown & Williams, 1985) and vocabulary recall and recognition (Daulton, 1998). Two corpus-based studies by Daulton (2003; 2008) examined the effect of loanwords in written production. In these investigations of Japanese college students' English writing, he found a strong preference for using loanwords compared with non-loanwords in the first 2000 most frequent word families in the BNC.

If Japanese learners of English do indeed prefer using loanword cognates, it is important to confirm this and broaden the scope of inquiry. The present study seeks to: (1) observe whether a loanword preference is evident in the writing of Japanese university English language majors, (2) observe any changes that occur over a one-year period, (3) observe any differences in loanword deployment between two genres, and (4) compare Japanese writers' loanword deployment patterns with those of NSs. The corpora are comprised of timed written responses, without use of dictionaries, to two prompts. The first prompt elicited a narrative by asking writers to tell the story of two friends who go shopping with one coming home sad and the other happy. The second elicited an argumentative writing sample by asking students to discuss advantages and disadvantages of studying abroad. The NNS writers were 170 Japanese university students aged 18-20, and a second sample was obtained from the same students one year later (57,701 words). Writing samples from 29 NSs (American university students), in response to the same two prompts, comprise another corpus for comparison (26,967 words).

These corpora were submitted to two vocabulary profile analyses: the first using the BNC 3000 (Nation, 2004) wordlists and the second using subsets of these three wordlists comprised only of corresponding loanwords (Daulton, 2008). From these two profiles, the proportion of loanwords, whose frequent deployment may be attributed to a 'loanword effect', can be determined. In addition, measures of lexical diversity, and frequency lists provide a closer description of the writing.

The results show a surprisingly consistent pattern of loanword deployment among the NNSs. While vocabulary profiles show variation between the frequency bands, the overall proportions of loanwords show no significant difference between genres or between the two points in time. Japanese writers do however exhibit a stronger preference for loanwords in their texts than NSs at both points in time and in both genres. Within NSs' writing samples, deployment of 'loanwords' differs with genre, with fewer loanwords present in argumentative writing. As expected, lexical diversity is much greater in NS writing, with NNSs relying on the repetition of fewer types.

The results can be interpreted as tentatively supporting Daulton's (2008) observed 'borrowed word effect' but not to the degree to which his data show. It is difficult to say what distribution of loanwords versus non-loanwords should be expected. However, the NS data offers a baseline with which the NNS data may be compared, and if this can be interpreted as a norm, then there is some preference for loanword cognates evident among Japanese NNSs. The analysis of lexical diversity shows greater diversity in NS writing (even within the loanword subset) as may be expected. So while the proportions differ moderately, there is heavy reliance on specific (loan) words by NNS revealed by word frequencies. Additionally, examination of some loanwords in context suggests cases of negative transfer in which Japanese usage patterns are apparent but inappropriate. Optimism about the usefulness of cognates as a resource for Japanese learners is worth consideration; but it appears that while they may contribute to fluency of production, they should be treated with caution.

References

Brown, J.B. and Williams, C.J. (1985). Gairaigo: A latent English vocabulary base? *Tohoku Gakuin University Review: Essays and Studies in English Eibungaku,* 76, 129-146

Daulton, F. E. (1998). Loanword cognates and the acquisition of English vocabulary. *The Language Teacher.* 20 (1), 17-25.

Daulton, F. E. (2003). The effect of Japanese loanwords on English written production – a pilot study. *JALT Hokkaido Journal*. 7, 4-14

Daulton, F. E. (2008). Japan'*s Built-in Lexicon of English-based Loanwords.* New York: Multilingual Matters.

Irwin, M. (2011). *Loanwords in Japanese*. Amsterdam: John Benjamins.

Nation, I. S. P. (2004). A Study of the most frequent word families in the British National Corpus. In P. Bogaards and B. Laufer (eds). *Vocabulary in a Second Language: Selection, Acquisition and Testing*. (pp. 3-13). Amsterdam: John Benjamins.

Masui, H. (1999). Katakanakotoba. In H. Inoue, (Ed.), *Nihongoyo doko e yuku* (pp. 111-127). Tokyo: Iwanamishoten.

Ohshima, K. (2004). The Movement of Gairaigo Usage: The case of the Asahi newspaper from 1952 to 1997. *Bunkyo Gakuin Daigaku Gaikokugo GakubuGakuin Daigaku Tankidaigaku Kio* 3, 91-102.

Tomoda, T. (1999) The impact of loan-words on modern Japanese. *Japan Forum* 11 (1), 231-253.

Tomoda, T. (2005). The Loanword (gairaigo) influx in the Japanese Language: Contemporary Perceptions and Responses. (Doctoral dissertation, School of Sociology, University of New South Wales).

Norsk andrespråkskorpus – A corpus of Norwegian as a second language

Tenfjord, Kari; University of Bergen; Kari.tenfjord@lle.uib.no

Meurer, Paul; Uni Research paul.meurer@uni.no

Ragnhildstveit, Silje University of Bergen silje.ragnhildstveit@lle.uib.no

In this demo, we will present the corpus design, the contents and the Web-based corpus management platform of Norsk andrespråkskorpus (ASK) which has been extensively used in the Askeladden project at the University of Bergen.

The Ask corpus is accessible in a newly designed and implemented corpus management platform (Corpuscle, developed at Uni Resarch AS) which is also in use for several other, quite diverse corpora at the University of Bergen, and which will eventually be open-sourced. The rationale for devising yet at new tool rather than adopting an existing corpus tool like CorpusWorkbench was our need for full support of hierarchically structured data (XML), improved query execution speed, and a seamless integration of manual corpus annotation and editing. In addition, there is built-in support for multi-valued and set-valued attributes, as well as full support of the Unicode Basic Multilingual Plane (BMP). Secure user authentication is provided via Feide (the Norwegian Identity provider for the educational sector), and the system comes with a fine-grained authorization scheme. The corpus platform is accessible through the European Clarin initiative (via the national Clarino centres).

ASK is a corpus of 1700 essays written by learners of Norwegian as a second language with ten different first languages. German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali. There is also a control corpus of 200 essays written under the same conditions by native speakers of Norwegian.

The essays are collected from two different tests in Norwegian, one intermediate level test and one advanced level test. This makes the textual data within each test level homogeneous when it comes to both test conditions and text type. The essays from the intermediate test level are mainly expository and the essays from the advanced level test is mainly argumentative. In a recent project most of the essays were reassessed for proficiency levels in accordance with the common European Framework of Reference for Languages by ten different assessors.

The corpus texts are annotated with grammatical categories, lemmas and error codes. They also contain metadata of each text including several personal variables such as age, home country, education, length of residence in Norway, type and intensity of language instruction, other second languages etc. The interface makes it possible to form different kinds of sub-corpora, to formulate different kinds of queries at the lexical, morphological and syntactical level, and to study correlations with personal variables. Personal variables can be used to form sub-corpora before the query, and can also be displayed in the concordance in columns.

Word level annotation (lemma, detailed part of speech and grammatical function) is performed automatically by the *Oslo-Bergen-Tagger,* and the POS tagging is manually post-edited. A set of error codes has been manually assigned together with suggested corrections. Specific guidelines for error tagging were developed to avoid theoretical bias and ensuring that the error codes only represent theory neutral descriptions of the learner language.

The inserted corrections are used for generating a searchable parallel corpus containing sentence-aligned original and corrected versions of the learners' essays. This allows powerful queries combining a specific construction in the learners' texts with a specific correction, or a specific error code, such as "M" (something missing) a missing preposition in learners' texts, may be corrected to the preposition *på*.

The new interface has been updated with an extended query language, a new concordance format and a new download function. In the Web-based user interface, queries can be formulated directly in a query language or can be composed through menu choices. The search menu allows

of the creation of sub-corpora based on first language, proficiency level, text topic or other variables, and also a combination of several variables. The menu also allows of the combination of different linguistic features in one position in the corpus or in sequences of positions.
The query results can be viewed in different types of concordances, also in a parallel concordance of learner language and the corrected version. There is a link between the concordance and the individual texts, which makes it possible to access the whole texts in a separate window.
The demonstration will be based on the fully functional corpus and its interface and will focus on query composition, including the combination of grammatical constructions with personal variables, the error codes and the proposed corrections.

References

Carlsen, Cecilie. (2012) Proficiency level - a fuzzy variable in computer learner Corpora. In: *Applied Linguistics;Volum 33 (2).* 161-183

Meurer, Paul. (2012) Corpuscle – a new corpus management platform for annotated corpora. In: Andersen, Gisle (ed.): *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian, Studies in Corpus Linguistics 49*, John Benjamins, .

Tenfjord, Kari; Hagen, Jon Erik & Johansen, Hilde. (2006) The hows and whys of coding categories in a learner corpus (or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy). *Rivista di Psicolinguistica Applicata (RiPLA) VI(3)*: 93-108.

Askeladden project: http://www.uib.no/fg/askeladden

ASK corpus: http://clarino.uib.no/ask

Oslo-Bergen-tagger: http://tekstlab.uio.no/obt-ny/

A longitudinal learner corpus investigation of vocabulary learning before, during, and after residence abroad

Tracy-Ventura, Nicole[1]; McManus, Kevin[2]; Mitchell, Rosamond [2]

University of South Florida[1]; University of Southampton[2]

nkt@usf.edu; K.McManus@soton.ac.uk; R.F.Mitchell@soton.ac.uk

Within the field of second language acquisition (SLA), the effect of residence/study abroad on language development has received a considerable amount of attention (see reviews by Collentine, 2009; Kinginger, 2011); yet, much of the research to date has suffered from major weaknesses in research design and statistical precision (Rees & Klapper, 2008). For example, most studies tend to measure linguistic development pre- and post a stay abroad, without investigating how participants' language develops while they are there. Additionally, many studies tend to focus primarily on oral skills (e.g., fluency) elicited through a proficiency-based interview or a picture-based narrative without investigating learners' language use across different task types, whether oral or written. Vocabulary development is one area of linguistic development that has received far less attention in residence/ study abroad research despite its popularity in second language research in general (e.g., Daller et al., 2007; Milton, 2009). Of the few studies investigating vocabulary development in the residence/study abroad context, most have adopted measures of receptive vocabulary knowledge (e.g., Dewey, 2008). One notable exception is Foster (2009) who investigated productive lexical knowledge using D, a measure of lexical diversity (Malvern & Richards, 2002), and found that compared to at-home learners, those who studied abroad performed similarly to native speakers. In recent research D has proven to be a more reliable measure of lexical diversity because it is less affected by text length than a measure such as type-token ratio (see McCarthy & Jarvis, 2007). It has also been shown to correlate with measures of general language proficiency (Yu, 2009; Tracy-Ventura et al., to appear).

With this in mind, the current study investigates productive vocabulary development before, during and after residence abroad through the use of a learner corpus approach and focuses on the following research questions:

1) Does learners' lexical diversity (as measured by D) change over time?

2) Are there differences in lexical diversity across tasks (oral vs written)?

3) Do task topics/prompts influence D scores?

Two new longitudinal learner corpora will be introduced which were compiled as part of a large-scale project on the acquisition of French and Spanish before, during, and after a 9- month stay abroad. Two parallel corpora (each around 300,000 words) include both oral and written data collected six times over 20 months, including three visits whilst abroad. Participants were students of L2 Spanish (n=27) and L2 French (n=29) at a British university spending their third year of a four-year degree in either France, Spain or Mexico. Native speakers (n=10 for each language) also completed all the same tasks at one time. Oral and written data were formatted in CHAT for use with the CLAN program (free from CHILDES, see MacWhinney, 2000). In addition to presenting the design of the corpus, we present a case-study demonstrating how these two longitudinal learner corpora can be used to investigate language learning during residence abroad. We investigate development of lexical diversity using D (Malvern & Richards, 2002; computable using the CLAN program) in three different tasks: 1) a semi-structured oral interview, 2) an oral picture-based narrative, and 3) a written argumentative essay. All three tasks were administered at each data collection cycle allowing us to investigate change in scores over time. The oral interview included different questions at each data collection cycle. In contrast, three prompts were used in sequence for the writing and three picture-based narratives were used in sequence in the oral retells. As a result, every prompt or story was repeated once each allowing us to also test for an influence of topics/prompts on D scores (following Foster, 2009; Yu, 2009).

By taking a longitudinal learner corpus approach we demonstrate that both sets of learners show significant gains in lexical diversity in their oral production from early on in their stay abroad whereas development in written production is slower to develop. In particular, the biggest change in D scores on the oral interview was between the pretest and visit 1 abroad, with little change during visits 1-3. In contrast, learners' D scores in written production decreased from the pretest to visit 1 abroad but changes were evident between visits 1-3. The results also demonstrate clear effects of task type (interview, narrative, argumentative essay) with the lowest D scores appearing overall on the narratives. Task prompts and the content of narratives also appear to influence D scores as evidenced from learners and native speakers alike. That is, certain writing prompts and narratives tended to elicit higher D scores than others. These results suggest that lexical diversity improves throughout a stay abroad and that when measuring development of lexical diversity longitudinally, the type of elicitation task and prompts used should be carefully chosen. Furthermore, having both oral and written language samples can provide the most reliable evidence of vocabulary development during a stay abroad.

References

Collentine, J. (2009). Study abroad research: Findings, implications, and future directions. In M. Long & C. Doughty (Eds.), The handbook of language teaching (pp.218-233). Maldon, MA: Wiley Blackwell.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds) (2007). Modelling and assessing vocabulary knowledge. Cambridge: Cambridge University Press.

Dewey, D. (2008). Japanese vocabulary acquisition by learners in three contexts. Frontiers: The Interdisciplinary Journal of Study Abroad, 15, pp.127-148.

Foster, P. (2009). Lexical diversity and native-like selection: the bonus of studying abroad. In B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application (pp. 91-106). Basingstoke: Palgrave Macmillan.

Kinginger, C. (2011). Enhancing language learning in study abroad. Annual Review of Applied Linguistics, 31, 58-73.

Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. Language Testing, 19, 85-104.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, P. & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. Language Testing, 24, 459–88.

Milton, J. (2009). Measuring second language vocabulary acquisition. Bristol: Multilingual Matters.

Rees, J. & Klapper, J. (2008). Issues in the quantitative longitudinal measurement of second language progress in the study abroad context. In L. Ortega & H. Byrnes (Eds.), The longitudinal study of advanced L2 capabilities (pp. 89-105). New York: Routledge.

Tracy-Ventura, N., McManus, K., Ortega, L., & Norris, J. (to appear, 2013). “Repeat as much as you can”: Elicited imitation as a measure of global proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds, Proficiency assessment issues in SLA research: Measures and practices. Multilingual Matters.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. Applied Linguistics, 31, 236-259.

Information Structural Development in the Writing of Very Advanced Learners of English: A cross-linguistic longitudinal study

de Haan, Pieter; van Vuuren, Sanne
Radboud University Nijmegen; Radboud University Nijmegen
p.dehaan@let.ru.nl; s.v.vuuren@let.ru.nl

Why is it that even at advanced levels of acquisition Dutch EFL learners produce texts which can instantly be recognized as having been written by a native speaker of Dutch, in spite of the absence of obvious errors in grammar or lexis? As Carroll and Lambert have noted, “the learning problem at advanced stages of learning is not one of linguistic form” (2003: 270). Rather, advanced learners differ from native speakers in the frequency with which they use structures available in the language and in the application of language-dependent principles of information structure (Carroll & Lambert 2003; Callies 2009).

One particularly problematic area for Dutch EFL learners seems to be the information status of pre-subject adverbials. Unlike English, Dutch has a multifunctional pre-subject position, often occupied by adverbials which connect the sentence to the preceding discourse. While in Dutch pre-subject adverbials can function as neutral discourse links, in English, if adverbials can be fronted at all, it is in most cases “a very marked choice” (Biber et al. 1999: 803).

In order to determine how exactly Dutch EFL student writers differ from native writers in their use of pre-subject adverbial phrases, we evaluated the use of pre-subject adverbial phrases in a longitudinal corpus of texts written by Dutch students of English between their first and fourth year at university. This corpus, which is the Dutch component of LONGDALE (Granger 2009), includes essays on various aspects of British and American literature and culture as well as shorter argumentative in-class assignments which had to be completed within thirty minutes, which we analyzed separately. We used two native speaker reference corpora: the Louvain Corpus of Native English Essays (LOCNESS), consisting of written work by native speaker students comparable in age and academic background, and the VU Native Speaker Published Research Article Corpus (VUNSPRAC) compiled by Philip Springer (2012), representing the level of professional academic writing our student writers are hoping to achieve. These corpora were parsed using the Stanford Parser (Klein and Manning 2003), after which pre-subject adverbials were filtered out with Corpus Studio (Komen 2012). This procedure resulted in a database of over 10,000 adverbials, which were then categorized according to their function label (e.g. 'instrument', 'addition', 'place') as well as their discourse status ('identity', 'inferred', 'assumed', 'new' or 'inert') and distance to their antecedent (-1, -2, etc.).

Initial results show that advanced Dutch learners of English overuse categories of pre-subject adverbials that link back to the directly preceding discourse, most notably (pronominal) addition adverbials such as 'apart from that', 'on top of that' and 'next to that', and place adverbials such as 'in the novel' or 'in his poem'. There are clear differences between the two text types: while the in-class assignments use twice as many 'addition' adverbials compared to the essays, the essays, in their turn, use over three times as many 'place' adverbials. Although there is a clear development in the direction of native writing, transfer of information structural features of Dutch can still be observed even after three years of extended academic exposure.

References

Biber, Douglas, Stig Johannsson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999) *The Longman Grammar of Spoken and Written English.* Harlow: Longman.

Callies, Marcus (2009) *Information Highlighting in Advanced Learner English: The Syntax-Pragmatics Interface in Second Language Acquisition.* Amsterdam/Philadelphia: John Benjamins.

Carroll, Mary and Monique Lambert (2003) 'Information Structure in Narratives and the Role of Grammaticised Knowledge: A Study of Adult French and German Learners of English'. In Information Structure and the Dynamics of Language Acquisition, ed. Christine Dimroth and Marianne Starren, 267-87. Amsterdam: Benjamins.

Granger, Sylviane (2009) LONGDALE, from https://www.uclouvain.be/en-cecl-longdale.html. (Date of access 14-02-2013).

Klein, Dan and Christopher D. Manning (2003) 'Accurate Unlexicalized Parsing'. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Komen, Erwin (2011) 'Coreferenced corpora for information structure research'. In *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. (Studies in Variation, Contacts and Change in English 10), ed. Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen. Helsinki: Research Unit for Variation, Contacts, and Change in English. Available online at <http://www.helsinki.fi/varieng/journal/volumes/10/komen/>

Springer, Philip (2012) *Advanced Learner Writing: A Corpus-based Study of the Discourse Competence of Dutch Writers of English in the Light of the C1/C2 Levels of the CEFR*. Oisterwijk: BOXpress.

Autonomous collocation error correction with a data-driven approach

Vincze, Orsolya; Alonso Ramos, Margarita
Universidade da Coruña
ovincze@udc.es; lxalonso@udc.es

We present an experimental study which aims at testing to what extent concordance lines can aid language learners to correct different types of collocation errors. The relation between corpus and collocation learning is twofold: on the one hand, corpus studies have singled out collocations as an important problem area for second language learning (Granger 1998; Nesselhauf 2005); on the other hand, corpora have also been suggested as an especially adequate resource for teaching collocations (e.g. Kilgarriff 2009; Chang et al. 2008). Furthermore, in the field of second language acquisition, it has been noted that classroom activities with corpora not only provide a valid context for language learning through the observation, selection and use of language items, but they also enhance learners' skills in and hence promote autonomous learning (e.g. Chambers and O'Sullivan (2004).

The specific case of using corpora as an aid for students to correct their own writing was examined by a number of authors (e.g. Chambers & O'Sullivan 2004; Wu et al. 2009). Our study is carried out in the framework of a research project which aims at the development of an active collocation learning environment including a writing aid tool for learners of Spanish. In order to better understand the nature of errors made by learners when using collocations, we annotated a part of the CEDEL2 learner corpus (Lozano 2009). One of the outcomes of the annotation process was a fine-grained typology of collocation errors (Alonso Ramos et al. 2010; Wanner et al. 2013). Since the system being developed in the framework of the project provides feedback to users in the form of concordances, we thought it was necessary to verify whether the errors included in our typology can be autonomously corrected by learners with the help of concordance lines. Accordingly, we designed an experimental study that aims to answer the following questions: 1) The correction of what error types poses more difficulty for the students when presented with the concordance lines? 2) What problems can we foresee in the case of each error type when offering automatic feedback? 3) What criteria should be observed when it comes to the automatic selection and presentation of feedback concordance lines in order to better assist students in the revision of collocation errors?

The experiment described in our study is carried out with students of Spanish as a second language. Subjects are given 20 sentences extracted from the annotated CEDEL2 learner corpus, each of which contains one or more collocation errors. The sample is put together in a way that it is considered to be representative of the different error types described in the error typology. In the course of the experiment, participants are asked to revise erroneous segments marked in the sentences, first without any aid and, second, with the help of concordance data either in the form of full sentences, or in the form of short segments representing commonly used patterns of the collocation. In the full version of our presentation, we will discuss the results of this experiment, and its implications for the design of the learning tool.

References

Alonso Ramos, Margarita, Wanner, Leo, Vincze, Orsolya, Casamayor, Gerard, Vázquez Veiga, Nancy, Mosqueira Suárez, Estela, & Prieto González, Sabela. (2010) Towards a motivated annotation schema of collocation errors in learner corpora. In: Calzolari, Nicoletta et al. (eds.) *Proceedings of the Seventh conference on International Language Resources and Evaluation* (pp. 3209–3214), ELRA.

Chambers, Angela, & O'Sullivan, Íde. (2004) Corpus consultation and advanced learners' writing skills in French. *ReCALL*, *16*(01): 158-172.

Chang, Yu-Chia, Chang, Jason S., Chen, Hao-Jan & Liou, Hsien-Chin. (2008) An automatic collocation writing assistant for Taiwanese EFL learners: a case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3): 283–299.

Granger, Sylviane. (1998) Prefabricated patterns in advanced EFL writing: collocations and formulae. In: Cowie, Anthony P. (ed.) *Phraseology: theory, analysis and applications* (pp. 145-160), Oxford: Oxford University Press.

Lozano, Cristobal. (2009) CEDEL2: Corpus escrito del español L2. In: Bretones Callejas, Carmen M. et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind* (pp. 197–212) Almería: Universidad de Almería.

Kilgarriff, Adam. (2009) Corpora in the classroom without scaring the students. Paper presented at the 18th International Symposium on English Teaching, Taipei. Retrieved from: http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc

Nesselhauf, Nadja. (2005) *Collocations in a learner corpus*. Amsterdam & Philadelphia: John Benjamins.

Wanner, Leo, Alonso Ramos, Margarita, Vincze, Orsolya, Nazar, Rofelio, Ferraro, Gabriela, Mosqueira, Estela, & Prieto, Sabela. (2013) Annotation of collocations in a learner corpus for building a learning environment. In: Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds.) *Twenty years of learner corpus research: looking back, moving ahead* Louvain-la-Neuve: Presses universitaires de Louvain.

Wu, Shaoqun, Witten, Ian. H., & Franken, Margaret. (2009) Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge. *ReCALL*, *22*(01): 83–102.

Essays on the one hand and research papers on the other hand: Variation in the use of “on the one hand”/ “on the other hand” in learner writing

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

One of the areas identified as problematic for advanced learners is the native-like use of multi-word units (MWUs), which often serve to express rhetorical functions (RFs) in written discourse (e.g. *as an example; in contrast to*) (e.g. Granger 1998; Paquot 2008, 2010). Learner Corpus Research has revealed that learners tend to use certain units more frequently than English native speakers and often make unidiomatic choices in terms of their use in discourse (e.g. Crew 1990; Granger & Petch-Tyson 1996). The general research methodology adopted in Learner Corpus Research (LCR) has been contrastive, 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 differences in learners’ and native speakers’ language use, this kind of approach, however, does not offer a comprehensive list of possible explanations of learners’ choices and thus fails to provide a full picture of learners’ language use in writing. Meanwhile, a variationist perspective on learner production considering a possible influence of variables (e.g. genre, task setting, etc.), has a potential to provide missing information on (hidden) systematicity in learners’ linguistic behaviour. Yet, studies combining contrastive and variationist perspectives in their analysis of written interlanguage are still scarce (cf. however, Ädel 2008; Paquot et al. 2011). Thus, for example, 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, abstracts, etc. (cf. however, Römer 2009; Wulff & Römer 2009).

This study serves to exemplify a possible way of combining contrastive and variationist frameworks to investigate German learners’ use of the MWUs *on the one hand/ on the other hand*, which often co-occur in a stretch of discourse, in argumentative essays and research papers, and addresses the following research questions:

1. Is there variation found as to learners’ preference of one of the variants (e.g. “correlative” *on the one hand/ on the other hand* or “non-correlative” *on the other hand* (Bell 2004)) in writing?

2. Are certain variants preferred over others in certain RFs and/ or in a particular sentence position in learner writing?

3. Is variation in learners’ use of these MWUs determined by genre/ text type as a plausible variable?

4. To what extent is learners’ use of these MWUs similar/ different to that of native speakers?

The analysis primarily draws on the German components of two types of learner corpora, i.e. the International Corpus of Learner English (ICLE) (Granger et al. 2009) and the pilot version of the *Corpus of Academic Learner English* (CALE), a Language for Specific Purposes learner corpus. Additionally, a selection of native-speaker texts will be examined. Preliminary results indicate a tendency of learners to prefer one of the two variants of the MWUs in expressing the RFs of contrast and listing in argumentative essays and research papers.

References

Ädel, Annelie. 2008. "Involvement features in writing: do time and interaction trump register awareness?" In: Gaëtanelle Gilquin, S. Papp. & M. B. Díez-Bedmar (eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam, Atlanta: Rodopi. 35-53.

Bell, David. 2004. "Correlative and non-correlative "on the other hand" ", *Journal of Pragmatics* 26. 2179-2184.

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

Granger, Sylviane. 1998. "Prefabricated patterns in advanced EFL writing: Collocations and formulae" In: Anthony P. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 145-160.

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

Granger, Sylviane, Erstelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Paquot, Magali. 2008. "Exemplification in learner writing: a cross-linguistic perspective" In: Sylviane Granger and F. Meunier (eds.), *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: Benjamins. 101-119.

Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. United States: Continuum Publishing Corporation.

Paquot, Magali, Hilde Håsselgard & Signe Oksefjell Ebeling. 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, Ute. 2009. "English in academia: Does nativeness matter? ", *Anglistik: International Journal of English Studies* 20(2): 89-100.

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

Author Index

Keyword Index