Learner Corpus Research 2019
Warsaw, 12–14 September

Book of Abstracts

Table of Contents

PLENARY TALKS

FULL PAPERS

WORK-IN-PROGRESS REPORTS

POSTERS

SOFTWARE DEMOS

Corpora & Cambridge Assessment English: A Widening Perspective

Mark Brenchley
Cambridge Assessment English
brenchley.m@cambridgeenglish.org

2016 saw the 10th anniversary of the *English Profile Programme*, itself a marker in the longstanding commitment of *Cambridge Assessment English* to corpus-based research (Barker, 2006, 2016). That commitment is based on at least two premises. Firstly, and especially given our wider commitment to a communicative view of language ability and language assessment, corpora represent an invaluable resource for understanding how learners use and develop their linguistic resources. Secondly, corpora represent a core means of expanding our technological capabilities; underpinning, for example, the development and validity of auto-marked tests such as *Linguaskill* and its newly launched variant *Linguaskill Business*.

Both premises are reflected in the wide range of practical purposes to which *Cambridge English* already puts corpora and corpus-based methods; from the ongoing process of test validation and test revision (e.g. Shaw & Weir, 2007; Elliott & Lim, 2016; Saville, 2003) through to the development of key assessment resources such as *English Profile*, the official English Reference Level Description for the *Common European Framework of Reference* (Harrison & Barker, 2015). They are also reflected in our commitment to the establishment and expansion of novel corpora, most notably the *Cambridge Learner Corpus* and the *Cambridge English Profile Corpus*, as well as our longstanding tradition of working within a wider community of researchers such as *ALTA* and *Cambridge University Press*.

So framed, the present talk will proceed in two parts. The first will provide a more detailed overview of current research activities at *Cambridge Assessment English*, outlining how they inform the development of exams such as *Linguaskill* and *Cambridge English Qualifications*. The second offers a wider perspective on prospects for expanding the practical role of corpora, including our plans for new resources such as the development of a spoken learner corpus to complement the written *Cambridge Learner Corpus*.

Central to the wider perspective outlined in the second half is the increasing dominance of computer-based testing. This dominance promises to be a boon for our corpus-based activities, providing for learner performances that are not only more extensively available, but in a format that makes them substantially easier to process. In turn, this availability is complemented by the value of this material as a means of further driving the quality of our assessment work, whether this be along the more technological dimensions of areas like auto-marking and plagiarism detection, or in terms of our capacity for continuing to develop an approach to assessment that is fully learner-oriented in the sense of Jones & Saville (2016).

At the same time, however, the sheer scale of material that computer-based testing is increasingly making available raises a number of important considerations. One of these is the challenge of ensuring that we are able to systematically process, organize, and integrate such a large, ever-expanding body of material so as to maximise its effectiveness. Another is the challenge of ensuring that our increasing practical use of such material does not become a validational straightjacket; rather, that we continue to interrogate how this material can best be analysed, whether on its own terms, in its relation to "external" corpora, or in relation to non-corpus based methods and sources of information. Each of these considerations represents a substantive challenge to the wider validity of corpus-based assessment work. Addressing them will be central to ensuring that *Cambridge English* continues to reap the benefits of its longstanding commitment to this form of research.

References

Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes 26*, 2–4.

Barker, F. (2016). The English Profile Programme 10 years on. *Research Notes 63*, 33–35.

Elliott, M., & Lim, G. S. (2016). The development of a new reading task: A mixed methods approach. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.) *Second Language Assessment and Mixed Methods Research*, Studies in Language Testing 43 (pp. 233–268). Cambridge: UCLES/Cambridge University Press.

Harrison, J., & Barker, F. (2015). *English Profile in Practice*, English Profile Studies 5. Cambridge: UCLES/Cambridge University Press.

Jones, N., & Saville, N. (2016). *Learner-Oriented Assessment: A Systemic Approach*, Studies in Language Testing 45. Cambridge: Cambridge University Press.

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. Weir & M. Milanovic (Eds.) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing 15 (pp. 56–120). Cambridge: UCLES/Cambridge University Press.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing*, Studies in Language Testing 26. Cambridge: UCLES/Cambridge University Press.

Expanding the scope of complexity research in SLA: a phraseological perspective

Magali Paquot
Université catholique de Louvain
magali.paquot@uclouvain.be

Usage-based research in corpus linguistics, psycholinguistics and cognitive linguistics has provided convergent evidence that lexis and grammar are inextricably intertwined and that word combinations, be they framed in terms of phraseological units, formulaic sequences or constructions, play crucial roles in language acquisition, processing, fluency, idiomaticity and change (e.g. Ellis, 1996; Sinclair, 1991; Wray, 2002; Schmitt, 2004; Goldberg, 2006). Second language research has been relatively slow to follow suit but phraseology, formulaic language and constructions are now at the forefront of debates in foreign language learning and teaching (Meunier & Granger, 2008; Polio, 2012). Learner corpus studies have already provided unique insights into the links between word combinations and L2 proficiency and development (e.g. Paquot & Granger, 2012; Ellis et al., 2015; Ebeling & Hasselgård, 2015). Research, however, is still extremely fragmented and not all domains of L2 research have navigated the transition. Interlanguage complexity research, most particularly, has remained impervious to current theoretical developments related to how words combine together to form meaningful units: Complexity has traditionally been narrowed down to syntactic complexity with a strong focus on clause-related measures (e.g. the number of coordinate or dependent clauses per T-unit), and lexical complexity is still very much regarded as its poor relation (Ortega, 2012). This is particularly unfortunate since complexity is considered one of the "major research variables in applied linguistic research" (Housen & Kuiken, 2009): measures of linguistic complexity are widely used to describe L2 performance, assess L2 proficiency, and trace L2 development.

Today, interlanguage complexity research stands at a crossroads. Ortega (2012) described complexity as "a construct in search of theoretical renewal". Measures of complexity have been repeatedly criticized for their lack of theoretical foundation and construct validity (i.e. how well they measure the construct that they are intended to measure) (e.g. Norris & Ortega, 2009; Biber et al., 2011; Pallotti, 2015). Leading researchers in the field have also called for an expanded view of complexity as a multifaceted and multidimensional construct that cannot be fully explored via just one of its dimensions (as is commonly done) but requires to be operationalized with a battery of measures (including new and more specific measures) tapping different properties of the construct in multivariate research designs (e.g. Ortega, 2012; Bulté & Housen, 2012). In Paquot (2019), I have argued that a successful renewal of the domain will also require a better appreciation of the phraseological dimension of language use.

In this presentation, I will report the first results of a 5-year FNRS research project (2016–2021) that aims to define and circumscribe the linguistic construct of phraseological complexity, i.e. "the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units" (Paquot, 2019), and to theoretically and empirically demonstrate its relevance for L2 complexity research, and more generally for theories of L2 use and development. The project centres around four main objectives: (1) determine the dimensions of phraseological complexity, (2) establish the construct validity of phraseological complexity measures automatically calculated using natural language processing (NLP) techniques and corpus data, (3) chart the development of phraseological complexity in L2 writing and speech, and (4) identify the best set of complexity measures to adequately capture the dynamics of phraseological complexity development over time.

To achieve these objectives, I have started investigating the diversity and sophistication of word combinations in a variety of cross-sectional and longitudinal written and spoken EFL learner corpora (e.g. the Varieties of English for Specific Purposes Database (VESPA) learner corpus, the Longitudinal Database of Learner English (LONGDALE), and the Trinity Lancaster Spoken Learner Corpus). In the presentation, I will briefly summarize some of the results and focus more particularly on the conceptual/theoretical and methodological issues faced.

I will round off my talk with a discussion of what I believe are the most important implications and most promising applications of this research programme.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly 45*(1), 5–35.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder, I. (Eds.), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency* (pp. 21–46). Amsterdam: John Benjamins.

Ebeling, S., & Hasselgård, H. (2015). Learner corpora and phraseology, In S. Granger, G. Gilquin & F. Meunier (Ed.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207–230). Cambridge: Cambridge University Press.

Ellis, N. C. (1996). Sequencing in SLA: Phonological Memory, Chunking and Points of Order. *Studies in Second Language Acquisition* 18, 91–126.

Ellis, N., Simpson-Vlach, R., Römer, U., O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 357–378). Cambridge: Cambridge University Press.

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

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

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

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics 30*(4), 555–578.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact* (pp. 127–155). Berlin, Boston: De Gruyter.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research 31*(1), 117–134.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

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

Polio, C. (Ed.), (2012). Special issue on formulaic language. *Annual Review of Applied Linguistics 32*.

Schmitt, N. (Ed.), (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.

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

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

Exploring Treelets in Learner Corpora

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

Syntactic annotation of native language corpora has a number of well-known applications in corpus-based research and natural language processing. The use of syntactic parsers in learner corpus research is complicated by the fact that treebanks used to develop such tools are, with some notable exceptions (Berzak et al. 2016), derived from samples of native language. As a result, the accuracy of automatic syntactic annotation is less predictable when applied to learner data. Despite these limitations, learner corpora have been parsed to enable searching over basic dependency structures (Pęzik, 2012, cf. http://pelcra.pl/PLEC/syntax.do) or estimate various syntactic and phraseological characteristics of learners' language.

This paper investigates some potentially new aspects of exploring dependency-annotated learner corpora. It introduces a software tool called *Treelets*, which can be used to generate and explore Automatic Combinatorial Dictionaries (ACDs) from syntactically parsed corpora. The tool implements a relational approach to extracting phraseology from dependency-parsed corpora as described by (Pęzik 2018). At the theoretical level, the approach is loosely inspired by the so-called Continuity Restraint, which predicts that phraseological units have underlying connected dependency structures (O'Grady, 1998). The entry structure of the ACD generated by *Treelets* is based on a list of lemmas found in the source corpus. Each entry contains information about recurrent subtrees of sentence dependency trees (called *treelets*) in which a given headword is found, including frequency, dispersion, strength of association and independence (a measure of how often they occur independently of larger recurrent treelets). The ACD therefore records explicitly typed, recurrent subtrees of arbitrary length, rather than just binary collocations, n-grams or skip-grams of contiguous word tokens. To illustrate, the binary collocations *deep breath* and *close look* are recorded in the ACD and marked as regularly subsumed by larger recurrent phrasemes such as *take a deep breath* and *take a close look*. The treelet take a close look is linked to its subsuming (and also recurrent) structure *take a close look at*, etc. The explicit marking of the subsumption relation between lower- and higher-order treelets makes it possible to observe subtle restrictions on their lexico-grammatical roles. For instance, using an ACD generated from a reference corpus of English, it is easy to see that the restricted collocation *deep breath* is not a very independent structure as it used almost exclusively as the direct object of just a handful of verbs such as *take* or *draw*.

After introducing *Treelets* as a corpus exploration tool, I will discuss its relevance to learner corpus research using example ACDs extracted from manually and automatically parsed learner corpora. I will also discuss the possibility of using ACDs derived from reference corpora of English to estimate the distribution of formulaic treelets in learner corpora. Finally, I will address two methodological problems inherent to relational phraseology extraction: the risk of looking at "impressions of language detail noted by people" (Sinclair 1991: 4) and the dichotomy between *recall from memory* and *recomposition* as two possible interpretations of recurrence in corpora.

References

O'Grady, W. (1998). The Syntax of Idioms. *Natural Language & Linguistic Theory 16*(2), 279–312.

Pęzik, P. (2012). Towards the PELCRA Learner English Corpus. In P. Pęzik (Ed.), *Corpus Data across Languages and Disciplines* (pp. 33–42). Frankfurt am Main: Peter Lang.

Pęzik, P. (2018). *Facets of Prefabrication. Perspectives on Modelling and Detecting Phraseological Units*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

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

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., Katz, B. (2016). Universal Dependencies for Learner English. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 737–746). Berlin: Association for Computational Linguistics. https://www.aclweb.org/anthology/P16-1070

Methodological Reform and Learner Corpus Research

Luke Plonsky
Northern Arizona University
lukeplonsky@gmail.com

Quantitative research methods in applied linguistics are currently undergoing a period of major reform. There are several causes or conditions that have led us here. For one, as the field began to apply meta-analysis in the last two decades as a means to understand empirical evidence in its aggregate form, many syntheses uncovered—whether by design or more incidentally—methodological challenges facing individual substantive domains. Developing alongside such observations is a growing set of tools for empirical examinations of 'study quality' (see e.g., Plonsky, 2013; Paquot & Plonsky, 2017), which had hitherto been largely assumed or de-emphasized in favor of theoretical and/or practical concerns. An enhanced awareness of methodological practices could also, rather simply, be argued to be a natural consequence of our maturity as an academic discipline (e.g., Marsden & Plonsky, 2018; Ortega, 2005).

Regardless of its origins, evidence of this movement can be observed in many distinct settings and venues. As we might expect, important steps have been taken by academic journals in the form of editorials (e.g., Trofimovich & Ellis, 2015), revised author guidelines (e.g., Norris, Plonsky, Ross, & Schoonen, 2015), and new procedures for and indicators of 'open science' (Marsden, Morgan-Short, Trofimovich, & Ellis, 2018). The movement is also manifesting itself through a variety of other activities both within learner corpus research (LCR) and adjacent domains. These include (a) workshops/bootcamps and methodologically oriented symposia, (b) studies of methodological literacy/training (e.g., Gonulal, Loewen, & Plonsky, 2017), (c) a newly added Research Methods strand at AAAL, (d) novel analytical techniques (e.g., bootstrapping in Gries, 2006, 2013; Bayesian data analysis in Norouzian, de Miranda, & Plonsky, 2018); and (e) methodological syntheses seeking to describe and evaluate research and reporting practices (e.g., Marsden, Thompson, & Plonsky, 2018). Paquot and Plonsky (2017), for instance, systematically reviewed 66 methodological features in a sample of 376 LCR studies. The results revealed a number of inconsistencies and infelicities ranging from corpus design and sampling to statistical analyses and data reporting practices. However, there is limited evidence that LCR has fully embraced the need for a number of changes necessary to maximize its potential for improving our understanding of L2 development, knowledge, and use.

This paper begins with an overview of the methodological reform movement taking place in applied linguistics, highlighting the notion of study quality and the motivations behind open science. The discussion will then apply these principles to LCR, exploring challenges and opportunities unique to the domain. Suggestions will also be put forth toward an agenda of methodologically-oriented work at the intersection of LCR and methodological reform.

References

Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad, & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219–228). Philadelphia, PA: John Benjamins.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *International Journal of Applied Linguistics* 168, 4–32.

Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad, & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219–228). Philadelphia, PA: John Benjamins.

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. (2018). Introducing Registered Reports at Language Learning: Promoting transparency, replication, and a synthetic ethic in the language sciences [Editorial]. *Language Learning* 68, 309–320.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics* 39, 861–904.

Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning* 68, 1032–1075.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning* 65, 470–476.

Ortega, L. (2005). Methodology, epistemology, and ethics in instructed SLA research: An introduction. *Modern Language Journal* 89, 317–327.

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*, 61–94**.** doi 10.1075/ijlcr.3.1.03paq

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition* 35, 655–687.

Trofimovich, P., & Ellis, N. (2015). Open science badges [Editorial]. *Language Learning* 65, v–vi.

Repeats in native and interlanguage speech:
Exploring traces of cross-linguistic influence and individual variation

Hege Larsson Aas, Sylvi Rørvik
Inland Norway University of Applied Sciences
hege.aas@inn.no, sylvi.rorvik@inn.no

This paper presents the results of a study investigating the use of repeats (e.g. *the the* and *it's it's*), with the aim of exploring interlanguage fluency variations and the potential for transfer of fluency behavior from native language (NL1) Norwegian to interlanguage (IL) English. The study seeks to answer the following research question: can IL speakers' use of repeats be reflected in their NL1 behavior, in terms of equivalent frequency, types and/or position?

While traces of the *disfluent* nature of this feature have been found in experimental studies of language comprehension (MacGregor et al., 2008), attempts to establish a relationship between the use of repeats and IL proficiency levels remain largely inconclusive (Molenda et al., 2018). Corpus-based fluency research has further revealed that repeats may serve communicative purposes beyond the planning of speech (Denke, 2005), and considerable disparity has been found in their distribution (Gráf, 2017; Götz, 2013).

To explore the research question, all sequential repeats were manually identified in six interviews from the (forthcoming) Norwegian component of the LINDSEI corpus (Gilquin et al., 2010) and comparable interviews in the speakers' NL1. They were further categorized according to type (word class(es)), number of repeated elements, and position. Our results confirm previous findings regarding the heterogeneous frequency patterns of this feature, with frequencies ranging from 0.19 to 1.80 repeats per hundred words in English, and 0.22 to 1.26 phw in Norwegian. Our results also show that the speaker who prefers this strategy the most among the six speakers in the material does so across languages, which may be an indication of individual preferences transferring to the interlanguage. The majority of the repeats found in our material were one-word repeats. Repeated segments of two or more words were more common in Norwegian, which may indicate a greater level of automatization in the NL1. Some traces of cross-linguistic influence were found, such as one speaker's preference for repeating the first person personal pronoun (*I/jeg*) in both languages. Our results thus support the idea that "the corpus-as-a-whole average may at times mask an amazing spectrum of individual competencies across the learners in a learner corpus" (Mukherjee, 2009, p. 215).

References

Denke, A. (2005). Nativelike performance: A corpus study of pragmatic markers, repair and repetition in native and non-native English speech. (Doctoral dissertation), Stockholm University, Department of English, Stockholm.

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

Gráf, T. (2017). Repeats in advanced spoken English of learners with Czech as L1. *Auc Philologica*, 2017(3), 65–78.

Götz, S. (2013). Fluency in native and nonnative English speech. Amsterdam: John Benjamins.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language* 111(1), 36–45.

Molenda, M., Pęzik, P., & Osborne, J. (2018). Self-repetitions in learners' spoken language: A corpus-based study. In V. Brezina, & L. Flowerdew (Eds.), *Learner Corpus Research: New Perspectives and Applications* (pp. 90–11). London: Bloomsbury Academic.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 203–230). Amsterdam: John Benjamins.

A longitudinal study of L2 spoken English: Development of fluency and pronunciation

Mariko Abe, Yusuke Kondo, Yuichiro Kobayashi, Akira Murakami, Yasuhiro Fujiwara
Chuo University, Waseda University, Nihon University, University of Birmingham, Meijo University
abe.127@g.chuo-u.ac.jp, yusukekondo@waseda.jp, kobayashi0721@gmail.com, a.murakami@bham.ac.jp, fujiwara@meijo-u.ac.jp

In this talk, we will provide an overview of the research project for compiling the Longitudinal Corpus of L2 Spoken English (LOCSE) and will show the main outcomes of our preliminary corpus-based analyses of learners' development in fluency and pronunciation. Our study aims to fill the gaps in our field by constructing and analysing a longitudinal L2 spoken corpus that includes speech samples by low-proficiency learners.

The data were collected twice or thrice a year for three consecutive years from 2016, with the total of eight data collection points, and were elicited from a group of 122 students. To collect data, we employed the Telephone Standard Speaking Test (ALC Press, 2016), a monologic speaking test. Three certified raters gave a holistic score to each speech sample based on various criteria. In order to transcribe the learners' utterances, automated speech recognition (ASR) technology was employed. The sound files of learners' speech were transcribed by the ASR and then manually checked by three human transcribers to correct any transcription errors.

When the data were collapsed across the eight time points, the oral proficiency of the learners ranged from Novice-Mid to Intermediate-Mid based on the scale of American Council on the Teaching of Foreign Languages Oral Proficiency Interview. The holistic scores ranged across five levels out of an eight-point scale, and the learners' overall scores rose across time.

With this corpus, we can investigate both cross-sectional and longitudinal development of learners' spoken performance. For example, examining global linguistic complexity and fluency measures, we found that the average length of the text best indicated the oral proficiency of learners. The mean length of utterances and the word bigrams presumably representing syntactic patterns were also associated with the L2 learners' spoken performance.

Various disfluency features of the learners' utterances were tagged, including silent pause, filled pause, repetitions, and non-verbal sound. To investigate the development in fluency, we compared the keyness (i.e. log likelihood ratio) of the trigrams between the eight data collection points. The results showed that the number of words between disfluency markers gradually increased, which indicates that learners' speech fluency improved across time.

Another dimension of the spoken performance that can be investigated is pronunciation. We examined the relationship between the word error rates of the ASR and the oral proficiency of learners. The differences between the ASR transcriptions and the human-corrected version were examined, and we regraded the difference (i.e., ASR transcription errors) as erroneous pronunciation of learners. The results revealed that the transcription error rates decreased from the lowest to the highest level. Since the pronunciation model in the ASR was based on the native speaker's pronunciation, this implies that learners' pronunciation gradually became closer to that of native English speakers as the oral proficiency rises from the lower to upper level.

In conclusion, we were able to identify fluency development and pronunciation improvement by using a longitudinal L2 spoken corpus and to provide evidence that the error rates of the ASR could be a useful index of learners' pronunciation proficiency. Our study results can be applied to establish more appropriate assessments for EFL learners' oral performances.

References

ALC Press. (2016). *Telephone Standard Speaking Test (TSST).* https://tsst.alc.co.jp/biz/en/

Italian-Russian Learner Corpus: a Study of Possessive Constructions and a Linguistic Norm Inquiry

Ksenia Balakina
University of Bologna
ksenia.balakina2@unibo.it

The aim of the present study is to demonstrate that parallel learner corpora and specifically, collections of learners' translations into L2, can not only improve the inverse translation teaching process, but also contribute to the study of the actual linguistic norm of the L2.

The present study is based on the data of the first learner Italian-Russian corpus that collects translations from Italian into Russian, performed by Italian-speaking students who study the Russian language as L2 at university.

Translations from the parallel learner corpus highlight some typological differences between the Italian and Russian languages, since the source Italian verb *avere* (*'to have'*) is translated into Russian in various ways, namely, by the predicative construction *y X est' Y* (literally translated as '*at X there is Y*') or by the transitive verb *imet'* (*'to have'*). Importantly, it is the predicative construction that represents the basic model for the expression of possession in the Russian language, whereas the transitive verb has some stylistic limitations and is mostly used with inanimate possessors (Isačenko, 1974; Ivanov, 1989: 176). So, the object of our study is to examine the use of the above mentioned possessive constructions in students' translations on one hand and to investigate the actual linguistic norm that governs the use of such constructions in the modern Russian language, on the other.

Before proceeding to the learner corpus data analysis, a brief overview of the previous study of possessive constructions in the Russian language is performed. According to some linguists, the use of the predicative construction is allowed only if the possessor is expressed by an animate noun (Ivanov, 1989: 176). Many studies, on the contrary, affirm that inanimate and abstract nouns should be followed by the transitive verb *imet'* (Adamec, 1960: 213; Činčlej, 1996: 107; Guiraud-Weber, Mikaelian, 2004: 65; Rakhilina, Weiss, 2002: 178), which can lead us to the conclusion that the two possessive constructions have the complementary distribution on the basis of the animacy of the possessor.

In order to verify the last affirmation, the author consults the reference data from the National Corpus of Russian Language. The analysis has shown that the use of the predicative possessive construction with inanimate possessors is much more widespread than one could expect, considering the previous studies summary.

The subsequent analysis of the learner corpus data showed that translation of possessive relations creates difficulties for the students who tend to choose between alternative target constructions in a totally chaotic way. Learner translations confirmed as well, that the use of the predicative construction with inanimate subjects is quite frequent and that its distribution over various context types corresponds to the tendency revealed in the National Corpus.

In the final part of the study, the author presents the preliminary results of the inquiry held among Russian mother-tongue speakers. It aims at registering the actual use of the possessive constructions and their alternatives in the modern Russian language and identifying the tendencies of an eventual linguistic norm fluctuation.

To conclude, the author discusses the spheres of application of the results of parallel learner corpora research, that allows not only to better comprehend specific characteristics of students' interlanguage, but also to improve the description of the actual rules of a language. The present study demonstrates, indeed, that it is the data of the learner corpus that casts doubts on the grammaticality of some possessive constructions' use and boosts inquiry of the confines of the linguistic norm in the modern Russian language.

References

Adamec, P. (1960). *K ekvivalentům sloves býti a míti v ruštině. Ruskočeské studie*, 191–213.

Benveniste, E. (1960/1966). "Etre" et "avoir" dans leurs fonctions linguistiques. In Benveniste, E. (1966) *Problèmes de linguistique générale*, Paris.

Činčlej, K.G. (1996). Pole posessivnosti i possesivnye situacii. Teorija funkcional'noj grammatiki. Lokativnost'. Bytijnost'. Posessivnost'. Obuslovlennost', 100–117. St. Petersburg: Nauka.

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

Granger, S. (2017). Learner Corpora in Foreign Language Education. Language, Education and Technology, 1–14.

Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger *Learner English on Computer* (pp. 3–18). London-New York.

Granger S., Tribble C. (1998). Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In Granger, S. (Ed.) *Learner English on Computer* (pp. 199–209). London/New York.

Guiraud-Weber M., Mikaelian I. (2004). V zaščitu glagola imet‘ In Arutjunovoj, N. D. (Ed.) *Sokrovennye smysly, slovo, tekst, kul’tura: sbornik statej v čest‘* (pp. 54–68). Moscow.

Isačenko, A. V. (1974). On ‘have’and ‘be’languages. A typological sketch. In *Slavic forum: Essays in linguistics and literature* (pp. 43–77). The Hague/Paris: Mouton.

Ivanov, V. (1989). *Kategorija posessivnosti v slavjanskih i balkanskih jazykah*. Moscow.

Kobozeva, I. M. (2015). O posessivnosti v russkom jazyke: posessivnye predikaty vs. genitiv. *Acta Linguistica Petropolitana. Trudy instituta lingvističeskih issledovanij*, 11(1).

Lado, R. (1957). Linguistics across cultures: applied linguistics for language teachers. Michigan.

Rakhilina, E.V. (2016). O novyh instrumentah opisanija russkoj grammatiki: korpus ošibok. *Russkij jazyk za rubežom*, (3), 20–25.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3).

Seliverstova, O.N. (1973). Semantičeskij analiz predikativnyh pritjažatel’nyh konstrukcij s glagolom „byt“. *Voprosy jazykoznanija* 5, 95–105.

Weiss, D., Rakhilina, E. (2002). Forgetting one's roots: Slavic and Non-Slavic elements in possessive constructions of modern Russian. *STUF–Language Typology and Universals* 55(2), 173–205.

Learner corpus-based lists and their validity: language use changes, lists remain

Bartosz Brzoza
Adam Mickiewicz University
bbrzoza@wa.amu.edu.pl

Frequency-based word lists produced by the publishers of didactic materials such as a learner Oxford 3000™ list or Coxhead's (2000) academic word list are long-known and frequently used in applied research and teaching practice. Using such didactic resources with students is particularly conducive to vocabulary learning (Cobb & Boulton, 2015). This conclusion can be derived from the evidence that language acquisition studies provide. According to many such studies, frequent words are easier learnt and processed by learners, as they correlate with the natural order of language acquisition. However, language use evolves dynamically and so the frequency with which people use certain lexical items concomitantly changes (e.g. Gries & Divjak, 2012). It seems uncontroversial then that there is some recent evidence (Brzoza, 2018) showing that corpus-based objective measures might not correlate as much with subjective frequency ratings provided by native speakers as they used to in previous research (Ringeling, 1984; Desrochers & Bergeron, 2000; Thompson & Desrochers, 2009).

The current contribution reports on the results of a study conducted to observe the relationship between the objective corpus-based and subjective participants-provided measures of word frequency at different times of measurement. Its aim is to investigate whether these results might have a bearing on the construction of frequency-based learner vocabulary lists, particularly whether these should be updated. This issue is up-to-date, as the existing lists of frequent vocabulary items do not evolve, and the new ones do not get published.

The study consisted of the comparison of the frequency values of used words during the time of constructing Oxford 3000™ and Coxhead's (2000) academic word list, and in a recent SUBTLEX-UK corpus (Van Heuven et al., 2014). The results of the correlational analyses between the frequency counts at these various measurements will be juxtaposed with the frequency ratings performed by native speakers of English in an online questionnaire. The hypothesis posits that there are discrepancies between then- and current frequency values of selected words from the lists. Another hypothesized relationship between scores is that the relationships between corpus-based and subjective participants' judgements of the same words is weaker for the past corpus values than for the current corpus values.

The present investigation considers the need for adjustment of the existing vocabulary lists. The results will be discussed in the light of changing lexical frequency values. I will offer some new alternatives or additions to corpus-driven frequency values for constructing frequency-based lists, such as combining objective with subjective frequency measures or turning to some new measures of wordhood dimensions, e.g. word prevalence (Brysbaert et al., 2016).

References

Brzoza, B. (2018). Word frequency counts: Linking corpus data to user's perception in linguistic research. *Lingvisticae Investigationes* 41(2), 224–239. doi:10.1075/li.00021.brz

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42, 441–458.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber, & R. Reppen (Eds.) *The Cambridge handbook of English corpus linguistics*, 478–497. https://doi.org/10.1017/CBO9781139764377.027

Coxhead, A. (2000). A new academic word list. *TESOL quarterly* 34(2), 213–238. https://doi.org/10.2307/3587951

Desrochers, A., & Bergeron, M. (2000). Valeurs de frequence subjective et d'imagerie pour un echantillon de 1,916 substantifs de la langue francaise. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 54(4), 274–325. https://doi.org/10.1037/h0087347

Gries, S. T. & Divjak, D. *Frequency effects in language learning and processing* 1. Berlin: De Gruyter. https://doi.org/10.1515/9783110274059

Oxford 3000™ Word List. Oxford University Press.

Ringeling, T. (1984). Subjective estimates as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin* 8(1), 59–69.

Thompson, G., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behaviour Research Methods* 41(2), 452–471. https://doi.org/10.3758/BRM.41.2.452

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M., (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Code-switching in multicompetent speakers:
A corpus study of German and Spanish EFL learners' writing at beginning and intermediate levels

Marcus Callies, María Belén Díez-Bedmar
Universität Bremen, Universidad de Jaén
callies@uni-bremen.de, belendb@ujaen.es

Code-switching is one of the most salient phenomena in second language production. Defined as the effortless alternation between two or more languages, it can extend from the insertion of single words to the alternation of languages for larger segments of discourse, as conceptualized in the difference between code switching (CS) and code mixing (CM) respectively (Ritchie & Bhatia, 2013). Several reasons have been put forward to explain the frequent occurrence of CS, such as filling linguistic gaps, expressing ethnic identity, or achieving particular discourse aims (Bullock & Toribio, 2009). Advanced L2 learners have been found to use CS as a special type of communication strategy to fill lexical gaps and for self-repair (e.g. Liebscher & Dailey-O'Cain, 2005; Nacey & Graedler, 2013; DeCock, 2015). However, despite the pervasiveness of CS and its importance for the study of SLA, studies based on learner corpora are still rare, partially because instances of CS are not regularly annotated in such corpora (Callies & Wiemeyer, 2017).

To fill this gap, this paper has two main objectives: a) to argue for the systematic annotation of CS in learner corpora by discussing its pros and cons, and b) to report on the results obtained from a quasi-longitudinal study on the use of CS by L1 German and Spanish beginning to intermediate EFL learners. The corpus data are taken from the *International Corpus of Crosslinguistic Interlanguage* (ICCI) (Tono & Díez-Bedmar, 2014) that mostly includes short narrative/descriptive texts written on a variety of topics, e.g. film, food, and money. The findings suggest that in the written corpus texts, the learners use intra-sentential CS with referential function to fill lexical gaps. The frequency of CS is generally much higher in texts written by Spanish EFL learners when compared to German EFL learners. We also observe largely similar developmental trends in both learner groups in terms of a decrease of CS with increasing proficiency.

References

Bullock, B. E., & Toribio, A. J. (2009). Themes in the study of code-switching. In B. E. Bullock, & A. J. Toribio (Eds.), *The Cambridge handbook of linguistic code-switching* (pp. 1–17). Cambridge: Cambridge University Press.

Callies, M., & Wiemeyer, L. (2017). Multilingual speakers, multilingual texts: The annotation of foreign elements in learner corpora of English. In A. Nurmi, T. Rütten, & P. Pahta (Eds.), *Are there monolingual corpora? Challenging the myth* (pp. 80–97). Amsterdam: Brill.

De Cock, S. (2015). *Foreign words in interviews with EFL learners: Bridging lexical gaps?* Poster presented at ICAME 35, Trier, Germany.

Liebscher, G., & J. Dailey-O'Cain (2005). Learner code-switching in the content-based foreign language classroom. *The Modern Language Journal* 89(2), 234–247.

Nacey, S., & Graedler, A.-L. (2013). Communication strategies used by Norwegian students of English. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 345–356). Louvain-la-Neuve: Presses universitaires de Louvain.

Ritchie, W. C. & Bhatia, T. K. (2013). Social and psychological factors in language mixing. In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism and multilingualism.* 2nd ed. (pp. 336–252). Chichester: Wiley.

Tono, Y. & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages. The ICCI corpus. *International Journal of Corpus Linguistics* 19(2), 163–177.

Are cross-linguistic influences underrated in ELF research?
The case of particle placement in multi-participant interactions

Sandra C. Deshors
Michigan State University
deshorss@msu.edu

This multifactorial corpus-based study on particle placement in English as a Lingua Franca (ELF) focuses on verb-object-particle (VOP) and verb-particle-object (VPO) alternations and identifies the usage patterns that characterize these alternations in ELF multi-participant interactions. Verb-particle constructions (VPC) have been approached differently across Englishes. In native English, studies have shown how all linguistic levels and cognitive processes of language acquisition influence the uses of VOP/VPO constructions. Similarly, in learner English (EFL), the uses of the two constructions have been shown to deviate from those in native English because of their syntactic and semantic complexity, processing demands, input effects, and the typology of speakers' L1. Further, analytically, VPC are explored semantically without considering VPO/VOP alternations nor the combined effects of linguistic and processing factors. Methodologically, quantitative studies still remain monofactorial and are not geared to handle the complexity of analyzing VPC based on multiple predictors, which is an important limitation since linguistic and extra-linguistic (i.e. socio-demographic and conversational) factors affect, independently, the structure of ELF. Approaching ELF theoretically as a complex adaptive system (CAS), the present study (i) makes a stronger connection between theory and method in ELF research, (ii) identifies the combinations of linguistic and extra-linguistic/conversational factors that influence constructional choices and (iii) aligns current ELF and EFL methodological approaches to VPC.

These points are addressed by investigating 585 VOP/VPO constructions from the Vienna-Oxford International Corpus of English (VOICE) annotated against individual phrasal verbs, length, determiner, complexity and type of direct object, concreteness of the referent of direct object, semantic use of the verb phrase, sex and age range of the speaker, his/her L1 and conversational role. These factors were analyzed statistically with regression modeling and stratified sampling random forests, conceptually compatible with the CAS framework. The approach involves a random forests analysis including a surrogate logistic regression model and interactions between linguistic, socio-demographic and conversational factors. I obtained models/trees with significant classification accuracies and *C*-scores and identified strongest predictors of alternations such as TYPE, COMPLEX, DET and LENGTH. Based on the global surrogate model, these factors also participate in several interactions that drive VOP/VPO alternations.

Overall, these interactions underscore the importance of integrating extra-linguistic factors to large-scale quantitative linguistic analyses of the structure of ELF and the usefulness of combining a CAS framework and a multifactorial methodology. Syntactically, the results indicate that as in native English and EFL, the type of direct object, the type of determiner and the degree of complexity of the direct object all play a significant part in the VOP/VPO alternation, suggesting that VOICE speakers are constrained by the same factors as native/EFL speakers. However, the influence of these factors emerges as a function of other co-occurring extra-linguistic factors such as the speakers' native language and conversational role. Overall, these results revive the question of what distinguishes ELF from EFL as linguistic systems and open the door for a discussion on bilingual processing in ELF, based on the significant influence of speakers' L1. Ultimately, these results urge us to pay close attention to the potential role of cross-linguistic influence in processing in ELF mode and invite us to reconsider the place of L1 typology in ELF research.

Development of L2 writing complexity:
Clause types, L1 influence and individual differences

Hildegunn Dirdal
University of Oslo
hildegunn.dirdal@ilos.uio.no

Structural complexity research has often focussed on global measures of subordination and attempted to find universal trends and benchmarks for developmental stages. Recently, researchers have pointed out the need to distinguish between different types of subordinate clauses as they may have different developmental trajectories (Lambert & Kormos, 2014, Vercellotti & Packer, 2016) and to take into account L1 influence (Lu & Ai, 2015, Ortega, 2015) and individual variation (Larsen-Freeman, 2009). The novelty of the present study consists in combining all these factors and making a finer distinction between clause types than what is normal in complexity research. The data come from texts collected for the TRAWL Corpus, a longitudinal corpus currently under construction, containing L2 texts written by Norwegian school children. A subset of the students have also contributed texts in their L1 Norwegian. The study addresses the following questions:

- How does Norwegian learners' L2 English writing complexity develop over time with respect to different types of subordinate clauses?
- Can L1 effects be detected in comparison with L1 English writers of similar ages?
- Are there individual differences, and can these be connected to differences in the learners' L1 writing?

The study focuses on five students who have contributed texts from lower secondary school and the first year of upper secondary school (age 13–17). The English and Norwegian texts produced by these students were manually coded for clause types and their syntactic function. A case study of five focal students allows for a detailed comparison of individual differences. To address the question of L1 influence, the data are compared with L1 English texts produced by writers of a similar age from the Growth in Grammar Corpus (Durrant & Brenchley, 2018). The authors have kindly made available the annotated part of the corpus, which contains information about clause types and functions.

A mean total of 1787 English clauses were coded for each learner (range: 1565–2010). The learners use a wide range of subordinate clauses when starting lower secondary school, although *wh*-clauses, *-ing* clauses and past participle clauses are less frequent than *that*-clauses, relative clauses, adverbial clauses and infinitive clauses. The use of all kinds of subordinate clauses increases over time, but there is a reduction in the use of *that*-clauses over the last year. Compared with the L1 writers in the Growth in Grammar Corpus, the L2 learners have a lower frequency of subordination overall, but the difference is most pronounced for *-ing* clauses, possibly due to cross-linguistic influence from Norwegian, which lacks a proper equivalent.

All five learners increase their use of subordination over time, but there are individual differences in the kinds of clauses involved. For example, one student has her largest increase in number of relative clauses, whereas another has a more even increase for all types of clauses in addition to being one of two students to show a more marked increase of *-ing* clauses in the final year.

The Norwegian L1 data will be used to determine whether the learners' lower subordination rate in English (compared to L1 English writers of similar ages) is due to influence from their L1 or an effect of lagging behind in the L2. The Norwegian data will also be used to examine whether individual differences in the L2 English data are due to individual preferences across languages.

The results confirm the need to take into account both L1 influence and individual differences in complexity research and for distinguishing between different types of subordinate clauses. They contribute to a more detailed picture of the development of L2 complexity that will also be of value for teachers and the developers of teaching materials.

References

Durrant, P., & Brenchley, M. (2018). *Growth in Grammar Corpus*. www.gigcorpus.com.

Lambert, C., & Kormos, J. (2014). Complexity, Accuracy, and Fluency in Task-based L2 Research: Toward More Developmentally Based Measures of Second Language Acquisition. *Applied Linguistics* 35(5), 607–614.

Larsen-Freeman, D. (2009). Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 579–589.

Lu, X., & Ai, H. (2015). Syntactic Complexity in College-level English Writing: Differences among Writers with Diverse L1 Backgrounds. *Journal of Second Language Writing* 29, 16–27.

Ortega, L. (2015). Syntactic Complexity in L2 Writing: Progress and Expansion. *Journal of Second Language Writing*, 29, 82–94.

Vercellotti, M. L., & Packer, J. (2016). Shifting Structural Complexity: The Production of Clause Types in Speeches Given by English for Academic Purposes Students. *Journal of English for Academic Purposes* 22, 179–190.

Indonesian EFL Learners' Argumentative Writing:
A Learner Corpus Study of Connector Usage

Nida Dusturia
University of Bremen
dusturia@uni-bremen.de

The use of connectors (a.k.a. linking adverbials, see Biber et.al. 1999: 875) has been found to be challenging for EFL learners (see e.g. Bolton & Nelson, 2002; Chen, 2006). Granger & Tyson (1996) conducted research on connector usage in the writing of native and non-native speakers of English and report over-, under, and misuse of some connectors. Several other studies (Crewe, 1990; Field & Yip, 1992; Martinez, 2004; Chen, 2006; Heino, 2010) obtain similar findings for both ESL and EFL learners in that they have problems in the use of conjunctions. Such findings mentioned have also been observed for Indonesian EFL learners (Swan & Smith, 2001; Ishak, 2002; Moehkardi, 2002; Marzuki & Zainal, 2004; Kurniyati, 2012; Antara, 2015). However, corpus data for Indonesian EFL learners is not widely available. This study is intended to fill this gap and examine Indonesian EFL learners' argumentative writing from a learner corpus perspective.

This research aims at investigating the use of connectors by Indonesian EFL learners in argumentative texts written at different proficiency levels (A.2. and B.1.2 of the *Common European Framework of Reference for Languages* (CEFR; Council of Europe, 2001) and native speakers of English. The data come from the *International Corpus Network of Asian Learners of English* (ICNALE; Ishikawa, 2013). The ICNALE includes essays based on two topics: "It is important for college learners to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country". The component produced by Indonesian EFL writers consists of 93.277 word tokens. In the compilation process, writing time, text length, and other conditions were controlled as strictly as possible, which leads to greater reliability in varied types of contrastive analyses.

The research questions addressed in this study are:

1. What semantic types of connectors are used in argumentative essay writing by EFL learners from Indonesia?
2. Do Indonesian EFL learners at different proficiency levels differ in the use of connectors in their argumentative essays in terms of frequency and semantic types?
3. How do Indonesian EFL learners and English native speakers compare as to over-/underrepresentation and misuse of connectors in argumentative essay writing?

The method used in this study is Contrastive Interlanguage Analysis (CIA; Granger, 2015) which involves two types of comparison. First, a comparison between the use of connectors in argumentative essays produced by Indonesian EFL learners (ICNALE_IDN) at the A.2. and the B.1.2 levels of the CEFR; and second, a comparison of connector usage in argumentative essay produced by Indonesian EFL learners and English native speakers (ICNALE_ENS). For the analysis, the connectors are classified into various semantic types according to their discourse function(s), such as Enumeration/Addition, Summation, Apposition, Result/Inference, Contrast/Concession, and Transition (Biber et. al., 1999). The annotation is carried out by means of *UAM Corpus Tool* (O'Donnell, 2015). When analyzing the data, quantitative and qualitative approaches are combined. The quantitative approach is used to examine potential over- and underrepresentation of connectors, while the qualitative approach is used for investigating potential misuses of connectors.

The quantitative results indicate that differences can be found in the use of connectors by Indonesian learners when compared to native speakers. The Indonesian learners of English at the A.2 level tended to use more contrastive and resultative connectors as the native speakers did, whereas the learners at the B.1.2 level more frequently used the additive and appositive types. The distribution of the different semantic categories was nearly identical in the Indonesian and the native-speaker data. Contrastive connectors was most frequently used, followed by the resultative, additive and appositive types. Additionally, Indonesian learners at both proficiency levels demonstrate misuse in the connector usage compare to the native speaker as shown in the following example which illustrates an unmotivated, non-target like use of the contrastive connector *on the other hand:*

1) At least we can help our parents with having a part - time job. *On the other hand* we can practice our skills too before we really do a job after we graduate. (IDN PTJ B12 066)

References

Antara, I Made. (2015). Keterampilan Menulis Wacana Argumentasi Berbahasa Inggris Dengan Metode Esa Pada Mahasiswa Stie Triatma Mulya Level Post Intermediate. Denpasar: Universitas Udayana.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. Essex: Pearson Education.

Bolton, K., & Nelson, G. (2002). Analyzing Hong Kong English. Sample texts from the International Corpus of English. In K. Bolton (ed.): *Hong Kong English. Autonomy and Creativity*. Hong Kong: Hong Kong University Press, 241–264.

Chen, C. W. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics* 11(1), 113–130.

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

Crewe, W. J. (1990). The Ilogical of logic connectives. *ELT Journal*, 44(4), 316–325.

Field, Y., & Yip Lee Mee, O. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal* 23(1), 15–28.

Granger S. (2015). Contrastive Interlanguage Analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7–24, Amsterdam and Philadelphia: John Benjamins.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15(1), 17–27.

Heino, P. (2010). Adverbial Connectors in Advanced EFL Learners' and Native Speakers' Student Writing. Bachelor degree project. Stockholm University.

Ishak, Abdulhak dkk. (2002). *Perencanaan Pengajaran Unit Pelaksanaan Teknis Program Pengalaman Lapangan*. Bandung: STKIP.

Ishikawa, S. (2013). *The International Corpus Network of Asian Learners of English.* http://language.sakura.ne.jp/icnale/index.html

Ishikawa, S. (2013). The ICNALE and sophisticated Contrastive Interlanguage Analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner Corpus Studies in Asia and the World* 1. Kobe: Kobe University Press, 91–118.

Kurniyati, D. (2012). Analisis Kesalahan Kohesi Dan Koherensi Paragraf Pada Karangan Siswa Kelas X Sma Negeri 3 Temanggung, *Universitas Negeri Yogyakarta* 1(2). http://eprints.uny.ac.id

Martinez, A. C. L. (2004). Discourse markers in the expository writing of Spanish university students. *IBERICA* 8, 63–80.

Marzuki, S., & Zainal, Z. (2004) *Common Errors Produced by UTM Students in Report Writing*. Malaysia: University Teknologi Malaysia.

Moehkardi, D. (2002). Grammatical and lexical English collocations: some possible problems to Indonesian learners of English, *Jurnal Humaniora* 14 (1), 53–62.

O'Donnell, M. (2015). *UAM Corpus Tool*. Version 3.1.17. Available from http://www.wagsoft.com/CorpusTool/index.html.

Swan, M., & Smith, B. (2001). Learner English: A Teacher's Guide to Interference and Other Problems. Cambridge: Cambridge University Press.

Tense and aspect in learner English: Frequency vs. accuracy

Robert Fuchs, Valentin Werner
University of Hamburg, University of Bamberg
robert.fuchs.dd@gmail.com, valentin.werner@uni-bamberg.de

This paper provides an extension of previous work (e.g. Fuchs, Götz, & Werner, 2016; Werner & Fuchs, 2017; Deshors, 2018; Götz, Werner, & Fuchs, forthcoming; Werner, Fuchs, & Götz, forthcoming) on established SLA principles relating to the acquisition of tense and aspect (TA) expressed through morphosyntactic means. Specifically, we consider (i) the order of acquisition of tense and aspect (OATA) and (ii) the Default Past Tense Hypothesis (DPTH), which to date both have largely been tested experimentally in smaller learner groups (cf. Bardovi-Harlig, 2000) but not on larger bodies of data.

Proponents of the OATA (see, e.g., Bardovi-Harlig, 2000; Svalberg, 2018) agree on an emergence of TA forms in learner English along the following lines: simple present/present progressive → simple past/past progressive → present perfect → present perfect progressive → past perfect → past perfect progressive. Proponents of the DPTH (e.g. Salaberry & Ayoun, 2005) predict that learners in early- intermediate stages will use a single morphological marker for past-time reference, which for English is the simple past.

Previous work tracked the emergence of TA forms as a function of frequency of usage and found that patterns for EFL learners largely are in accordance with both the OATA and the DPTH (Werner & Fuchs, 2017; cf. Collins, 2002). In the present paper, we enrich this perspective with accuracy ratings and error annotations to explore whether and to what extent an increase in the frequency of usage corresponds to an increase in accuracy.

To this end, we use a quasi-longitudinal research design, and rely on the *International Corpus of Crosslinguistic Interlanguage* (ICCI; Tono & Díez-Bedmar, 2014) and the *International Corpus of Learner English* (ICLE; Granger et al., 2009) to assess tense-aspect acquisition in (tutored) learner writing from the beginning to the advanced level in four typologically different L1 language backgrounds (Germanic: German, Sinitic: Chinese, Slavic: Polish, Romance: Spanish). A subsample (50 occurrences per TA form, learner L1 and grade) was rated by two native speakers and disagreements were resolved by a third rater.

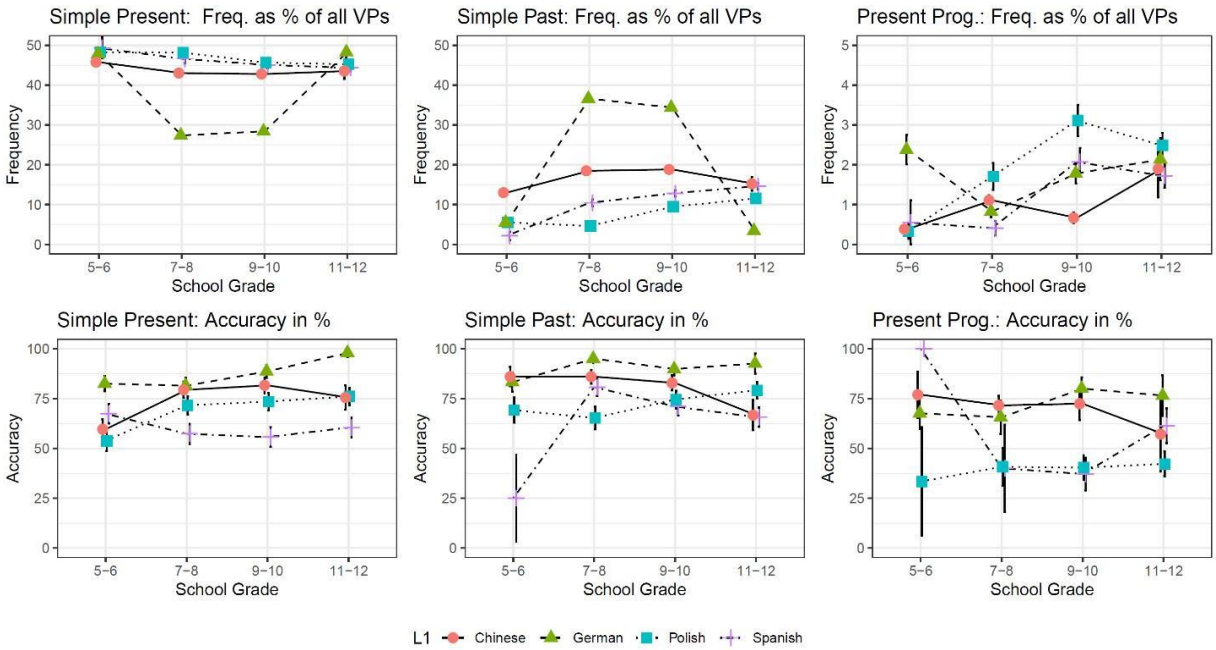


Fig. 1. Frequency and accuracy of usage of selected TA forms (note difference in scale in top right panel)

Results indicate that accuracy of usage does not linearly increase with frequency of usage or proficiency. As Fig. 1 shows for learners at school level (ICCI data), accuracy (correct usages as a percentage of overall frequency) increases in the use of the simple present for L1 German, Polish and Chinese learners, but at different times in the learning process. The usage of the simple past simultaneously becomes more frequent and more accurate for

L1 Polish learners, but not the other groups. For low-frequency TA forms such as the present progressive, accuracy may even drop for some groups (here: L1 Chinese) as frequency increases.

However, in the larger picture, the analysis of accuracy of usage again confirms the predictions of the OATA, as did the analysis of frequency of usage in previous work (see above). Findings indicate that simple forms are used (i) earlier and more frequently and (ii) more accurately than complex forms at any stage in the acquisition process. In the present paper, we will further enrich this analysis by investigating (i) accuracy of usage in terms of "false negatives" (e.g. using a present simple where a present progressive is required) and (ii) particular error types (functional errors – i.e. confusion of TA forms – and formal errors – e.g. omission of 3rd person singular *-s* in the present).

In light of the findings, we argue for a nuanced view of the interaction between frequency and accuracy of usage instead of a simple correlational pattern. The delayed target-like form-function mapping observable for less frequent and structurally more complex TA forms further has practical implications as it is an issue that could be explicitly or implicitly addressed in EFL education. From a methodological perspective, we further suggest that learner corpus research benefits from using a quasi-longitudinal design, using both meta-data and manual annotation to exploit its potential, informing the broader domain of SLA studies (e.g. as regards the accountability of stage/acquisition order models; cf. Hulstijn, Ellis, & Eskildsen, 2015) as well as investigations of tense and aspect and its development in learner language in particular (see contributions to Fuchs & Werner, 2018).

References

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. Malden: Blackwell.

Collins, L. (2002). The roles of L1 influence and lexical aspect in the acquisition of temporal morphology. *Language Learning* 52(1), 43–94.

Deshors, S. (2018). Does the passé composé influence L2 learners' use of English past tenses? *International Journal of Learner Corpus Research* 4(1), 23–53.

Fuchs, R., Götz, S., & Werner, V. (2016). The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In V. Werner, E. Seoane, & C. Suárez-Gómez (Eds.), *Re-assessing the present perfect* (pp. 297–337). Berlin: Mouton de Gruyter.

Fuchs, R, & Werner, V. (2018). *Tense and aspect in second language acquisition and learner corpus research*. Special issue of the *International Journal of Learner Corpus Research* 4(2).

Götz, S., Werner, V, & Fuchs, R. (Forthcoming). Temporal adverbials in the acquisition of past-time reference: A cross-sectional study of L1 German and Cantonese learners of English. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicholas (Eds.), *Widening the scope of learner corpus research*. Louvain-la-Neuve: Presses universitaires de Louvain.

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

Hulstijn, J. H., Ellis, R., & Eskildsen, S. W. (2015). Orders and sequences in the acquisition of L2 morphosyntax, 40 years on. *Language Learning* 65(1), 1–5.

Salaberry, M. R., & Ayoun, D. (2005). The development of L2 tense-aspect in the Romance languages. In D. Ayoun & M. R. Salaberry (Eds.), *Tense and aspect in Romance languages* (pp. 1–33). San Diego: Academic Press.

Svalberg, A. M. L. (2018). Mapping tense form and meaning for L2 learning – from theory to practice. *International Review of Applied Linguistics in Language Teaching.* doi.org/10.1515/iral-2016-0105

Tono, Y., & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics* 19(2), 163–177.

Werner, V., & Fuchs, R. (2017). Acquisition of tense and aspect in learner English: A corpus-based cross-sectional perspective. Paper presented at the *4th Learner Corpus Research Conference*. Bolzano.

Werner, V., Fuchs, R., & Götz, S. (Forthcoming). L1 influence vs. universal learning mechanisms: An SLA-driven corpus study on temporal expression. In B. Le Bruyn, & M. Paquot (Eds.), *Second language acquisition and learner corpora.* Cambridge: Cambridge University Press.

Investigating the scope of textual metrics for learner level discrimination and learner analytics

Thomas Gaillat, Nicolas Ballier
University of Rennes, University of Paris-Diderot
thomas.gaillat@univ-rennes1.fr, nicolas.ballier@univ-paris-diderot.fr

This paper focuses on textual metrics that can be used in ICALL systems as criterial features. Empirical research approaches to learner corpora include the identification of criterial features linked to learners' proficiency levels. Numerous metrics have been developed to measure lexical sophistication (Kyle, Crossley, and Berger 2018), readability (Francois 2011) and syntactic complexity (Lu 2010, 2014). With a view to developing ICALL systems aimed at giving feedback on the level of proficiency, it is necessary to identify which metrics are significant to discriminate learners at a given level (Crossley et al. 2011; Hawkins and Filipović 2012; Arnold et al. 2018; Pilán and Volodina 2018; Kim and Crossley 2018; Khushik and Huhta 2019). However, the metrics need to be self-intuitive for learners in their meta-cognitive learning processes. For that purpose, they should be interpreted in terms of scope rather than functionality e.g. readability or lexical diversity.

Our research question is to investigate the significance of a scope-oriented typology of metrics. We propose a typology in which metrics are related to constituents, from syllables to text grammar. Depending on their formula, metrics rely on different types of frequencies and have syllable, word, clause and sentence scopes. Our purpose is to investigate how metrics of different scopes correlate with different proficiency levels.

We have followed a modelling approach in which we test metrics of different scopes in relation to the scores obtained by students at the DIALANG test (Alderson and Huhta 2005), as a proxy to the levels of the Common European Framework of Reference (CEFR). We put to the test the typology with the classification of 272 French learners of English. Essays were collected through MOODLE (Dougiamas and Taylor 2003), resulting in 86,000 tokens. Data was processed with the {quanteda} R package (Benoit et al. 2018). The resulting dataset was made up of lexical diversity, readability and syntactic complexity features. For the modelling method we used the {randomForests} package (Liaw and Wiener 2002) with default parameters (ntree=500 and mtry=6) to discriminate which metrics could best predict the level of learners. We divided the dataset in training (80% of the data, randomly collected) and test (20%) sets. Evaluation was conducted on the test set.

Preliminary investigations show mitigated results with a mean accuracy of 55.35% across the six classes on the test set. When classifying according to three aggregated A, B and C levels, accuracy is 75% with most confusion between A and B levels (see Figure 1).

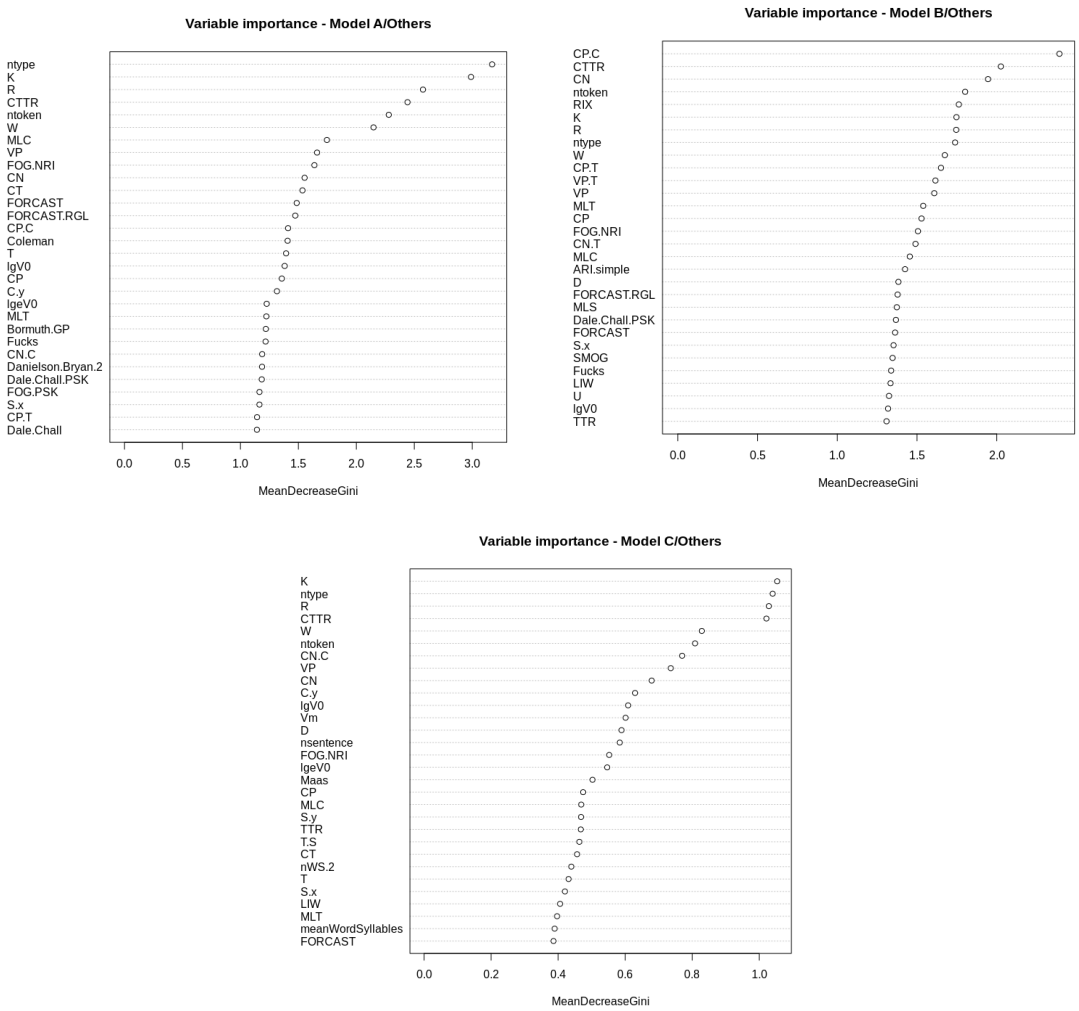
		True classes		
		A	B	C
Predicted classes	A	7	5	0
	B	7	34	2
	C	0	0	1

Figure 1: Confusion matrix for the three aggregated CEFR levels in the test set

We conducted model explanation by extracting important variables with the Gini Index measure, averaging the loss of impurity of the data linked to specific variables. Yule's K, R (Guiraud's root TTR), ntype (number of types), CTTR (Carrol's Type Token Ratio) and ntoken (number of tokens) metrics are reported to have the highest level of importance. These metrics have a text rather than a sentence scope as their formulae rely on type and token frequencies within entire texts. In our models, metrics with a text scope appear to provide more significant information for decisions. Conversely, sentence level metrics come second in the ranking, which includes a series of syntactic complexity metrics relying on clauses.

We also investigated the correlation between variables and specific CEFR levels. We used a binary iterative approach in which data points of each level are modelled against the others. Results show that the B level is impacted by sentence-scope variables as it highly correlates with Coordinate Phrases per Clause (CP.C). This trend is also observed in A-level learners, albeit to a lesser extent. It is absent among C-level learners (see Figure 2). This may suggest that B1 and B2 learners make disproportionate uses of coordinating conjunctions in their

writings, indicating a strong preference for parataxis. As they complexify their speech, learners might try to be as



informative as possible within each clause.

Figure 2: Importance of variables depending on CEFR levels

Identifying important metrics contributes to the research on learner language criterial features. Interpreting them within the framework of a learner-intuitive typology can be used in visualisation techniques aimed at learners in ICALL systems. We hope to guide learners with scope-focused advice suggesting confidence intervals as "targets" for their productions.

References

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22(3), 301–320.

Arnold, T., Ballier, N., Gaillat, T., & Lissòn, P. (2018). *Predicting CEFRL levels in learner English on the basis of metrics and full texts*. ArXiv:1806.11099 [Cs].

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing* 28(4), 561–580.

Dougiamas, M., & Taylor, P. (2003). Moodle: Using Learning Communities to Create an Open Source Course Management System. Proceedings of the EDMEDIA 2003 Conference, Honolulu, Hawaii, 171–178.

Francois, T. (2011). *Les apports du traitement automatique du langage à la lisibilité du francais langue étrangère*. Université Catholique de Louvain, Louvain.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. United Kingdom: Cambridge University Press.

Khushik, G. A., & Huhta, A. (2019). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, amy064.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing* 37, 39–56.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods* 50(3), 1030–1046.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.

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

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Dordrecht: Springer.

Pilán, I., & Volodina, E. (2018). Investigating the importance of linguistic complexity features across different datasets related to language learning. *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, 49–58.

Fluency in Advanced Spoken Learner Language:
A Contrastive Interlanguage Analysis across L1s, Task Types and Learning Context Variables

Sandra Götz, Christoph Wolk, Katja Jäschke
Justus Liebig University Giessen
Sandra.Goetz@anglistik.uni-giessen.de, Christoph.B.Wolk@anglistik.uni-giessen.de,
Katja.Jaeschke@anglistik.uni-giessen.de

Planning pressure in spontaneous speech production is naturally very high when speaking in a foreign language. Strategies to overcome such planning phases (or "fluencemes"; Götz 2013) include the use of filled or unfilled pauses (e.g. eh, ehm, er, erm; e.g. Gilquin 2008; Götz 2013), discourse markers (e.g. you know, like, well; e.g. Müller 2005; Crible 2018) or smallwords (e.g. sort of, kind of; Hasselgren 2002). Previous (learner corpus) research investigating learners' use of such fluencemes revealed that even advanced learners heavily underuse discourse markers and show a tendency of using filled or unfilled pauses instead (e.g. Gilquin 2008; Götz 2013; Dumont 2017). However, previous corpus-based research on fluency has mainly focused on one learner variety in particular, whereas contrastive interlanguage analyses on learners' use of fluencemes from different L1 family backgrounds have only rarely been undertaken. On top of that, the effect of task types and learning context variables on learner fluency is also just beginning to be investigated (e.g. Dumont 2017; Crible 2018; Götz 2019). In order to systematically test if learners from different L1 backgrounds use fluency-enhancing strategies differently (both, from each other and from native speakers), we would like to present a "Contrastive Interlanguage Analysis 2.0" (Granger 2015) that investigates four types of fluencemes (viz. filled and unfilled pauses, discourse markers and smallwords) in four components of the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010). Each subcorpus contains interviews with advanced learners of English from four different language backgrounds (i.e. German, Japanese, Bulgarian and Spanish), which we compared to the Louvain Corpus of Native English Conversations (LOCNEC; DeCock 2004). In our study, we analyze these corpora in order to answer four research questions: We test if learners from four different L1 backgrounds (1) still deviate from native speakers in the way they establish fluency (and where they have already approximated to the native target norm), (2) establish fluency in different ways (e.g. by showing preferences of using different fluencemes over others to establish fluency), (3) use fluencemes in different positions in the utterance, and (4) if their use of fluencemes can be predicted by extra-linguistic parameters such as age, gender or the task type during the interview.

Methodologically, we use an application that automatically extracts these fluencemes from the five corpora and shows them in their communicative context, which makes it easier and more convenient to disambiguate their use (e.g. well as a discourse marker vs. an adverb). After the automatic extraction and a manual post-editing of these fluencemes, we analyze them using multivariate regression modelling (e.g. Gries 2013) using the software package R (R Development Core Team 2017) in order to answer our research questions. The preliminary findings derived from these analyses suggest that, while there is considerable variation between all four learner varieties compared to the native speaker data, all learners show a heavy underuse of discourse markers and smallwords and a heavy overuse of filled and unfilled pauses. The position of dysfluencies such as filled or unfilled pauses shows a tendency to be quite similar across the learner data, with showing many disfluencies within clauses or even constituents. Also, taking into consideration extra-linguistic parameters predicts an increase in fluency after a stay abroad as well as a significant effect of certain sociolinguistic variables (e.g. an increased use of filled pauses is predicted by an increase in the learners' age). These findings will be discussed in the light of (1) the benefits and limitations of using automatic data extraction applications, (2) the relevance of taking into consideration extra-linguistic variables when analyzing fluency in learner corpora and (3) their implications for L1-specific vs. universal features of describing (dis-)fluency in advanced learner language.

References:

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*. Amsterdam: John Benjamins.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures*, 2,225–246.

Dumont, A. (2017). The contribution of learner corpora to the substantiation of fluency levels. In P. de Haan, S. van Vuuren, & R. de Vries (Eds.), *Language, Learners and Levels: Progression and variation* (pp. 281–308). Louvain-la-neuve: Presses Universitaires de Louvain.

Gilquin, G. (2008). Hesitation markers among EFL learners: pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics: A Mutualistic Entente* (pp. 119–149). Mouton de Gruyter: Berlin, Heidelberg, New York.

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

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.

Götz, S. (2019). Do learning context variables have an effect on learners' (dis)fluency? Language-specific vs. universal patterns in advanced learners' use of filled pauses. In L. Degand, G. Gilquin, L. Meurant, A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (pp. 177–196). Louvain-la-neuve: Presses Universitaires de Louvain.

Granger, S. (2015). Contrastive Interlanguage Analysis. A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

Gries, S. (2013). Statistical tests for the analysis of learner corpus data. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data (pp. 287–309). Amsterdam: John Benjamins.

Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 143–173). Amsterdam: John Benjamins.

Müller, S. (2005). *Discourse Markers in Native and Non-Native English Discourse*. Amsterdam: John Benjamins.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing. Foundation or Statistical Computing*. Vienna, Austria. <http://R-project.org> (accessed December 2018).

Permitting a middle ground: an improvement of the MuPDAR method for learner corpus studies

Sandra C. Deshors, Stefan Th. Gries
Michigan State University, University of California at Santa Barbara & Justus Liebig University Giessen
deshorss@msu.edu, stgries@linguistics.ucsb.edu

Corpus-based studies of learner language and English varieties have become more quantitative and increasingly use regression-based methods and classifiers. A development is the MuPDAR approach which improves traditional regression- or tree-based approaches by training a model on the native speaker (NS) reference, by, then, using this model to predict what a reference speaker would have produced in the situation the non-native speaker (NNS) target is in. Crucially, step 3 consists of determining whether the NNS made a nativelike choice or not and explore that variability with a second statistical analysis.

MuPDAR has attracted quite some interest in the LCR community and led to many interesting results, but most applications have one potential shortcoming: Constructional choices by the NNS are categorised as nativelike or not, but the approach has no mechanism to state 'in this context, a NS would be fine with either constructional choice'. This means that current implementations of MuPDAR might be too inflexible in not recognising the real possibility that in certain contexts both constructions are perfectly acceptable; this in turn leads to current MuPDAR potentially being overeager to label a NNS's constructional choice 'non-nativelike' when it is actually nearly just as likely to be used as the competing construction.

In this paper, we introduce a way to remedy this potential problem, which in the above hypothetical would return a prediction of 'either', indicating that the NS would accept either construction, so that our improvement would label even a NNS choice of a prepositional dative as 'nativelike'. We exemplify this using 2234 instances of the dative alternation (ditransitives and prepositional datives with *to*) representing 10 verb lemmas (verbs preferring the ditransitive, the prepositional dative or neither) from two NS corpora (LOCNESS, LOCNEC) and two NNS (ICLE, LINDSEI). These instances were annotated for variables affecting the alternation including lengths and animacy of patient and recipient plus verb lemma and form and L1 and L1 family. We then applied a random forest to the 406 native speaker data points. However, rather than making only a binary prediction – ditransitive vs. prepositional dative – for the 1828 NNS datapoints, we now also permitted predictions that a NS would accept either construction.

Specifically, we discuss two kinds of results: First, a qualitative analysis of the cases that this approach flags as acceptable but whose wrongness the traditional MuPDAR approach would have exaggerated; these involve inanimate patients, the verb forms *gives* and *bring*, but *not* literal transfer, mostly involved recipients a bit shorter than patients and pronominal/definite recipients and indefinite/quantified patients, which actually offers implications for both NS and NNS usage. Second, we compare the traditional kind of results to those of a second random forest based on the new predictions. We find that, while some effects are virtually identical to the results from the more traditional approach, others, notably some involving the L1 of the learners, are now different. For instance, the new analysis shows that the learners with Germanic L1s exhibit verb-specific effects (being better than before with *give*, being worse now with *show*). Also, the new approach shows that the constructional choices in borderline cases by Germanic L1 learners are preferably ditransitives, whereas Chinese and Romance L1 learners prefer prepositional datives, which we suggest can be interpreted in terms of L1-family specific default fall-back choices.

A Corpus Based Study of the L2 Acquisition of (Norwegian) Perfect Notions: the Effects of Semantic, Frequency, and Intralingual Contrast

Ann-Kristin H. Gujord
University of Bergen
ann-kristin.gujord@uib.no

The overall aim of the study is to identify different types of factors that affect the use of the present perfect and acquisition of perfect notions in Norwegian L2, and the interaction among the various factors. The study builds on previous research on how L2 learners learn to express temporal relations on verbs: Research on the importance of verb semantics (e.g. Bardovi-Harlig, 2000), research on crosslinguistic influence (e.g. Jarvis and Pavlenko, 2008), specific studies of the acquisition of the perfect category (e.g. Gujord, 2017), and more recent research of input frequency in L2 acquisition (e.g. Ellis, 2002). With a few exceptions (e.g. Wulff et al., 2009), most studies of L2 acquisition of temporal morphology have typically looked at the effects of various factors in isolation.

Altogether 1189 clauses have been extracted from an electronic learner corpus of Norwegian (ASK). These are distributed across 495 texts written as responses to two different official tests of Norwegian for adult immigrants by learners with seven different L1 backgrounds (English, Polish, Russian, Somali, Spanish, German, Vietnamese) at different proficiency levels (A2-C1 in CEFR). The 1189 occurrences have been coded for the writer's L1, form, temporal context, correctness, erroneousness, verb lexeme, verb lexeme frequency, lexical aspect, syntactical properties, adverbial presence and type, text type and various background information about the writer.

Research questions:

- Do proficiency level in L2 affect overall use, correct and incorrect use of the present perfect?
- Do L1 background affect overall use, correct and incorrect use of the present perfect?
- Do input frequency of the verb lexeme (main verb in the phrase) affect overall use, correct and incorrect use of the present perfect?
- Do the lexical aspectual content of the verb phrase affect of overall use and correct use of the present perfect and the past perfect?
- Do the syntactical properties of the verb phrase (main, subordinate, coordinate) affect overall use, correct and incorrect use of the present perfect?
- Do the presence or absence of an adverb/adverbial phrase affect overall use, correct and incorrect use of the present perfect?

The preliminary results indicate that proficiency level, L1 background, verb semantics and knowledge of English affect the use and acquisition of perfect notions in Norwegian L2.

References

Bardovi-Harlig, K. (1998). Narrative structure and lexical aspect. Conspiring factors in second language acquisition of tense-aspect morphology. *Studies in second language acquisition* 20(4), 471–508.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. Studies in Second Language Acquisition 24, 143–188.

Gujord, A. K. H. (2017). The "perfect candidate" for transfer: A discussion of L1 influence in L2 acquisition of tense-aspect morphology. In A. Golden, S. Jarvis, & K. Tenfjord (Ed..), *Crosslinguistic influence and distinctive patterns of language learning. Findings and insights from a learner corpus* (pp. 29–63). Bristol: Multilingual Matters.

Scott, J., Pavlenko, A. (2008). Crosslinguistic influence in language and cognition. New York: Routledge.

Tenfjord, K., Meurer, P., & Hofland, K. (2006). The ASK Corpus–a Language Learner Corpus of Norwegian as a Second Language. Paper read at Proceedings from 5th International Conference on Language Resources and Evaluation (LREC), Genova.

Wulff, St., Ellis, N. C., Römer, U., Bardovi-Harlig, K., & Leblang, Ch. J. (2009). The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *The Modern Language Journal* 93, 354–369.

Writer/reader visibility in young learner writing: A study of the TRAWL corpus of secondary school texts

Ingrid Kristine Hasund, Hilde Hasselgård
University of Agder, University of Oslo
kristine.hasund@uia.no, hilde.hasselgard@ilos.uio.no

A pervasive finding in learner corpus research is that advanced EFL learners tend to overuse interactional features of writer/reader (W/R) visibility in their academic written texts, including 1st and 2nd person pronouns, private verbs, expressions of modality, evaluation and subjective stance, imperatives and direct questions (see e.g. Ädel, 2008; Aijmer, 2002; Gilquin & Paquot, 2008; Granger & Rayson, 1998; Hasselgård, 2009; Paquot et al., 2013; Petch-Tyson, 1998; Ringbom, 1998; Virtanen, 1998). Very little research has been done on younger learners, however. The present study investigates the use of W/R visibility features across genres in a corpus of EFL texts written by lower secondary school pupils in Norway (age 13–16). At this stage, pupils move from writing predominantly personal/narrative texts to other types of genres, hence this level can give interesting insights into writing development.

Several researchers have asked *why* advanced EFL learners overuse W/R features in their academic writing, and possible explanations include the influence of spoken English, L1 transfer, teaching-induced factors and developmental factors (Aijmer, 2002; Fossan, 2011; Gilquin & Paquot, 2008; Granger & Rayson, 1998). For the present purpose, teaching-induced factors and developmental factors are of particular interest. The former have to do with how task type and setting influences the use of interactional features in learner texts, whereas the latter include the extent to which pupils have acquired the necessary skills to be able to employ an impersonal, formal style when required. Our research questions concern the nature, frequency and distribution of W/R visibility features across levels and tasks. Comparisons will be made with more advanced levels on the basis of work done by Paquot et al. (2013).

The data are drawn from the TRAWL (Tracking Written Learner Language) corpus, currently being compiled in Norway. TRAWL is a longitudinal corpus of authentic texts written by school-age Norwegian learners (age 10-19) (Dirdal et al. 2017). It includes authentic EFL texts written as part of regular class work, assignments/tasks, teacher comments, and metadata about pupils and texts. The subcorpus used in the present study comprises all English texts written by 13 pupils in one class from the beginning of year 8 to the end of year 10, about 50,000 words in total.

Using the Lancaster University corpus toolbox, LancxBox, we will search for features of W/R visibility across genres and school year. Preliminary results indicate that the pupils are highly visible writers, even when producing factual texts. For instance, the collocation *I think* is almost twice as frequent in our material as in ICLE-NO (cf. Paquot et al. 2013). We thus expect to find even more evidence of W/R visibility in TRAWL than in more advanced learner writing. Intriguingly, the frequency of *I think* increases from year 8 to year 10 despite a general decrease of *I*, which can possibly be related to differences in genre and writing tasks. It is therefore important to examine the collocates of first-person pronouns to assess the roles of the authorial *I* (Fossan 2011).

Linking the use of W/R visibility features to teaching-induced factors and developmental factors, we hypothesise 1) that there will be a gradual decrease of W/R features overall in the texts from year 8 to year 10, as pupils develop their academic writing skills, and 2) that there will be a gradual decrease of tasks inviting personal and informal language from year 8 to year 10; i.e. that the teacher offers the pupils an increasing amount of academic writing tasks as they grow older.

References

Ädel, A. (2008). Involvement features in writing: do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M.B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research*, 35–53. Amsterdam and New York: Rodopi.

Aijmer, Karin. (2002). Modality in advanced learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 55–76. Amsterdam: Benjamins.

Dirdal, H., E.M.D. Drange, A.-L. Graedler, T.M. Guldal, I.K. Hasund, S.L. Nacey, S. Rørvik. (2017). "Tracking Written Learner Language (TRAWL): A longitudinal corpus of Norwegian pupils' written texts in

second/foreign languages." Poster presentation at the Third Learner Corpus Research conference, Bolzano. https://susannacey.hihm.no/wp-content/uploads/2017/09/Abstract-LCR-4-Poster-Dirdal-et-al..pdf

Fossan, H. (2011). The writer and the reader in Norwegian advanced learners' written English: A corpus-based study of writer/reader visibility features in texts by Norwegian learners of English and native speakers of English. MA thesis, University of Oslo. http://urn.nb.no/URN:NBN:no-30872

Gilquin, G., & M. Paquot. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction* 1:1, 41–61.

Granger, S., & P. Rayson. (1998). Automatic lexical profiling of learner texts. In S. Granger (Ed.) *Learner English on Computer*, 119–131. London: Longman.

Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.), *Corpora and Language Teaching*, 121–140. Amsterdam and Philadelphia: John Benjamins.

Paquot, M., Hasselgård, H., & Ebeling, S.O. (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 S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, 377–387. Louvain: Presses Universitaires de Louvain.

Petch-Tyson, S. (1998). Reader/writer visibility in EFL persuasive writing. In S. Granger (Ed.), *Learner English on Computer*, 107–118. London: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer*, 41–52. London: Longman.

Investigating the development use of Lexical Bundles and Keyness in B2 and C1 EFL Learners' Academic Writing

Hattan Hejazi
University of Liverpool
Hattan@Liverpool.ac.uk

Lexical bundles (LBs) are word combinations which has been defined as continuous multi-word sequences that recur frequently to satisfy specified frequency and dispersion thresholds, for example, occurring at least 20- 40 times per million words in five texts or at least 10% of the texts (Biber and Barbieri, 2007, Chen and Baker, 2016). The considerable attention has been given to LBs within the area of corpus linguistics has increased since it has been widely agreed that LBs are widespread in spoken and written registers, serve as "building blocks of discourse and frequent use of these bundles is indicative of fluency in linguistic production. To researcher's knowledge, only little research has been done to investigate whether learners from different proficiency levels groups exhibit the same behaviour in the use of LBs or not. This research investigates whether there is a relationship between the use of three- and four-word LBs and language competence. The study conducts both quantitative and qualitative analyses to see whether learners from different CEFR levels groups exhibit the same behaviour in the use of lexical bundles. An additional aspect, this study examines the development of LBs across the proficiency levels. Therefore, it compares between two different levels B2 and C1 in term of frequency, structures and functions of LBs to give an overview of some of the linguistics features to differentiate between the levels.

This study was first concerned about the relationship between the use of LBs and the academic performance, thus, it compared between B2 and C1 sub-corpora (frequency, structures and functions of LBs) of EFL learners. The data used come from written essays equivalent to the IELTS test in term of the title by intermediate and advanced EFL learners who have studied in the UK. The procedure for determining the CEFR levels originates from the manual for Relating Language Examinations to the Common European Framework of Reference for Languages (Europe, 2003).

In the second stage, the study comes under second language development research which compares learners' language across proficiency levels (CEFR levels). A longitudinal study investigated the 3 months' development of two EFL learners use of lexical bundles in their academic essays across the levels to give a picture of the increases of the proficiency levels.

The analysis used was provided by wordsmith computer software (Scott, 2012). Due to the smaller sub-corpora size in this study, the low-frequency cut-off point 4 times per 100,000 (40 times per million words) was selected to include highly used LBs in the analysis and eliminate low-frequency parameters. In addition, a bundle has to be found in at least 3-5 texts (Biber and Barbieri, 2007, Chen and Baker, 2010) or in at least 10% of the texts (Hyland, 2008) to avoid focusing on idiosyncratic uses by individual speakers of the texts.

A major finding from the analysis shows that generally EFL learners favoured to use more signalling bundles in their writing, three-word bundles turned out to be the most frequent bundles in EFL sub-corpora. Moreover, a significant progress has found in the variability of the structures and functions of LBs, C1 writers are found to have used various structures and functions as professional writers in their academic writing.

For development of LBs in relation to the CEFR levels. The findings clearly indicate that there is no significant relationship between the increase use of LBs and the academic performance. However, multiple regression analysis revealed that there is a direct proportionality between variations of the use of LBs and the CEFR levels, as higher-level students (C1) act as professional writers and used variant structures; and functions than lower level (B2).

References

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

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics* 37**,** 849–880.

Hyland, K. (2008a). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1)**,** 41–62.

Scott, M. (2012). WordSmith Tools (Computer Software. Version 6.0). Liverpool: Lexical Analysis Software.

Intensifying compounds in the Diasystem of Belgian French-speaking learners of Dutch and English

Isa Hendrikx
Université catholique de Louvain
isa.hendrikx@uclouvain.be

The present study focuses on the acquisition of adjectival intensification: $[[X]_{INT}\ [Y]_{ADJ}]_{ADJ/AP}$ ↔ 'very Y' (e.g. *very proud*). The diversity of constructions (such as degree adverbs, intensifying prefixes, compounds, etc.) and the language-specific preferences for particular types of intensification (Hoeksema 2011, 2012; Rainer 2015) may complicate the acquisition of intensifying constructions for second language learners (Lorenz 1999). Specifically intensifying adjectival compounds (henceforth IAC) (e.g. *ice-cold*) are expected to be difficult to acquire. While these constructions are a productive means to express intensification in Dutch and in English, in French this construction is hardly productive. In consequence, French-speaking learners may encounter difficulties acquiring IAC in L2 Dutch/English.

This study is situated within the theoretical framework of Construction Grammar (cf. Tomasello 2003; Ellis & Cadierno 2009 a.o.). More specifically, the results are interpreted taking the approach of *Diasystematic Construction Grammar* (DCxG) (Höder 2012, 2014) which conceptualizes the linguistic competence of multilingual speakers as an 'interlingual network of constructions with different degrees of schematicity' (Höder 2012: 255). Analyzing the interlanguage of French-speaking learners of Dutch and English through the lens of DCxG allows one to identify the diasystematic links between intensifying constructions in French (L1) and the target languages of these learners. In this contribution I will address the following research question: Does more target language exposure provided through *Content and Language Integrated Learning* (CLIL) lead to a deeper entrenchment of (more) diasystematic constructions and target language idioconstructions?

Within the context of a research project on CLIL in French-speaking Belgium (cf. Hiligsmann et al. 2017), I assess the impact of CLIL input on the acquisition of IAC in the L2. The sample consists of French-speaking 12th grade pupils (aged 17-19), in CLIL and non-CLIL settings, learning Dutch (CLIL n=132; non-CLIL n=100) or English (CLIL n=90; non-CLIL n=90). A corpus study on written productions of these learners revealed that the CLIL students display a greater written proficiency in terms of lexical diversity among others (Bulon et al. 2017) and a more target-like use of intensifying constructions (Hendrikx et al. 2019). Since IAC are infrequent in the learner corpora, the present study uses a multiple-choice exercise to evaluate the learners' receptive knowledge of IAC. In this manner both the learners' productive use and receptive knowledge of IAC are analyzed. In order to distill the effect of CLIL, other target language exposure variables are included in the analysis (i.e. the number of years of target language learning and the current informal contact with the target language). I also analyze the impact of measures of receptive L2 vocabulary knowledge (PPVT-IV or PPVT-III-NL) and of productive L2 vocabulary knowledge (Measure of Textual Lexical Diversity), as predictors for a learner's receptive knowledge of IAC.

Preliminary results indicate that CLIL pupils develop greater receptive knowledge of IAC, both for L2 Dutch and L2 English. Interpreting the findings within the framework of DCxG, different levels of linkage between the L1 and the target language can be observed. On the one hand, cross-linguistic similarities lead to entrenched diasystematic constructions, for instance $[ADV_{booster}+ADJ_{scalar}]$ (instantiated by e.g. *heel leuk* / *very nice*). On the other hand, despite different degrees of productivity between particular native and target language constructions, input can favor the formation of diasystematic links, illustrated by the CLIL learners' greater productive use and receptive knowledge of target language IAC.

References

Bulon, A., Hendrikx, I., Meunier, F., & Van Goethem, K. (2017). Using global complexity measures to assess second language proficiency Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. *Papers of the Linguistic Society of Belgium* 11*(*1), 1–25.

Ellis, N., & T. Cadierno (2009). Constructing a Second Language. *Introduction to the Special Section. Annual Review of Cognitive Linguistics* 7, 111–139.

Hiligsmann, Ph., Van Mensel, L., Galand, B., Mettewie, L., Meunier, F., Szmalec, A., Van Goethem, K., Bulon, A., De Smet, A., Hendrikx, I., & Simonis, M. (2017). Assessing Content and Language Integrated Learning

(CLIL) in French-speaking Belgium Linguistic, cognitive and educational perspectives. *Les Cahiers de Recherche du Girsef* 17(109), 1–25.

Hendrikx, I., Van Goethem, K., & Wulff, S. (2019). Intensifying constructions in French-speaking L2 learners of English and Dutch: cross-linguistic influence and exposure effects. *International Journal of Learner Corpus Research,* 5(1), 63–103.

Höder, S. (2012). Multilingual constructions: a diasystematic approach to common structures. In K. Braunmüller, & Chr. Gabriel (Eds.), *Multilingual individuals and multilingual societies* (*Hamburg studies on multilingualism* 13), Amsterdam/Philadelphia: Benjamins, 241–257.

Höder, S. (2014). Constructing diasystems: Grammatical organisation in bilingual groups. In T. A. Åfarli, & B. Maehlum (Eds.), *Studies in Language Companion Series* 154. Amsterdam: Benjamins, 137–152.

Hoeksema, J. (2011). *Bepalingen van graad in eerste-taalverwerving*. *TABU* 39(1/2), 1–22.

Hoeksema, J. (2012). Elative compounds in Dutch: Properties and developments. In Oebel, G. (Ed.), *Intensivierungskonzepte bei Adjektiven und Adverben im Sprachvergleich*. Hamburg: Verlag Dr. Kovac, 97–142.

Lorenz, G.R. (1999). *Adjective intensification. Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi BV.

Rainer, F. (2015). 77. Intensification. In P.O. Müller (Ed.), *Word-Formation: An International Handbook of the Languages of Europe*. Berlin/Boston: De Gruyter Mouton, 1339–1351.

Tomasello, M. (2003). *Constructing a language: A Usage-Based Theory of Language Acquisition*. Boston: Harvard University Press. Van Goethem, K. (2009). Choosing between A+N compounds and lexicalised A+N phrases: The position of French in comparison to Germanic languages. *Word Structure* 2, 241–253.

In search of a gold standard for error annotation: lexical errors

Tim Hoffmann
University of Marburg
timhoffmann@staff.uni-marburg.de

Error analysis has been firmly established as an important part of learner corpus research and spawned numerous studies contributing to our understanding of interlanguage and its development. Important theoretical assumptions affecting the choice of *concrete* error categories and their application are, however, not often discussed publicly. This is surprising for two reasons: First, the categories formed during the creation of an error annotation scheme greatly affect the outcome of any study. Second, it is often difficult for readers to properly interpret the results of a study if the annotation guidelines are unknown to them. Compounding these issues is the fact that there are few guidelines for *practical* work with error categories, leading to a proliferation of independent tag sets.

Therefore, using lexical errors as an example, this study seeks to open up a dialog about 'best practices' for the creation and application of error annotation schemes. One key area of interest lies in finding borderline cases where an error might fit several categories. For a simple example, consider (1):

(1) *[...] *I have never heared bevore from this*

Is **heared* best described as an orthographic or morphological error? A case could be made for either choice, yet the annotator's choices are often left implicit. In other instances, it might not even be clear what the proper correction should look like, as is the case in (2):

(2) *[...] *the respect others are spending you ahead* [...].

One way around this ambiguity lies in the use of multiple target hypotheses for the same error, as proposed by Lüdeling et al. 2005. While this approach has great merit and should be considered the default, individual target hypotheses are still subject to the above issues. Ambiguity can also be countered via the creation of finely-grained taxonomies, such as the one used in the *TREACLE* (O'Donnell et al. 2009). This poses another question: is there a 'sweet spot' for the level of detail in general error annotation? The third issue is the underrepresentation of intermediate learners and annotation problems arising from their errors. The above can be summed up more concisely in these research questions:

- RQ1: Which lexical errors are most problematic in terms of spanning several error categories?
- RQ2: Can we pinpoint a level of granularity for error categories that offers a compromise between depth of description and ambiguity?
- RQ3: How does the learners' proficiency impact the difficulties in error annotation?

This study aims to answer these questions by applying existing tag sets (i.e. those used by the *ICLE*, *TREACLE*, *CLC*) to a longitudinal subsample taken from the *MILE* (Kreyer 2015). First, troublesome errors are analyzed regarding their treatment in each of the tested tag sets, thus revealing their respective strengths and weaknesses. A look at error type frequencies serves to quantify the impact problematic errors have on the outcome of error analysis. Secondly, this quantitative approach provides data to help answer the question of how much granularity is desirable in general error annotation. Together, these insights could be used to motivate future efforts towards the creation of a gold standard for error annotation.

References

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins. 13–34.

Lüdeling, A, Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics* 2005, Birmingham.

O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., Stuart, K., & Boquera, M. (2009). *Exploring the proficiency of English learners: The TREACLE project*. Proceedings of the Fifth Corpus Linguistics, Liverpool.

Distinguishing between learner vs. novice writing features: a crosslinguistic approach

Pauline Jadoulle
Université catholique de Louvain
pauline.jadoulle@uclouvain.be

In recent decades, studies in Learner Corpus Research have highlighted features that are considered to be typical of English as a foreign Language learner writing, such as register unawareness and overuse of reader/writer visibility features (e.g. Gilquin et al., 2007; Paquot, 2010). There is, however, a debate surrounding this tendency to portray these features as "learner-typical". A number of studies found similarities between L2 and L1 (English) novice writing, and therefore emphasize that "*expertise* is a more important aspect to consider than nativeness" (Römer, 2009: 99). Academic writing might thus be better described in terms of *novice* writing vs. expert writing.

If the aforementioned features are in fact characteristic of novice academic writing rather than learner writing, it could be hypothesized that they are shared by novice writers across languages. However, to date hardly any studies have compared novice academic writing across languages to better tease apart features of learner vs. novice academic writing. This is what this study seeks to investigate via a crosslinguistic approach to lexical bundles in novice L1 French, novice L1 English and French learner L2 English.

The choice to work on lexical bundles was primarily motivated by that the fact that, while some authors state that "phraseology is one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners" (Granger & Bestgen, 2014: 229), others claim that phraseological features uncovered in EFL writing are better explained by the noviceness of these writers rather than by their non-nativeness (Römer, 2009). The challenge, however, is to take into account the typological differences between French and English, which raise interesting questions concerning bundle extraction and analysis.

In this first case study, I focus on two research questions:

RQ1: To what extent do the lexical bundles used by L1 French and L1 English students resemble the lexical bundles found in comparable corpora of expert academic writing?

RQ2: To what extent do features of novice writing in L1 French and L1 English share commonalities?

To answer these questions, I zoom in on 2- to 4-word bundles with personal pronouns and analyze them in terms of frequencies, structures and functions in corpora of novice and expert writing in the discipline of linguistics. To answer RQ1, I draw intralanguage comparisons between novice and expert academic writing. First, the *French Academic wRiting corpus* (FAR; 216,470 words), a corpus of French novice writing, is compared with the KIAP-LING-FR, a corpus of French expert academic writing (339,490 words). For the English counterpart, samples drawn from the *British Academic Written English* (BAWE) corpus and the *Michigan Corpus of Upper-Level Student Papers* (MICUSP), two corpora of L1 English novice academic writing, are compared to the *Louvain Corpus of Research Articles* (LOCRA-LING; 1,071,494 words). Lexical bundles in the two novice corpora are then compared to answer RQ2.

Results will also be compared with lexical bundles as used in the *Varieties of English for Specific Purposes dAtabase* (VESPA-LING-FR; 413,161 words), thus starting to address a third research question:

RQ3: To what extent do French EFL learner writing resemble L1 French vs. L1 English novice writing?

By questioning the status of features put forward as learner-typical in the literature, my PhD project will help to gain deeper understanding of the interplay between non-nativeness and noviceness in academic writing. It also aims to inform the development of the Louvain EAP dictionary (LEAD) by determining whether there is a need for more EFL-specific usage notes or, on the contrary, whether users would be better served if "more emphasis be put on expertise than on nativeness" (Römer, 2009: 99).

References

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In Wilson, A., Rayson, P., & McEnery, T. (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71–92). Frankfurt am Main: Peter Lang.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319–335.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching* 52, 229–252.

Louvain EAP dictionary (LEAD), https://leaddico.uclouvain.be/login (last accessed 15 January 2019)

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

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

Using 'what already works' to 'bridge the gap' between corpus research and corpora in schools

Barry Kavanagh
Inland Norway University of Applied Sciences
barry.kavanagh@inn.no

Researchers have discussed the need to 'bridge the gap' between corpus linguistics and the direct application of it in teaching (e.g. Mukherjee, 2004). This paper asks the question *What does bridging the gap mean in practice?* It is not just a case of educating teachers in corpus linguistics, as it has been observed that what pre-service teachers learn about it in their studies is often absent from their teaching practice (Breyer, 2009: 156; Callies, 2019; Leńko-Szymańska, 2014: 261). Recognizing the 'need for institutionalized teacher-training courses devoted to or featuring the applications of corpora in language instruction' (Leńko-Szymańska, 2014: 261), studies have developed and described instruction for pre-service teachers (Breyer, 2009; Farr, 2008; Leńko-Szymańska, 2014; Zareva, 2017), but we lack at present knowledge of how teachers go on to use corpora in service. In order to explore what is different about the in-service teaching situation, I focus on English in Norway, where it has been stated that corpus methods have not been applied in language teaching to a great degree (Cardona et al., 2014: 1).

The results presented in this paper are from a questionnaire taken by 210 in-service teachers of English in Norway, and in-depth interviews with 3 of them. The questionnaire was designed to map the informants' general familiarity with corpora, and also to discover what those informants who are familiar with corpora use in their teaching. The subsequent interviews were designed to explore the practices of some of the informants who claimed to use corpora.

The findings indicate a low level of corpus use. While 55 informants claimed to be 'fairly familiar with corpus linguistics' and 34 claimed to 'have already done some work with corpora', only 12 claimed they had either used corpora-based materials in teaching or introduced corpora to pupils. The interview data reveals corpora as being used for teaching materials, in written feedback, and in the classroom. The interviewees use corpora primarily for teaching vocabulary, using online interfaces as their main tools (they did not download concordancers). There is a 'light' use of corpus tools in teaching, regardless of the age of the pupils. In upper secondary school, there is less focus on the English language itself in the English subject (and more on, for example, argument structure), and pupils are not marked solely for language in exams. At lower levels, where there is more focus on linguistic aspects of the subject, corpus tools were said by the interviewees to lack the necessary 'user-friendliness', and also collocation is not taught to less advanced pupils. One particular interviewee who studied corpus linguistics at master's level does not use corpus methods in his teaching – this may indicate that bridging the gap is not necessarily a case of making pre-service teachers corpus linguists.

Bridging the gap could mean an exploration of what already works for English teachers who use corpus methods, and feeding that back into the corpus training that pre-service teachers are provided with. This may be a more realistic way to spread corpus methods, identifying (1) technology to expose pre-service teachers to: 'simple' online tools rather than more complicated tools for corpus analysis; (2) corpus methods for teachers when pupils focus on language issues, such as lower secondary in Norway, where pupils are to 'use central patterns for orthography, word inflection, sentence and text construction' (Udir 2013: 9); and (3) corpus methods for teachers when language teaching no longer focuses explicitly on grammar, such as upper secondary in Norway, where pupils are to 'understand and use an extensive general vocabulary and an academic vocabulary related to one's education programme' (Udir, 2013: 10).

References

Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers, *Computer Assisted Language Learning* 22:2, 153–172. doi: 10.1080/09588220902778328

Callies, M. (forthcoming 2019) In *Studies in Corpus Linguistics* series: https://benjamins.com/catalog/scl). 'Integrating corpus literacy into language teacher education: The case of learner corpora', manuscript provided by the author on 23 November 2017.

Cardona, M.D., Didriksen, A.A., & Gjesdal, A.M. (2014), Korpusbasert undervisning I fremmedspråkene: La elevens nysgjerrighet sette dagsorden, *Acta Didactica Norge* 8(2), Art. 4, 1–26.

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

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education, *ReCALL* 26(2), 260–278.

Mukherjee, J. (2004). Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In U. Connor, & T. A. Upton (Eds.), *Applied Corpus Linguistics – A Multidimensional Perspective.* Amsterdam/New York: Rodopi.

Udir (Norwegian Directorate for Education and Training). (2013). *English Subject Curriculum.* http://data.udir.no/kl06/ENG1-03.pdf?lang=eng.

Zareva, A. (2017). 'Incorporating corpus literacy skills into TESOL teacher training', *ELT Journal* 71(1), 69–79. doi:10.1093/elt/ccw045

"Applying the right statistics": Can advanced L2 learners acquire register-specific distributional statistics?

Elma Kerz, Daniel Wiechmann, Marcus Ströbel
RWTH Aachen University, University of Amsterdam
elma.kerz@ifaar.rwth-aachen.de, d.wiechmann@uva.nl, marcus.stroebel@ifaar.rwth-aachen.de

Emergentist approaches to the acquisition of language highlight the experientially adaptive nature of linguistic knowledge (e.g. Christiansen & Chater, 2016). In these approaches, language learning heavily relies on learning the statistical regularities and distributional patterns inherent in linguistic input. This lifelong process brings about changes in language representation in response to the statistics in linguistic input (e.g. Chang et al., 2012). These experientially-driven adaptive processes occur on multiple linguistic levels and apply to the acquisition of new structures, the modification and adjustment of already learned representations. However, the growing body of research on linguistic adaptation has primarily focused on comprehension and has been confined to laboratory studies. Further, this research has only looked at short-term adaptation processes within the time period of an experiment and to the stimulus material used in experimental settings.

In this paper we demonstrate how the combined use of learner corpora and NLP techniques can make a unique and important contribution to this line of research. Focusing on distributional frequencies of relative clause (RC) constructions, which have played a pivotal role in debates on language processing and acquisition, we investigate whether to what extent L2 learners of English can acquire the distributional statistics of a target register as a result of long-term adaptation processes. The data used in this study come from three corpora: (1) an L2 academic writing corpus currently compiled at RWTH Aachen University, (2) the BAWE corpus on similar topics, and (3) the COCA corpus with its the five registers (academic, fiction, magazine, news, spoken). In a first step, all texts were parsed using the Stanford CoreNLP dependency parser and all instances of the RC types distinguished in Roland et al. (2008) were automatically extracted, yielding a total of 8,286 RCs from the L2 learner corpus and a total of 5.56 million RCs from the reference corpora. The resulting RC frequency distributions were analyzed using unsupervised machine learning techniques to assess the degree of (dis)similarity between the RC frequency distributions of the learner corpora and the COCA reference corpora. In a second step, we conducted regression analyses to determine whether an individual learner's similarity to the academic target registers could be predicted from their similarity to the other registers that they may encounter as input. Our results indicate that, at the group-level, L2 learner production was most similar to academic COCA components and most dissimilar to the spoken and fiction ones, suggesting that the tested L2 learner group has successfully adapted to the distributional patterns of the target register. However, we observed considerable variation in performance. Results from the regression modelling revealed that this variation was strongly associated with learners' similarity scores to non-target registers, suggesting that their RC usage in the academic register was based on the statistics derived from other sources of language input. The implications of our findings for current theories of language acquisition are discussed.

References

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, 6(5), 259–278.

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3), 348–379.

Non-Canonical Syntax in Learner Languages: A Contrastive Interlanguage Analysis

Kathrin Kircili
Justus Liebig University Giessen
kathrin.kircili@anglistik.uni-giessen.de

The English language offers seven basic, canonical, clause patterns (Quirk et al., 1985). Particularly in writing, however, they do often not suffice to convey all the communicative interests that writers have in mind, and it is the restructuring of sentence elements by means of non-canonical structures that enables them to do so. The range of patterns is versatile and includes, among others, *fronting*, *dislocations* or *introductory-it*. While a considerable amount of research has been dedicated to their description in the ENL context (e.g. Quirk et al., 1985; Birner and Ward, 1998; Biber et al., 1999), in EFL the majority of studies focus on individual phenomena (one of the laudable exceptions being Callies 2009) as well as learner backgrounds rather than providing comprehensive overviews.

Against this backdrop, the present paper will report on the pilot study findings of a Contrastive Interlanguage Analysis (CIA; cf. Granger, 2015) on seven non-canonical sentence patterns (*fronting*, *inversion*, *existential-there*, *introductory-it*, *right* and *left dislocation* as well as *clefting*) among learners of four different L1 backgrounds (German, Spanish, Turkish and Japanese). The study is based on the respective components of the *International Corpus of Learner English* (ICLE; Granger et al., 2009) along with both British and American native-speaker essays of the *Louvain Corpus of Native English Essays* (LOCNESS).[1]

In total, 1,800 sentences (300 per target- and interlanguage) were manually annotated for various syntactic and pragmatic variables, including, among others, sentence length, information status, type(s) and function of non-canonical phenomena. The data were subjected to regression modelling to test for possible predictors of non-canonical patterns in the learner vs. the native-speaker data. The analysis showed that all varieties exhibit commonalities, like the preference for fronting as the most frequent phenomenon. Identified differences include the choice of non-canonical patterns that are also in use in the learners' L1, such as the German speakers' frequent employment of dummy pronouns (1) or the common use of left-dislocations in the Spanish data (2):

(1) It was nearly impossible for her to live an own life […]. (ICLE-GE-DRE-0023.1)
(2) And France, well that is a special case: (ICLE-SP-UCM-0010.1)

These findings will be discussed in the light of the ongoing debate on whether EFL performances can be traced back to an L1 transfer (cf. Jarvis, 2000) or whether the employment of non-canonical structures is rather influenced by language universals (cf. e.g. Gass, 1984).

References

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

Birner, B., & Ward, G. (1998). Information Status and Noncanonical Word-Order in English. Amsterdam/Philadelphia: John Benjamins.

Callies, M. (2009). Information Highlighting in Advanced Learner English: The syntax- pragmatics interface in second language acquisition. Amsterdam/Philadelphia: John Benjamins.

Gass, S. (1984). A Review of Interlanguage Syntax: Language Transfer and Language Universals. *Language Learning* 34(2), 115–132.

Granger, S., Dagneaux, E., & Meunier, F. (2009). International Corpus of Learner English. Louvain: UCL.

Granger, S. (2015). Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1(1), 7–24.

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

[1] Cf. http://www.learnercorpusassociation.org/resources/tools/locness-corpus/

False friends in upper-intermediate and advanced learner language: Evidence from learner corpora

Jasmin Knepper, Robert Fuchs
University of Münster, University of Hamburg
jasmin.knepper@gmx.de, robert.fuchs@uni-hamburg.de

The appropriate and correct use of a varied vocabulary is of crucial importance when learning a second or foreign language. A lack of sufficient vocabulary may cause comprehension problems and even lead to a breakdown in communication. False friends (FFs), words that look similar in a learner's first and the target language but differ in meaning, are commonly thought to be a major source of errors (e.g. German 'aktuell' ('current') is an FF of English 'actual' ('real')). However, existing evidence on the scale of the problem and factors that influence it is limited (Roca-Varela 2012, Ambrozova 2014), despite the popularity of this topic in instructed second language acquisition (SLA).

The present study addresses this research gap and investigates the usage of FFs in the spoken and written learner English of university students of English with a wide range of first languages (German, Dutch, Spanish, Italian, French). We perform a contrastive interlanguage analysis, relying on the corpora ICLEv2 (Granger et al. 2009) and LINDSEI (Gilquin et al. 2010). We investigate the influence of a number of factors: first language, word class, word concreteness, mode (spoken vs. written), and relative frequency in the target language and L1. In total, 41 high-frequency FFs (among the 3000 most frequent words in English) were investigated.

Results indicate that most FFs are rarely used erroneously. In written language, this was found in 7.7% (139/1806) of all cases, whereas in spoken language only 3.4% (19/562) of all occurrences were incorrect. A mixed effects regression analysis (with LEMMA as random factor) indicates that FFs whose equivalent occurs frequently in the learner's L1 are significantly more likely to be used inaccurately than those that are infrequent in the L1 (examples include 'actual', 'classic', 'place'). Word concreteness is another significant factor, with more abstract FFs more likely to be used erroneously than more concrete items. Contrary to expectations, errors are more frequent in written than in spoken learner language. As regards L1 background, speakers of Italian and Spanish were significantly more likely to make errors than speakers of German and Dutch, likely due to distinct traditions of learning English as a foreign language in these countries. However, frequency in the target language (i.e. English) and word class do not significantly influence error rate.

Based on these results we make recommendations for second language pedagogy: Teaching should focus more on abstract than concrete FFs, and, in particular, on the limited number of FFs that are frequently used incorrectly by learners.

References

Ambrozova, R. (2012). *Between True and False Friends: Corpus Analysis of Students' Translations*. (Master's thesis).

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage.* Louvain-la-Neuve: Presses Universitaires de Louvain.

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

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, 1-6. Cambridge: Cambridge University Press.

Hill, R. J. (1982). *A dictionary of false friends*. London: Macmillan.

Roca Varela, L. (2012). *New insights into the study of English false friends: Their use and understanding by Spanish learners of English*. (Doctoral dissertation)

The Effects of Study Abroad on Oral L2 Development: Results from a Learner Corpus Study

Zeynep Köylü
İzmir Bakırçay University
zeynep.koylu@bakircay.edu.tr

The advantages of the study abroad context (SA) on L2 development have gained much attention in SLA literature (Llanes, 2011). Scholars have reported numerous benefits of SA especially on oral measures, such as fluency, accuracy, and syntactic and lexical complexity (Pérez-Vidal, 2014). Compared mostly to local immersion contexts (IM) or at home foreign language contexts (AH), SA has been discussed in relation to high amounts of input and interaction opportunities available helping learners develop their L2 even in shorter periods of time (Llanes, 2011). However, very few studies have attempted to categorize the SA in Europe as Anglophone and Non-Anglophone when it comes to learning English as an L2 considering the highly available student mobility programs such as ERASMUS (Köylü, 2016). Given the lingua franca status of English, this new non-Anglophone context is operationalized as English as lingua franca study abroad (ELFSA) in the current study, the effects of which have not yet been studied in SLA to date.

Motivated by this gap in the literature, this longitudinal study investigated the effects of the two study abroad contexts as SA and ELFSA (a European ERASMUS country where English has a lingua franca status) on L2 oral development as compared to the AH foreign language context, as part of a larger study. The participants were 50 Turkish undergraduates, 33 of whom took the 16-week ERASMUS semester either at a university in England or a European country where the native language is one other than English. Following a quasi-experimental mixed-methods pretest-posttest design, an Elicited Oral Imitation Test (EIT, Ortega et al., 1999) was utilized to find out pre-departure and arrival proficiencies especially to see if it was necessary to control initial proficiency levels. The oral data were collected via a one-minute spoken task to determine linguistic complexity, accuracy, and fluency gains. Accordingly, a learner corpus of 100 minutes of spoken data was collected and analyzed. All the data were transcribed and coded into the CHAT format following annotation conventions by Hilton (2009) via CLAN (MacWhinney, 2000) to facilitate measures of fluency, accuracy, and complexity (CAF). Following Skehan (2009) oral fluency was determined through utterance fluency measures which were categorized as speed fluency, breakdown fluency, and repair fluency. As for lexical complexity D measure (MacWhinney, 2000) was calculated for each participant's performance. As for oral syntactic complexity, clauses per analysis of speech (AS) unit (CL/ASU) were determined. Finally, errors per AS-unit (ERR/AS) were determined for oral accuracy. The data from the EIT were scored using the original rubric from Ortega et al. (1999). To discern the intragroup and intergroup development over time, a series of two-way mixed between-within subjects ANOVAs were utilized to discuss the influence of context on oral development. As being one of the most striking findings of the study, no significant main interaction effect was found between the SA and the ELFSA on spoken fluency, yet the latter group had the highest mean gains in terms of two subconstructs; (1) breakdown fluency and (2) speech rate, which might suggest that the ELFSA participants had less hesitation and faster oral production after a semester abroad. The results, which are partially in line with the current SA literature, help us question the scope and necessity of SA programs, and more importantly if Anglophone SA contexts have any additional benefits over the English as a lingua franca context in Europe in terms of oral L2 development.

References

Köylü, Z. (2016). *The influence of context on L2 development: The case of Turkish undergraduates at home and abroad*. (Unpublished doctoral dissertation). The University of South Florida, Tampa, FL.

Llanes, À. (2011). *The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience*. The International Journal of Multilingualism, 3, 189–215.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J.M. (1999) *A Multilanguage Comparison of Measures of Syntactic Complexity* [Funded Project]. Honolulu, HI: University of Hawaii, National Foreign Language Resource Center.

Skehan, P. (2009). *Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis*. Applied Linguistics 30(4), 510–532. doi:10.1093/applin/amp047

The phraseology of core vocabulary in expert and learner data: The case of *thing(s)*

Tove Larsson, Sylviane Granger
Université catholique de Louvain
tove.larsson@uclouvain.be, sylviane.granger@uclouvain.be

A range of learner-corpus-based studies have pointed to an overuse of *core vocabulary* (i.e. the basic, high-frequency words of a language) in students' academic writing (Ringbom, 1998; Källkvist, 1999). As such words tend to be particularly frequent in spoken production, they are often perceived as informal and as a sign of novice writing that needs to be remedied (e.g. Hasselgren, 1994). However, these studies have tended to focus on single words (e.g. *make*) and thus disregard the wide range of productive multiword units that these high-frequency words tend to generate (e.g. *make a contribution, make decisions*), which is unfortunate given the fact that such units are frequently found in expert writing.

The objective of our study is to address this weakness by focusing on the phraseological uses of the high-frequency core word *thing(s)* in second-language (L2) writing as well as in written and spoken expert data. By using written and spoken production, we are able to situate the learners' usage on the formal-informal continuum, on which spoken and written production can be seen as representing end points (Larsson & Kaatari, 2019). *Thing(s)* stands out as an especially interesting core vocabulary item to investigate as it has been shown to be particularly frequent in student writing (Tåqvist, 2016) and as it is often mentioned in academic textbooks as an example of a word that must be avoided (cf. Swales & Feak, 2012).

The data used come from three large corpora: the Corpus of Academic Journal Articles (CAJA; Kosem, 2010), the Spoken BNC2014 (Love et al., 2017) and the International Corpus of Learner English (ICLE; Granger et al., 2009). The ICLE data used represent learners with 10 different first-language (L1) backgrounds. We extracted and analyzed four-word lexical bundles (Biber et al., 1999) with *thing(s)* in each corpus and used exploratory statistics to further study the quantitative results. The research questions used to guide the analysis are as follows:

- To what extent is the learners' usage of the most frequent lexical bundles with *thing(s)* similar to that of the experts, and where on the informal-formal continuum is the learners' usage thus situated?
- How do the different L1 groups in the learner data cluster in terms of the frequency of use of these bundles?

The results showed that out of the top 15 most frequent bundles, only one third were shared between the academic expert data and the learners; examples of shared bundles include *no such thing as* and *of the things that*. In addition, the main discourse functions performed differed across the corpora: while an important function of the bundles in the expert data was *comparing and contrasting* (e.g. *the same thing as, it is one thing*), this function was absent in the learner data, where *placing emphasis* was the primary function (e.g. *the most important thing*). Nonetheless, when the spoken data was added to the analysis, it became clear that although the learners' usage is far from identical to that of the academic experts, it is still found closer to the formal end of the continuum than to the informal. Only minor differences were noted across the L1 groups.

With regard to teaching implications, the results show that multi-word units including *thing(s)* are surprisingly productive in expert academic writing, thus suggesting that teaching students to avoid using *thing(s)* completely is reductive and potentially counterproductive. Instead, we should perhaps aim to raise learners' awareness of the stylistic preferences of multi-word units and move away from talking about "taboo words" that must be avoided towards a more nuanced view which acknowledges that high-frequency words occur in a wide range of constructions, some more formal than others.

References

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). International Corpus of Learner English. Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Hasselgren, A. (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.

Källkvist, M. (1999). Form-class and task-type effects in learner English. A study of advanced Swedish learners. *Lund Studies in English* 95. Lund: Lund University Press.

Kosem, I. (2010). Designing a model for a corpus-driven dictionary of academic English. PhD thesis, Aston University.

Larsson, T., & Kaatari, H. (2019). Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. *International Journal of Learner Corpus Research*, 5(1), 33–62.

Love, R., Dembry, C., Hardie, A. Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3), 319–344.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In Granger, S. (Ed.) *Learner English on Computer* (pp. 41–52). London & New York: Addison Wesley Longman.

Swales, J.M., & Feak, C.B. (2012). Academic writing for graduate students: Essential skills and tasks (3rd Ed.). Ann Arbor: University of Michigan Press.

Tåqvist, M. (2016). "Another thing": Discourse-organising nouns in advanced learner English. Karlstad University studies. Karlstad: Universitetstryckeriet.

Lexical indices as developmental measures of lexical competence and proficiency: a meta-analysis

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

The development of a range of automatically-computed gauges of lexis in learners' written and/or spoken production constitute an important line of research on second language vocabulary acquisition and assessment. Such metrics take the form of various – simpler or more complex – mathematical formulas describing vocabulary of texts or speech samples compiled in learner corpora. Over the last 30 years a host of such indices have been proposed which have been claimed to tap various aspects of lexical competence and proficiency. They have been applied in automatic essay scoring systems used in second language testing on the one hand, and as a yardstick of learners' linguistic ability and development in SLA research, on the other.

One way of validating these measures has been to compare them with other methods of assigning levels to students, for example according to the length of their language study or their results on an independent (vocabulary) test. Another way of confirming the validity of these gauges has been to juxtapose them with scores attributed to (the lexical aspects of) learners' speech or writing by human raters. All such studies have attempted to demonstrate a statistical relationship of one or several measures of lexis in learners' texts with other measures either of their linguistic proficiency or of the quality of their spoken/written performance.

The numerous studies carried out in the last 30 years have produced rather contradictory results about the applicability of lexical metrics as developmental measures. Some prove the discriminatory power of the lexical gauges, some other cannot replicate such an effect on their data. Various papers advocate different indices as best correlates of speech or text quality rated by human judges or by having the best predictive power in regression models.

The aim of this paper to present the results of a meta-analysis of 27 studies in this area and synthesize their findings. For the qualitative summary of the studies, the following information has to be taken into account:

- type of analysed production: speech or writing
- analysed indices: how many, which ones
- study design: group comparisons, correlations with human raters' scores, regression analyses
- participants: different groups of L2 learners (their levels), learners and native speakers

The quantitative synthesis will follow the steps outlined by Oswald & Plonsky (2010).

The results of the meta-analysis reveal which measures have proved to be most effective in capturing L2 learners' growth of lexical competence across various studies.

References

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in Second Language Research: Choices and Challenges. *Annual Review of Applied Linguistics*, 30, 85–110. https://doi.org/10.1017/S0267190510000115

On relativizer use in French learners of English: a corpus-based study

Paula Lissón, Nicolas Ballier, Kim Gerdes
Universität Potsdam, Université Paris Diderot, Université Sorbonne Nouvelle
paula.lisson@uni-potsdam.de, nicolas.ballier@univ-paris-diderot.fr, kim.gerdes@sorbonne-nouvelle.fr

In standard English, the use of an overt relativizer in object relative clauses (ORCs) is optional (Quirk, 1957). In sentences such as "That's the music that I like", the speaker may use an overt relativizer (*which/that),* or a null relativizer (*ø,* also known as *zero).* There are some psycholinguistic theories that address relativizer omission in L1 speakers of English, such as ambiguity avoidance (Temperley, 2003) the Predictability Hypothesis (Wasow, Jaeger, & Orr, 2011), or entrenchment (Wiechmann, 2015), among others. These theories have been essential for many corpus-based studies dealing with the detection of features that condition relativizer use, such as definiteness of the antecedent, complexity, lexical density, or priming. In the learner corpora community, however, relativizer omission literature is scarce: as far as we know, only a few studies deal with relativizer omission/alternation in a quantitative way (e.g. Lester, forthcoming; Olofsson, 2009).

The present work reports new insights on learner use of relativizers. In particular, our study focuses on whether there is a preference in the use of overt vs. null relativizer, and whether different overt relativizers are used with the same frequency and within the same contexts. We also investigate which features correlate with the use of the null relativizer (*ø*) vs. two overt relativizers (*that*/*which*). We address three major questions:

a) which are the factors or patterns correlated with the absence or presence of a relativizer?
b) which are the factors or patterns correlated with the use of *which* and the use of *that*?
c) are the factors or patterns distinctive enough to be different from the ones previously attested in L1 use?

In order to investigate these questions, we downloaded a subset of the EFCAMDAT corpus[2] (Geertzen, Alexopoulou, & Korhonen, 2013). The texts were parsed with SpaCy (Honnibal & Johnson, 2015), and ORCs were extracted using a Python script[3]. The final dataset contained 1,675 ORCs from 560 learners, across 24 tasks. We adapted part of the methodology followed in previous studies on relativizer use (e.g. Fox & Thompson, 2007; Grafmiller, Szmrecsanyi, & Hinrichs, 2016; Hundt, Denison, & Schneider, 2012). Each of the extracted RC was coded with the following features: definiteness, number, length of the antecedent, length of the sentence containing the RC, task, type-to-token ratio (TTR), number of sentences of the text, mean length of the sentences, and learner ID.

A random forest model was fit to the data (see Table 1), with all the aforementioned features as predictors. The model shows that the most relevant features related to the variability in use of the relativizers *which*, *that*, and *zero* are related to external predictors, i.e., stylistic factors of the text where the relative clause was found. A closer analysis of the most important variables within the classification model revealed that TTR[4] scores and mean sentence length values were higher for most of the texts where *zero* relative clauses were found. This, contrary to the complexity principle by Rohdenburg (1996), indicates that learners tend to produce the *zero* relativizer in complex contexts. Interestingly, this was also found in Lester (forthcoming) for learners' spoken use of relativizers.

The analysis of the properties of the antecedents will be also presented and compared with previous studies on English L1 (see Table 2), and the role of the task will be discussed: the frequency and features of the antecedents could be the result of the elicited tasks. We will also discuss the difficulties in the automatic retrieval and annotation of ORCs from a learner corpus, as well as the specific details of the random forest model.

Table 1: Confusion matrix

	that	which	zero
that	223	0	0
which	0	35	0
zero	29	7	1381

[2] Productions corresponding to the B2 level of the CEFR from the French learners of the corpus.
[3] https://github.com/kimgerdes/SUD/tree/master/tools
[4] Used as a proxy, however questionable, for lexical complexity.

Table 2: Comparison of antecedents highly correlated with relativizers in Wasow et al. (2011)

Head-noun	Expected relativizer	Results in this study (number of occurrences)
people	overt	that (1) zero (16)
stuff, little	overt	not enough occurrences
all	zero	that (4) zero (20)
way	zero	that (1) zero (53)
time	zero	that (2) zero (38)
first, last, every	zero	not enough occurrences

References

Fox, B. A., & Thompson, S. A. (2007). Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language"*, 31(2), 293–326.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum.* Somerville, MA: Cascadilla Proceedings Project.

Grafmiller, J., Szmrecsanyi, B., & Hinrichs, L. (2016). Restricting the restrictive relativizer. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2016-0015

Honnibal, M., & Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. (pp. 1373–1378). Presented at the EMNLP.

Hundt, M., Denison, D., & Schneider, G. (2012). Relative complexity in scientific discourse. *English Language & Linguistics*, *16*(2), 209–240.

Lester, N. (forthcoming). That's hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research.*

Olofsson, A. (2009). The Gift of the Gap: A Study of Dutch and Swedish Learners' Use of the English Zero Relativizer. *English Studies*, 90(3), 333–344.

Quirk, R. (1957). Relative clauses in educated spoken English. *English Studies*, *38*(1–6), 97–109.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics (Includes Cognitive Linguistic Bibliography)*, *7*(2), 149–182.

Temperley, D. (2003). Ambiguity avoidance in English relative clauses. *Language*, 464–484. https://doi.org/10.1353/lan.2003.0189

Wasow, T., Jaeger, T. F., & Orr, D. (2011). Lexical variation in relativizer frequency. *Expecting the Unexpected: Exceptions in Grammar*, 175–195. https://doi.org/10.1515/9783110219098.175

Wiechmann, D. (2015). *Understanding relative clauses: A usage-based view on the processing of complex constructions* (Vol. 268). Walter de Gruyter GmbH & Co KG.

A corpus-based study on the factors affecting the use of referential expressions in L1 English-L2 Spanish writing in CEDEL2

Fernando Martín-Villena, Cristóbal Lozano
Universidad de Granada
fmartinvillena@ugr.es, cristoballozano@ugr.es

In topic continuity (TC) contexts where a referent is maintained throughout clauses, L1 English–L2 Spanish learners have been shown to overuse/overaccept overt subject referential expressions (REs) (i.e. overt pronouns and NPs) regardless of the context under investigation (e.g. coordination/subordination).

These results have been accounted for by:

1) the residual deficits that learners show according to the Interface Hypothesis (Sorace & Filiaci, 2006),

2) the low accessibility of a referent following the Accessibility Theory (Ariel, 1990),

3) the Interpretability Hypothesis (Tsimpli et al., 2004), among others.

However, intrinsic factors such as the distance and the number/gender of potential antecedents (Lozano 2016) have also accounted for the overexplicitness found in the written production of L2 learners, although they have not been explored in detail.

Since the vast majority of studies on the production/comprehension of subject REs in L2 Spanish are experimental in nature (inter alia, Rothman 2009; Filiaci 2010), and mostly focus on the dichotomy null vs. overt pronouns, this paper aimed to explore how subject REs (i.e. null/overt pronouns, but also NPs) produced by L1 English–L2 Spanish learners compare to those produced by Spanish native speakers in TC contexts, which have proved to be highly problematic for learners. We also analysed the REs used in contexts of coordination against the rest of contexts since, crucially, null pronouns are only allowed in coordinated clauses with coreferential subjects in L1 English (Ryan 2012), coinciding with the pragmatically felicitous option in L1 Spanish. Finally, this study explored two additional factors: the role played by the gender of potential antecedents and whether a higher number of potential/activated antecedents might result in the use of more explicit REs (i.e. overt pronouns or NPs) in order to avoid ambiguity.

A linguistically-motivated, corpus-based approach is taken to analyse the aforementioned factors in the written compositions of three groups of L1 English–L2 Spanish learners (beginners, intermediates and advanced) vs. a comparable Spanish native control corpus from CEDEL2. Each RE in TC was assigned different tags following a fine-grained tagset implemented in the UAM Corpus Tool: 1) form of the RE (null/overt pronoun/NP), 2) the syntactic patterns in which it occurs (coord./non-coord.), and 3) the number/gender of potential antecedents.

The results show that learners overuse overt pronouns and NPs in TC, although to a lesser extent as proficiency increases in line with previous studies (Rothman 2009; Lozano 2016). This overuse of overt forms is minimised in contexts of coordination with coreferential subjects even at very early stages of acquisition, possibly because English does not require subjects to be overtly realised in such contexts. In addition, the number of potential antecedents seems to modulate the grammar of only intermediate and advanced learners: a higher number of potential antecedents in the production of intermediates results in a higher production of overt forms. By contrast, advanced learners produce less overt REs when the number of potential antecedents increases. Interestingly, the effect of the number of potential antecedents seems to be more pronounced in contexts that do not involve coordination. This could be explained in terms of the parallelism between the null pronouns used both in L1 English and L1 Spanish in coordinated contexts. Finally, the overproduction of overt pronouns and NPs can be explained in terms of the gender of potential antecedents: when the gender is the same, NPs are more likely to be produced so as to avoid ambiguity, whereas overt pronouns are more common with different gender antecedents.

References

Ariel, M. (1990). Accessing Noun Phrase Antecedents. London: Routledge.

Bel, A., García-Alcaraz, E., & Rosado, E. (2016). Reference comprehension and production in bilingual Spanish: The view from null subject languages. In A. A. de la Fuente, E. Valenzuela, & C. Martínez Sanz (Eds.), *Language Acquisition Beyond Parameters* (pp. 37–70). Amsterdam: John Benjamins.

Filiaci, F. (2010). Null and overt subject biases in Spanish and Italian: a cross-linguistic comparison. In C. Borgonovo, M. Español-Echevarría, & Ph. Prévost, (Eds.), *Selected Proceedings of the 12th Hispanic Linguistic Symposium* (pp. 171–182). Somerville, MA: Cascadilla Proceedings Project.

Lozano, C. (2009). Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In N. Snape, Y-k. I. Leung, & M. Sharwood-Smith (Eds). *Representational Deficits in Second Language Acquisition* (pp. 127–166). Amsterdam: John Benjamins.

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus. In M. Alonso Ramos (ed.), *Spanish Learner Corpus Research: Current Trends and Future Perspectives* (pp. 236–265). Amsterdam: John Benjamins.

Rothman, J. (2009). Pragmatic deficits with syntactic consequences? L2 pronominal subjects and the syntax-pragmatics interface. *Journal of Pragmatics*, 41, 951–973.

Ryan, J. (2012). *Acts of reference and the miscommunication of referents by first and second language speakers of English*. Unpublished doctoral thesis. University of Waikato, Hamilton, New Zealand.

Sorace, A., & Filiaci, F. (2006). *Anaphora resolution in near-native speakers of Italian. Second Language Research*, 22(3), 339–368.

Tsimpli, I.-M., Sorace, A., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: A study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism*, 8(3), 257–277.

Identifying Interactional Features Accompanying the Requests in Shopping Role Plays

Aika Miura
Rikkyo University
aika_miura@rikkyo.ac.jp

This study aims to present how interactional features are addressed in annotating requests in a spoken learner corpus. The author developed a multi-layered annotation scheme, investigating shopping role-play interactions between interlocutors (or trained Japanese-speaking interviewers) and test-takers (or 68 learners at A1, 114 at A2, and 55 at B1 level of the CEFR) from the NICT JLE Corpus, using the UAM CorpusTool.

The study addresses the following research questions:

(1) What kinds of interactional features accompany the core of requests produced by learners at different proficiency levels?

(2) Are there any interruptions by the interlocutors in the learners' utterances?

(3) Are there any strategies adopted relating to the negotiation of meaning such as correction, repetition, and elaboration of requests?

As Figure 1 shows, the requests were divided into either *main*, *supporting*, or *combined repair feature* segments. The main segments were categorised into requestive *head acts* (e.g., "Could I use a credit card?") and *internal modification* (e.g., "please") based on the Cross-Cultural Speech Act Realization Project coding scheme (Blum-Kulka, House, & Kasper, 1989). The optional categories of supporting and combined repair feature were constructed to identify discontinuity, interruption, and requestive head acts spreading over multiple turns.

The identification of supporting segments helped avoid multiple counts of requests; a learner's request "My favorite maker is" was categorised as a main segment and "Edwin" as a supporting one since the learner completed his utterance after the interlocutor's interruption by nodding "Mh-hmmm". The supporting segments were classified into *continued/continuing* utterance (e.g., "Edwin"); *alert* (e.g., "Excuse me"); *self-corrected head act* (e.g., "I want to" in "Now, so I want to could you show me some wire key?", where a learner rephrased a head act from "I want to" to "could you"); *confirming* (e.g., "Is that possible?" after the head act "I thought I could exchange this into the other color", which elicits a hearer signal); and *responded yes please* (e.g., "Yes, please", where a learner responded affirmatively to the interlocutor's offer).

The *combined repair feature* signified a series of more than two head acts, which occurred due to *repetition* (e.g., "So this time, I try it. Can I try this on?"), *elaboration* (e.g. "I want I want a basketball shoes. And its color is black. And err size er is Japanese size is err twenty-four size."), and *prompted correction*, which was annotated when a learner rephrased a previously uttered head act, prompted by the interlocutor.

The author manually identified a total of 597, 1,168, and 412 segments, containing 452 (i.e., 87.4%), 895 (89.3%), and 263 (70.5%) requestive head acts for A1, A2, and B1 learners, respectively. No significant differences were found between the three proficiency groups in terms of the distribution of occurrences of each category. However, A1 and A2 learners tended to show higher ratios of combined repair features, exhibiting 27 segments (accounting for 6.0% of the total head acts) and 40 segments (4.5%), respectively, while B1 exhibited only 5 segments (1.9%). Among the three types in this category, elaboration was the most frequent. The self-corrected head act was also the least frequent in B1 as there were 17 (i.e., accounting for 3.8% of the total head acts), 38 (4.2%), and 5 (2.7%) by A1, A2, and B1 learners, respectively.

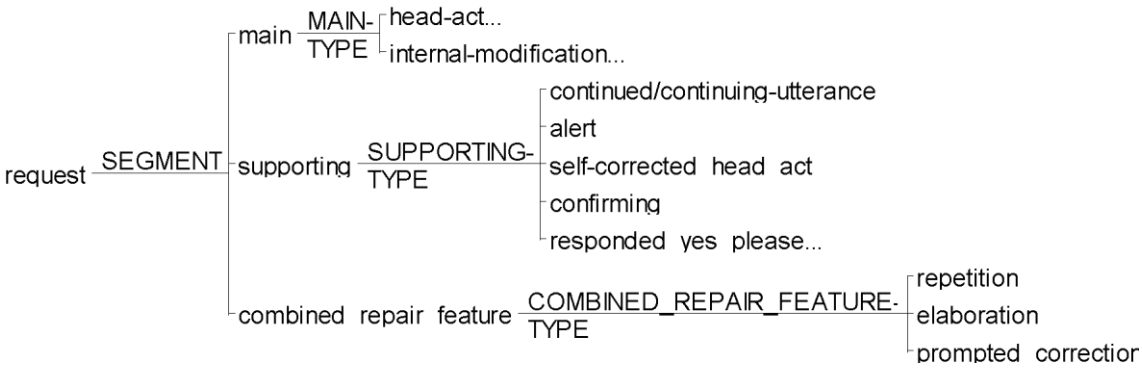


Figure 1. The annotation scheme for identifying the interactional features accompanying the core of requests

References

Blum-Kulka, S., House, J., & Kasper, G. (1989). Investigating cross-cultural pragmatics: An introductory overview. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics: requests and apologies* (pp. 1–34). Norwood, NJ: Ablex.

The effects of frequency and contingency on the accuracy of L2 English grammatical morphemes

Akira Murakami, Nick Ellis
University of Birmingham, University of Michigan
a.murakami@bham.ac.uk, ncellis@umich.edu

The constructionist account of language acquisition holds that (second) language acquisition is strongly influenced by the properties of linguistic input, and that learners are sensitive to the distributional properties such as frequency and contingency of linguistic features (e.g., Ellis, 2002, 2006; Goldberg, 2006). To empirically examine it in the context of English grammatical morphemes as a second language (L2), Guo and Ellis (2018) employed elicited imitation tasks and tested the effects of availability (i.e., surface-form frequency), reliability (i.e., the proportion of a particular inflectional form out of all the occurrences of the corresponding lemma), and formulaicity (i.e., how formulaic the context in which the morpheme occurs is) on the accuracy of imitation. They found that all the three distributional features contribute to the accurate imitation of grammatical morphemes.

The study reported in this talk complements and extends part of Guo and Ellis's (2018) experiments by drawing data from a large-scale partially error-tagged learner corpus. Compared to experimental work, the analysis of a large-scale corpus allows us to target a larger number and range of words and learners, leading to a study with a larger scope and a more fine-grained picture of the effects of relevant factors. Specifically, the study examined (i) whether the use of grammatical morphemes is more accurate in more available and more reliable words and (ii) whether the effects of availability and reliability interact with other factors such as learners' proficiency.

The study employed EF-Cambridge Open Language Database (Geertzen, Alexopoulou, & Korhonen, 2014). The corpus includes learners' writings submitted to Englishtown, the online school formerly run by EF Education First. The course in Englishtown consisted of 16 Levels, each of which covered eight Units. At the end of each unit was a free-writing task on a variety of topics. Each writing received manual feedback that included the correction of grammatical morphemes and was used to calculate accuracy in the present study. The corpus included learners from a range of proficiency levels and nationality backgrounds. The subcorpus used in this study included 122,192 writings by the 2,962 learners with 30 or more error-tagged writings.

The target morphemes of the present study were the same as those targeted in Guo and Ellis (2018); past tense *-ed*, progressive *-ing*, third person *-s*, and plural *-s*. The analysis included the words whose number of obligatory contexts plus overgeneralisation errors was 10 or more in each band of the Englishtown levels corresponding to the Common European Framework of Reference levels A1-C1. The availability and reliability of each form were calculated based on the Corpus of Contemporary American English. In each morpheme, whether a particular (non-)use of the morpheme was erroneous or not was modelled by an additive mixed-effects binary logistic regression model as a function of learners' proficiency, their longitudinal development, the availability of the inflectional form, its reliability, and their two-way interactions, as well as by-learner, by-nationality, by-lemma, and by-topic random intercepts.

The series of statistical models indicated that reliability is a strong predictor of morpheme accuracy, with more reliable forms used more accurately in each of the target morphemes. Its effect, however, was modulated by other variables. Specifically, in past tense *-ed*, the effect of reliability was weaker in higher proficiency learners, while the reverse was true in progressive *-ing*. In plural *-s*, reliability exerted a weaker influence in higher frequency words. Contrary to Guo and Ellis's (2018) experimental findings, availability was not a significant predictor of accuracy in any of the target morphemes. Graphical analyses pointed towards the possibility that the difference is in part due to the difference in the frequency range of the target words. In the current study that sampled words from a much wider range of frequency, availability appeared to be associated with accuracy only in high-frequency words, which Guo and Ellis (2018) primarily targeted.

Overall, our corpus-based study reinforces the view that L2 learners are sensitive to the distributional properties of linguistic input, particularly to the contingency of inflectional form and lemma. To the best of our knowledge, this study and Guo and Ellis (2018) are the first that demonstrated it in the L2 use of English grammatical morphemes. Our study further highlights the value of triangulating experimental findings with learner corpus research.

References

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24. https://doi.org/10.1093/applin/ami038

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Millar, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, A. Tseng, … D. Walter (Eds.), *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines* (pp. 240–254). Cascadilla Proceedings Project.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Guo, W., & Ellis, N. (2018, July). *Second language (L2) knowledge of English morphology: A usage-based account*. Paper presented at the Tenth International Conference on Construction Grammar, Paris.

Syntactic complexity as a part of learner Finnish proficiency

Taina Mylläri
University of Jyväskylä
taina.myllari@jyu.fi

Learner language development is often analysed measuring complexity, accuracy and fluency based on linguistic features, while learner language proficiency is typically assessed using CEFR levels based on communicative competences. The present study seeks to link these two by analysing the development of syntactic complexity in written learner Finnish on different proficiency levels. In this paper, I will address the following research questions:

1. How do clause and sentence structures develop in written learner Finnish?
2. How is the development reflected in proficiency assessments based on the CEFR level descriptions?

Syntactic complexity is usually studied using quantitative measures of length, such as mean length of clause or mean length of T-unit, or measures of subordination, such as mean number of clauses per T-unit or mean number of dependent clauses per clause (e.g. Bulté & Housen 2012, Ortega 2003, Wolfe-Quintero et al. 1998). However, the results of these measures have been inconsistent or even contradictory (e.g. Lu 2010, Ortega 2015), and concerns have been raised about their ability to catch the development of complexity (e.g. Bulté & Housen 2012, Biber et al. 2011).

The present study explores how clauses and sentences in learner Finnish develop and how this development correlates with the perception of learners' language skills. First, the corpus has been annotated with information on the borders and structure of clauses, sentences, and T-units. Next, these production units have been analysed using seven quantitative measures of syntactic complexity: mean length of sentence, mean length of clause, mean length of T-unit, mean number of T-units per sentence, mean number of clauses per T-unit, mean number of clauses per sentence, and mean number of dependent clauses per clause. According to preliminary results, only one of the measures, i.e. mean length of clause, develops linearly from one proficiency level to the next in most task types. Also, there are differences in the timing and the degree of change in complexity measures between the adult and young L2 learners. Similar differences have been reported in previous studies using the same data and focusing on the use of existential sentences (Kajander 2013), indirect references (Seilonen 2013), and transitive constructions (Reiman 2014).

The corpus in the present study comprises 667 texts (48,876 tokens) written by 481 adult L2 Finnish learners, 411 texts (16,590 tokens) written by 212 adolescent L2 learners between 12 and 16 years of age, and 453 texts (19,826 tokens) written by 175 L1 Finnish adolescent in school years 7, 8 and 9, corresponding to the age group of the young L2 learners. The pseudo-longitudinal data was compiled and assessed on CEFR proficiency levels during the CEFLING project at the University of Jyväskylä (for the reliability of the assessment see Huhta et al. 2014). The adult L2 learner texts cover all CEFR proficiency levels from A1 to C2, the young L2 learner texts levels from A1 to B2. The tasks used to elicit the data include writing informal messages, formal messages and argumentative texts. The corpus makes it possible to measure the changes on proficiency levels from beginners to advanced learners, to compare adult and adolescent language learner texts and to compare L2 and L1 texts.

The preliminary results show that there is a need for new measures of syntactic complexity in learner Finnish. Potential indicators of development could be changes in the variety in clause types or other clause-level structures, e.g. in verb phrases, or changes in the use of connectors to combine clauses or the variety of dependent clause types used. The results provide new empirical evidence on syntactic complexity and its development in a language that is structurally very different from those more frequently studied.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development?, *Tesol Quarterly* 45(1), 5–35.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins Publishing Company, 21–46.

CEFLING = *Linguistic Basis of the Common European Framework for L2 English and L2 Finnish*, https://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/cefling/ (accessed January 2019).

Council of Europe. (2001). *Common European Framework of Reference for: Learning, Teaching, Assessment*. Available at: https://rm.coe.int/1680459f97 (accessed January 2019).

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study – Validating the rating process across tasks, scales and languages, *Language testing* 31(3), 307–328.

Kajander, M. (2013). Suomen eksistentiaalilause toisen kielen oppimisen polulla. *Jyväskylä Studies in Humanities* 220. Jyväskylä: Jyväskylän yliopisto.

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

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion, *Journal of Second Language Writing* 29, 82–94.

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

Reiman, N. (2014). Yläkoulun S2-oppilaiden transitiivi-ilmausten käyttö Eurooppalaisen viitekehyksen taitotasoilla, *Lähivõrdlusi. Lähivertailuja* 24, 183–220.

Seilonen, M. (2013). Epäsuora henkilöön viittaaminen oppijansuomessa. *Jyväskylä Studies in Humanities* 197. Jyväskylä: Jyväskylän yliopisto.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity. Technical report No. 1*. Honolulu: Second Language Teaching and Curriculum Center.

Metaphors in high-stakes language exams

Susan Nacey
Inland Norway University of Applied Sciences
susan.nacey@inn.no

Lakoff and Johnson's Conceptual Metaphor Theory advances the view that metaphor is a fundamental cognitive process defining our understanding of reality: "the essence of metaphor is understanding and experiencing one kind of thing [e.g. love] in terms of another [e.g. a journey]" (Lakoff & Johnson, 1980, p. 5). Such metaphors in thought (*conceptual metaphors*) are reflected as metaphors in language, i.e. by the words and expressions produced (*linguistic metaphors*). Empirical research has since confirmed that linguistic metaphor is ubiquitous in both L1 and L2 language (see e.g. Steen et al., 2010).

While metaphor therefore necessarily plays an important role in language learning, some scholars suggest that processing figurative language may pose challenges for L2 speakers of a language, who are less familiar with cultural conventions and connotations, and lack a figurative language repertoire (see e.g. Littlemore & Low, 2006). Metaphor is thus thought to be "difficult" for L2 learners, although as Low pointed out already in 1988 (p. 137) "it would be helpful to know whether the ways in which learners learn to cope with metaphor are similar from person to person".

This paper addresses that acknowledged need by investigating how L2 learners of Norwegian respond to a task requiring them to interpret a literary metaphor and incorporate that metaphor in a text about their own lives: a task involving both receptive and productive metaphorical competence. The empirical data consists of 22 texts (approximately 10000 words) collected in the Norwegian Second Language Corpus (ASK), written by L2 Norwegian learners as part of the *Test in Norwegian – higher level*. This test is a high-stakes language test primarily intended for immigrants to Norway who need to document their language skills for employment or for admission to Norwegian universities and colleges. The learners were instructed to write a text incorporating their own opinions and experiences of friendship with the message(s) in the Kolbjørn Falkeid poem *Det er langt mellom venner* [It is far between friends]. At the poem's core is metaphorical simile steeped in the background of the Norwegian "hyttekultur", the tradition of enjoying cabins in the wilderness as a means of temporarily escaping from the demands of daily life.

This study focuses on the degree to which the informants themselves produce metaphor in their response manifesting their understanding of the poem (even though the exam instructions included no explicit mention of metaphor). Metaphor identification in the L2 texts is carried out using the Scandinavian version of MIPVU, which requires analysis of each word for metaphorical status (Nacey et al., forthcoming). Subsequent analysis focuses upon metaphor density (i.e., how much metaphor is produced), as well as the role of the identified metaphor clusters (i.e., what is the function of metaphor). Preliminary results indicate three main approaches to the task, with the 'interpretation' stage of understanding in particular involving either 1) absence of metaphor, 2) repetition of Falkeid's metaphor without elaboration, or 3) alternative metaphors and/or extension of Falkeid's metaphor through added entailments.

References

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

Littlemore, J., & Low, G. (2006). *Figurative thinking and foreign language learning*. Basingstoke: Palgrave Macmillan.

Low, G. (1988). On Teaching Metaphor. *Applied Linguistics*, 9(2), 125–147.

Nacey, S., Dorst, A. G., Krennmayr, T., & Reijnierse, W. G. (Eds.). (forthcoming). *Metaphor identification in multiple languages: MIPVU around the world*. Amsterdam: John Benjamins.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). Metaphor in usage. *Cognitive Linguistics*, 21(4), 757–788.

Do bilingual immersion programmes affect the use of referring expression in discourse? A corpus-based study on L2 English learners

Teresa Quesada, Cristóbal Lozano
University of Granada
teresaquesada@ugr.es, cristoballozano@ugr.es

Information status factors (e.g., topic continuity and topic shift) constrain the form of anaphors and referring expressions in discourse (zero and overt pronominal subjects, as well as full NPs). Anaphora Resolution (AR) has been shown to be problematic in L2 acquisition at the syntax-discourse interface (Sorace, 2011) as learners are overexplicit (i.e., they use fuller forms than is required) (Leclercq & Lenart, 2013; Ryan, 2015). Importantly, AR has been mostly studied in psycholinguistic studies but there is a lack of linguistically informed corpus-based studies that focus on real discourse.

An unexplored area in AR is whether additional L2 exposure (classroom immersion) is beneficial for the acquisition of AR at the syntax-discourse interface. It has been argued that CLIL (Content and Language Integrated Learning) immersion programmes are beneficial in terms of general L2 proficiency (Lasagabaster, 2008; Ruiz de Zarobe & Jiménez Catalán, 2009, but see Bruton, 2011), but those benefits may not always extend to specific morphosyntactic areas (García Mayo & Villarreal Olaizola, 2011; Martínez-Adrián & Gutiérrez Mangado, 2015). Crucially, it is not known yet whether immersion is beneficial for AR.

The main aim of this developmental study is to use real discourse production (corpus data) to determine i) the factors that constraint the use of REs in discourse and ii) whether additional exposure (CLIL) is beneficial for the L2 acquisition of AR at the syntax-discourse interface. We used the written Corpus of English as a Foreign Language (COREFL) (Lozano, Díaz-Negrillo, & Callies, forthcoming) and analysed samples from L1 Spanish – L2 English CLIL vs mainstream EFL learners at several proficiency levels (A1, A2, B1, B2) and an equivalent English native control corpus (N=119 texts). Data come from the classic frog story previously used in L2 studies of AR (e.g., Kang 2004). We created a linguistically-informed tagset in the UAM Corpus Tool software and tagged multiple factors that previous independent L2 studies have shown to affect the use of REs, e.g., information status (topic continuity/shift), referential forms (null/overt pronominal subjects, NPs), syntactic environment, number of potential antecedents, amongst others.

Overall, preliminary results show that there are several factors that affect the use of REs in AR and it is important to distinguish them in order to determine if learners are overexplicit. If we focus on information status of the REs show that in topic-continuity contexts, even though CLIL learners initially outperform non-CLIL learners (who overuse NPs), non-CLIL eventually catch up and outperform CLIL. Regarding topic-shift contexts, both CLIL and non-CLIL produce more NPs than overt pronouns, as English natives do, but it is again the non-CLIL group that eventually attains native-like levels at B2. Therefore, even though CLIL groups seem to perform better at beginner levels, both CLIL and non-CLIL groups behave similarly at intermediate levels and even the non-CLIL B2 outperforms the CLIL B2. A closer inspection of the syntactic contexts where REs appear (coordinate vs subordinate clauses) reveals that the discourse of CLIL learners is syntactically more complex and elaborate that non-CLIL learners, amongst other findings.

In short, we reveal that additional exposure through CLIL does not affect the syntax-discourse interface but CLIL learners perform better at a discursive level.

References

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, *39*(4), 523–532.

García Mayo, M. del P., & Villarreal Olaizola, I. (2011). The development of suppletive and affixal tense and agreement morphemes in the L3 English of Basque-Spanish bilinguals. *Second Language Research*, *27*(1), 129–149.

Lasagabaster, D. (2008). Foreign Language Competence in Content and Language Integrated Courses. *The Open Applied Linguistics Journal*, *1*(1).

Leclercq, P., & Lenart, E. (2013). Discourse Cohesion and Accessibility of Referents in Oral Narratives: A Comparison of L1 and L2 Acquisition of French and English. *Discours. Revue de Linguistique, Psycholinguistique et Informatique*, (12).

Lozano, C., Díaz-Negrillo, A., & Callies, M. (forthcoming). Designing and compiling a learner corpus of written and spoken narratives: COREFL. In *C. Bongartz, & J. Torregrossa (Eds.), What's in a narrative? Variation in story-telling at the interface between language and literacy.* Frankfurt: Peter Lang.

Martínez-Adrián, M., & Gutiérrez Mangado, M. J. (2015). L1 Use, Lexical Richness, Accuracy and Syntactic Complexity in the Oral Production of CLIL and NON-CLIL Learners of English. *Atlantis. Journal of the Spanish Association for Anglo-American Studies*, *37*(2), 177–199.

Ruiz de Zarobe, Y. R. de, & Jiménez Catalán, R. M.-A. J. (2009). *Content and Language Integrated Learning: Evidence from Research in Europe*. Multilingual Matters.

Ryan, J. (2015). Overexplicit Referent Tracking in L2 English: Strategy, Avoidance, or Myth? *Language Learning*, *65*(4), 824–859.

Sorace, A. (2011). Pinning down the concept of 'interface' in bilingualism. *Linguistic Approaches to Bilingualism*, *1*(1), 1–33.

Corpus-based transfer study of grammatical gender in Norwegian as a second language

Silje Ragnhildstveit
Western Norway University of Applied Sciences
siljera@hvl.no

This paper reports results from a PhD study of transfer when it comes grammatical gender in Norwegian as a second language. In languages with grammatical gender, every common noun has an inherent gender. In Norwegian either masculine (m.), feminine (f.) or neuter (n.). The gender of the noun determines the form of words that agree in gender, for instance as with determiners and adjectives as in the case with Norwegian.

Central questions about L1 transfer:

1. Is it facilitating to have a three gender language like Norwegian (a similar relation between L1 and L2) compared to a two gender language (a different relation between L1 and L2)?
2. Is it facilitating to have an L1 *with* gender, compared to an L1 *without* gender (a unique relation between L1 and L2)?

These questions are investigated by studying the use of correct and incorrect gender agreement with respect to indefinite articles (*en* (m.), *ei* (f.), *et* (n.)) and adjectives alone, and three different types of noun phrases that require gender agreement: "Type I phrase": gender marking shown on the indefinite article and adjective: ***en fin*** *bil* (m.) (a nice car), ***et fint*** *hus* (n.) (a nice house), "Type II phrase": gender marking shown on the definite suffix and the possessive *bil**en min*** (m.) (my car), *hus**et mitt*** (n.) (my house), and "Type III phrase": gender marking shown on the demonstrative and definite suffix ***denne*** *bil**en*** (m.) (this car), ***dette*** *hus**et*** (n.) (this house).

The theoretical grounding is the usage-based approach, included functional learning theory and the Competition Model of Bates and MacWhinney (1987; 1989). The source of data are texts written by adult learners, sitting for a Norwegian language test, extracted from the electronic learner language corpus *ASK – Norsk andrespråkskorpus* (The ASK corpus – learner corpus of Norwegian as a second language). Jarvis' (2000; 2010) methodological framework is used in the investigation of L1 transfer. I compare second language learners of Norwegian with L1s that differ with respect to grammatical gender:

	Target language	**Native language**				
Language	**Norwegian**	**German**	**Dutch**	**Spanish**	**Vietnamese**	**English**
Gender system	Masculine Feminine Neuter	Masculine Feminine Neuter	Common Neuter	Masculine Feminine	– Classifier language	–

There are 100 informants/ texts, from each language group. The ASK Corpus also contains information about various person data. In addition to L1 transfer, this study investigates the role of CEFR level,[5] years of residence in Norway and use of Norwegian (daily, seldom or never). The language data were extracted using different types of word, lemma, pos and morphological queries. The study makes use of different types of inferential statistics (for example Mann Whitney U Test, Kruskal-Wallis H Test and Fishers Exact Test).

One result with respect to L1 transfer is that there is no significant difference between the gender language groups German, Dutch and Spanish, in their interlanguage with respect to the use of gender. The results are ambiguous with regard to the unique relation: The Vietnamese L1 group, but not the English L1 group, evidences more correct use of grammatical gender in most of the analyses compared to the other groups. This result partly supports the Competition Model that claims that a unique relation will not make the acquisition difficult because this relation will not cause competition between the structures in L1 and L2. In general, when controlling for the other learner related factors, the effect of transfer is present. The interaction between the factors are minimal. Input frequency, however, seems to be an important factor, which supports the usage-based approach.

[5] CEFR: *Common European Framework of Reference for Languages*.

References:

Bates, E. & MacWhinney, B. (1987). Competition, Variation, and Language Learning. In B. MacWhinney (Ed.), *The Mechanisms of Language Acquisition* (pp. 157–193). Hillsdale, New Jersey, London: Lawrence Erlbaum Associates.

Bates, E. & MacWhinney, B. (1989). Functionalism and the Competition Model. In B. MacWhinney & E. Bates (Eds.), *The Crosslinguistic Study of Sentence Processing* (pp. 3–76). New York: Cambridge University Press.

Carlsen, C. (2012). Proficiency Level – a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics*, 33(2), 161–183.

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. *EUROSLA Yearbook* 10, 169–192.

Meurer, P. (2012). Corpuscle – a new search engine for large annotated corpora. In G. Andersen (Ed.), *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian* (pp. 29–50). Amsterdam: John Benjamins.

Ragnhildstveit, S. (2017). *Genus og transfer når norsk er andrespråk. Tre korpusbaserte studier*. PhD Thesis, University of Bergen, Norway.

Tenfjord, K., Hagen, J.E. & Johansen, H. (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 on big comparative fallacy). *Rivista di Psicolinguistica Applicata (RiPLA)*, 3, 93–108.

Tenfjord, K., Hagen, J.E. & Johansen, H. (2009). Norsk andrespråkskorpus (ASK) – design og metodiske forutsetninger. *NOA. Norsk som andrespråk*, 1, 52–81.

Tenfjord, K. (2004). ASK – A Computer Learner Corpus. In P.J. Henrichsen (Ed.), *CALL for the Nordic Languages. Tools and Methods for Computer Assisted Language Learning* (pp. 147–158). Copenhagen: Copenhagen Business School.

Tenfjord, K., Meurer, P. & Hofland, K. (2006). The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language. *Proceedings from 5th International Conference on Language Resources and Evaluation (LREC)*, 6, 1821–1824.

Collocation issues in a learner corpus of final-year dissertations by Spanish undergraduates

Noelia Ramon, Ana Frankenberg-Garcia
University of Leon, University of Surrey
noelia.ramon@unileon.es, a.frankengerg-garcia@surrey.ac.uk

In a globalized world, scholars and students from all areas of expertise are increasingly expected to use written academic English. Indeed, the status of English as a lingua franca in academia nowadays is unquestionable (Jenkins, 2014). Users of academic English outside English-medium universities may struggle to produce the phrases, lexical bundles, formulaic sequences or collocations - to name but a few of the terms used in the literature to describe this phenomenon (Howarth, 1998; Wray, 2002) - that make texts idiomatic (Peters & Pauwels, 2015). Additionally, psycholinguistic research has "shown that language is to a great extent, acquired, stored and processed in chunks." (Granger & Meunier, 2008: 247), making texts that conform to predictable, idiomatic combinations of words more readable (Hoey, 2005).

Given the importance of idiomaticity in academic language, new resources are being developed to help novice writers and non-native speakers of English when writing academic texts in English (for example, the Academic Collocations list by Ackermann and Chen (2013), LEAD by Granger and Paquot (2015) and ColloCaid by Frankenberg-Garcia and her team (2019). These aids can be enhanced with data from learner corpora, and the present study explores how learner data on academic English collocations by Spanish undergraduates can be used to this effect.

At Spanish universities, undergraduate students in their last year need to submit a final-year dissertation to complete their studies, and this is increasingly being done in English. To investigate how the word combinations in the texts by these students differ from the conventional collocations used in general academic English, we compiled a learner corpus consisting of 102 final year papers written in English by Spanish students pursuing a degree in English at the University of León, Spain, between 2013 and 2018. As English students, their level of proficiency in the language corresponds to the C1-C2 CEFR level, and is higher than that of students taking other degrees. The corpus, Sp-ACE (Spanish Academic Corpus of English), contains 1,392,855 running words.

Sp-ACE was compiled, stored and explored using Sketch Engine (Kilgarriff et al., 2014). This tool provides an array of options for extracting linguistic data, including wordlists and word sketches displaying the main collocates of a lemma sorted by grammar relation. In addition, Sketch Engine provides association scores based on logDice statistics, which are arguably more relevant to our work than other measures such as the T-score and mutual information (Rychlý, 2008; Gablasova et al., 2017).

This paper will present an analysis of the noun work, the most frequent noun in Sp-ACE that was also listed in the Academic Keyword List (Paquot, 2010). Its analysis serves as a model of how other academic lemmas in this (and other learner corpora) can inform user-centred lexicographic resources and language teaching.

A word sketch for the noun work was retrieved to display the main lemmas surrounding it in Sp-ACE. The analysis focused on (a) its modifiers, (b) verbs taking work as an object, (c) verbs taking work as a subject, and (d) prepositions following work. The data were then compared with the logDice association scores of the same word combinations in the 71,372,972 million-word Oxford Corpus of Academic English (OCAE), used in the framework of the Oxford Learner's Dictionary of Academic English (Lea, 2014), as a standard reference for expert academic English. The reason for choosing a corpus like OCAE as a reference rather than a corpus like the British Academic Written English (BAWE) with student writing (Nesi, 2011) was that our aim was to not so much to investigate what is different about Spanish and comparable British student use of academic English, but rather to develop materials for helping Spanish students improve their use of collocations.

The results of our analysis disclose a number of discrepancies between how Spanish students and expert users of academic English employ the noun work in context, focussing on collocation error, underuse and overuse. Our methodology can be employed to analyse other academic lemmas so as to help develop customized pedagogical materials for Spanish EAP users. Similarly, Sp-ACE can be extended to include dissertations by students from other fields, as well as texts by other Spanish users of academic English, such as postgraduate students and academics.

References

Ackermann, K., & Chen, Y. (2013). Developing the academic collocations list (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, 235–247.

ColloCaid (no date). http://www.collocaid.uk/ (accessed on 11/01/2019).

Frankenberg-Garcia, A., Lew, R., Roberts, J.C., Rees, G.P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23–39.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in Corpus-based Language Learning Research: Identifying, Comparing and Interpreting the Evidence. *Language Learning*, 67(1), 155–179.

Granger, S. & Meunier, F. (2008). Phraseology in language learning and teaching: where to from here? In F. Meunier, & S. Granger (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 247–252). Amsterdam/Philadelphia: John Benjamins.

Granger, S., & Paquot, M. (2015). Electronic lexicography goes local. Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), 118–141.

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

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 4–44.

Jenkins, J. (2014). *English as a Lingua Franca in the International University*. London: Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M, Kovvář, V., Michelfeit, J., & Suchomel, V. (2014). *The Sketch Engine: Ten years on*. Lexicography, 1(1), 7–36.

Lea, D. (Ed.). (2014). *Oxford Learner's Dictionary of Academic English*. Oxford: Oxford University Press.

LEAD (no date) https://uclouvain.be/en/research-institutes/ilc/cecl/lead.html (accessed on 11/01/2019).

Nesi, H. (2011). BAWE: an introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New Trends in Corpora and Language Learning* (pp. 213-228). London: Continuum.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing. From Extraction to Analysis*. London: Bloomsbury.

Peters, E., & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes*, 20, 28–39.

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN* (pp. 6-9). Brno: Masaryk University.

Sketch Engine (no date). https://www.sketchengine.eu (accessed on 11/01/2019).

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

Matching the CEFR with Linguistic Measures. A Pilot Study Based on Vocabulary Measures in a Corpus of German-speaking Learners of French as a Foreign Language

Katia Rey, Anita Thomas
Université de Fribourg
katia.rey@unifr.ch, anita.thomas@unifr.ch

Since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has played a fundamental role in the teaching of modern languages in Europe (Goullier, 2008). The six levels (A1, A2, B1, B2, C1, C2) are widely used by school authorities, textbook and exam developers as well as language teachers (Forel & Gerber, 2013). The different scales that describe what learners can do with their language(s) were developed by combining qualitative, quantitative as well as intuitive methods (Council of Europe, 2001). However, it appears that the six levels lack some empirical foundation, especially from second language (L2) learners' production data (Hulstijn, 2014). Studies comparing the CEFR's scales with traditional linguistic measures that reflect second language proficiency have shown that they are often difficult to match. For example, in a study comparing L2 French, English and Italian written productions, Gyllstad, Granfeldt, Bernardini & Källkvist (2014) found only weak to medium-strong correlations between measures of syntactic complexity and the assigned CEFR levels. This means that there is a need for further studies on the relationship between the CEFR levels and linguistic measures.

Based on a corpus of L2 learners at beginner and intermediate levels of proficiency, the aim of this paper is to discuss in what way communicative abilities as described in the CEFR scales are reflected by L2 learners' linguistic skills. We will concentrate on the development of vocabulary in oral production with a particular focus on the vocabulary range scale (Council of Europe, 2001, p. 112). In this scale, the terminology appears vague and therefore may be difficult to apply in practice (Milton, 2010). Some questions that arise are: What are the "basic words" expected at level A1? What does "sufficient vocabulary" mean at level B1? How "good" should the "command of idiomatic expressions and colloquialism" be at level C1? Based on these questions, we developed the following two research questions: 1) How can the descriptors proposed by the CEFR be operationalized through linguistic criteria? 2) Do learner texts on different CEFR levels differ with respect to the linguistic criteria?

In order to explore these questions, we used traditional measures of vocabulary development in second language research (Bulté & Housen, 2012; Read, 2000). Measures of lexical diversity (*MTLD* – McCarthy, 2005), lexical density (*the number of content words has been divided by the total number of words* – Michel, Kuiken, & Vedder, 2007) and lexical sophistication (*Guiraud Advanced* – Daller, Van Hout, & Treffers-Daller, 2003), as well as the use of formulaic sequences (Forsberg, 2006) were investigated. The study is based on data from a corpus of ten German-speaking learners of French as a foreign language at secondary school in an oral interaction task. The learners' productions have previously been evaluated by CEFR experts as having reached the following five levels of competence regarding the vocabulary range: A1.1, A1.2, A2.1, A2.2 and B1.1. The first results show that while the most advanced learners of the sample are those who score the highest, the measures of lexical density, diversity and sophistication indices do not increase gradually across levels. For instance, the production evaluated at level A1.1 obtains a better score of lexical diversity than the one evaluated at level A2.1. Based on earlier comparative studies, these results are not completely unexpected. One explanation could be that the experts in charge of evaluating the learners' productions considered different aspects of vocabulary development to be important. They may also have been more sensitive to criteria related to other linguistic phenomena than to vocabulary such as the quality of oral production and especially fluency. A final explanation could be that these measures do not allow a distinction to be made between such fine levels (*see* Prodeau, Lopez, & Véronique, 2012).

References:

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Language Learning & Language Teaching* (pp. 21–46). Amsterdam, Pays-Bas: John Benjamins Publishing Company.

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

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical Richness in the Spontaneous Speech of Bilinguals. *Applied Linguistics*, *24*(2), 197‑222. DOI: 10.1093/applin/24.2.197

Forel, C., & Gerber, B. (2013). L'apprentissage des langues au-delà de la linguistique : le CECR. *Cahiers Ferdinand de Saussure*, *66*, 81–95. Retrieved from https://www.jstor.org/stable/24324191?seq=1#metadata_info_tab_contents.

Forsberg, F. (2006). *Le langage préfabriqué en français parlé L2 : Étude acquisitionnelle et comparative* (Thèse de doctorat présentée devant le Département de Français, d'Italien et de Langues classiques). Université de Stockholm, Stockholm, Suède.

Goullier, F. (2008). La mise en œuvre du Cadre européen commun de référence pour les langues en Europe. Une réalité différenciée dans ses finalités et dans ses modalités. *Revue internationale d'éducation de Sèvres*, *47*, 55‑62. DOI: doi.org/10.4000/ries.367

Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*, 1–30. DOI: 10.1075/eurosla.14.01gyl

Hulstijn, J. H. (2014). The Common European Framework of Reference for Languages: A challenge for applied linguistics. *ITL - International Journal of Applied Linguistics*, *165*(1), 3–18. DOI: 10.1075/itl.165.1.01hul

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* (Thèse de doctorat). University of Memphis, Memphis, USA.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3). DOI: 10.1515/iral.2007.011

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211-232). Eurosla Monograph Series I.

Prodeau, M., Lopez, S., & Véronique, D. (2012). Acquisition of French as a Second Language: Do developmental stages correlate with CEFR levels? *Apples - Journal of Applied Language Studies*, *6*(1), 47–68. Retrieved from https://jyx.jyu.fi/handle/123456789/40865

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press

Annotation of phonetic errors in Portuguese L2 texts

Iría del Río, Adelina Castelo, Rita Santos and Maria João Freitas
CLUL, University of Lisbon
igayo@letras.ulisboa.pt, adelina.castelo@gmail.com, rnazares@gmail.com, joaofreitas@letras.ulisboa.pt

This work presents an annotation schema for phonetic errors in written L2 Portuguese. We describe the categories and their motivation, and we present the results of an IAA test. These results show that the schema is consistent and reliable and indicate that further refinements are necessary for the annotation guidelines.

Most error annotation tag sets for L2 written data do not consider the phonetic level (Díaz-Negrillo & Fernández-Domínguez, 2006) and focus on levels more traditionally linked to writing, as grammar or spelling. For the annotation of errors in COPLE2 (Mendes et al., 2016) we were interested in analyzing the potential interaction between the phonetic output forms of words and the written errors produced by the learners. With this goal, we developed an annotation schema for written errors triggered by phonetic aspects of European Portuguese. We developed the schema considering the European Portuguese phonological system and error classifications previously proposed by different authors (Pinto, 1997; Castro & Gomes, 2000; Sousa, 1999; Leiria, 2001; Horta & Martins, 2004; Gonçalves, Guerreiro & Freitas, 2011). In the development phase, we validated its consistency and adequacy through two annotation experiments with data from the COPLE2, and we made the necessary adjustments. Finally, we ran an IAA experiment to evaluate its performance.

The phonetic level works in parallel with the error annotation system described in (del Río & Mendes, 2018). Phonetic errors are identified whenever a written word should be pronounced with a non-target phonetic form due to its misspelling. There are 5 categories of phonetic errors: substitution, deletion, addition, transposition and stress. The first four categories consider the phonetic segment as their unit (e.g. [p], [o], [j]), although several segment errors also enable identifying syllable structure problems. The last category is based on the stressed/unstressed syllable contrasts. In substitution errors, a phonetic segment is replaced in the misspelled word's pronunciation (*confução [s] instead of *confusão* [z]). Deletion errors present the suppression of a phonetic segment (*s_pportei [Ø] instead of *suportei* [u]). Addition errors exhibit the inclusion of a new segment, (*saludações [l] instead of *sa_udações* [Ø]). In transposition errors the order of two segments is reversed (*pregunta instead of *pergunta*). Finally, stress errors correspond to a change in word stress position due to the misspelling (*inicio instead of *início*). We expect that these categories may be useful both to researchers in L2 phonology and Portuguese L2 teachers. They can be used by L2 phonologists to cue phonological deviations from the target system in terms of segments, syllable structure and stress. Moreover, they may be used to identify and characterize different stages of the learners' interphonology. Portuguese L2 teachers may also consider these categories, while choosing the phonetic topics to deal with in their didactic materials and approaches. For example, substitution errors may show which phonetic segments are not well represented and are competing in the learners' interphonological system.

To test the reliability of the schema, we ran an experiment using spelling errors. Two annotators tagged 234 words with spelling problems, using the five categories plus an extra category ("zero") to identify the cases where the misspelling did not affect the phonetic level. We used two IAA measures: Fleiss kappa and Krippendorff alpha. We got the same number for both measures, 0.72, which can be considered a good result. The main discrepancies in the annotation appear when the word produced by the learner exhibits several substitutions, additions, etc., at the same time. This problem shows that we have to pay particular attention to this aspect in future annotation guidelines.

References

Castro, S. L., & Gomes, I. (2000). *Dificuldades de Aprendizagem da Língua Materna,* Lisboa, Universidade Aberta.

Del Río, I., & Mendes, A. (2018). Error annotation in the COPLE2 corpus. *Revista Da Associação Portuguesa De Linguística*, (4), 225–239.

Díaz-Negrillo, A. & Fernández-Domíguez, J. (2006). Error Tagging Systems for Learner Corpora. *RESLA*, 19: 83–102.

Gonçalves, F., Guerreiro, P., & Freitas, M. J. (2011). *O Conhecimento da Língua: Percursos de Desenvolvimento*, Lisboa, ME-DGIDC.

Horta, I. V., & Martins, M. A. (2004). Desenvolvimento e aprendizagem da ortografia: Implicações educacionais. *Análise Psicológica*, n.º 1 (XXII), pp. 213–223.

Leiria, I. (2001): *O léxico, aquisição e ensino do Português Europeu língua não materna,* Dissertação de Doutoramento. Universidade de Lisboa.

Mateus, M. H., & Andrade, E. (2000). *The Phonology of Portuguese.* Oxford UP.

Mendes, A., Antunes, S., Janssen, M. & Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. In *Proceedings of LREC 2016*, Portorož, Slovenia.

Pinto, M. G. L. C. (1997). A ortografia e a escrita em crianças portuguesas. *Revista da Faculdade de Letras do Porto – Línguas e Literaturas*, 14, pp. 7–58.

Sousa, O. C. (1999). *Competência ortográfica e competências linguísticas*, Lisboa, ISPA.

"I believe we can assume with some certainty": the functions of singular and plural first-person pronouns in master's theses

Sylvi Rørvik
Inland Norway University of Applied Sciences
sylvi.rorvik@inn.no

Disciplinary differences means that courses in academic writing need to take into account the requirements of the specific disciplinary communities of which the students are aspiring members (Bruce, 2016, Jiang, 2017), and previous research has shown that student writers may struggle to comply with disciplinary conventions regarding the frequency and function of first-person pronouns (Hyland, 2002, Ädel, 2006, Callies, 2013, Leedham & Fernandez-Parra, 2017). This paper examines the functions of singular and plural first-person pronouns in master's theses written in English and Norwegian within the three disciplines of chemistry, literary studies, and sociology, in comparison with the use of such pronouns in PhD dissertations and published research articles. The overarching research questions are:

1. What are the differences and similarities in the functions of first-person pronouns between master's theses written in English and Norwegian within the disciplines of chemistry, literary studies, and sociology?
2. Does the use of first-person pronouns in master's theses in these disciplines differ from the way first-person pronouns are used in comparable PhD dissertations and published research articles, and if so, how?

In total, the dataset comprises approximately 12,500 instances of first-person pronouns divided between the three disciplines, the two languages, and the three text categories (master's theses, PhD dissertations, and research articles). Each occurrence was categorized according to a framework based on Sheldon (2009).

Preliminary results indicate that there are disciplinary similarities among the master's theses in the sense that the most frequent function for both singular and plural pronouns in both languages is that of reference to the conductor of research. However, the disciplines differ in how frequently this function occurs, and there are also cross-linguistic differences when it comes to which discipline has the highest frequencies of this function. For the English singular pronoun, references to the conductor of research occur most frequently in chemistry, followed by sociology and literary studies. For the Norwegian singular pronoun, the highest frequency is found in sociology, followed by chemistry and literary studies. Similarly, for the plural pronouns, in English the highest frequency is found in chemistry, followed by literary studies and sociology. In Norwegian, the highest frequency is found in literary studies, followed by chemistry and sociology. Further results will be reported on in the presentation.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Bruce, I. (2016). Constructing critical stance in University essays in English literature and sociology. *English for Specific Purposes* 42, 13–25.

Callies, M. (2013). Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics* 18(3), 357–390.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics* 34, 1091–1112.

Jiang, F. (2017). Stance and voice in academic writing: The "noun + that" construction and disciplinary variation. *International Journal of Corpus Linguistics* 22(1), 85–106.

Leedham, M., & Fernandez-Parra, M. (2017). Recounting and reflecting: The use of first person pronouns in Chinese, Greek and British students' assignments in engineering. *Journal of English for Academic Purposes*, 26, 66–77.

Sheldon, E. (2009). From one *I* to another: Discursive construction of self-representation in English and Castilian Spanish research articles. *English for Specific Purposes*, 28, 251–265.

Assessing the cross-linguistic validity of phraseological complexity measures as indices of L2 proficiency

Rachel Rubin, Alex Housen, & Magali Paquot
Vrije Universiteit Brussel, Université catholique de Louvain
rachel.rubin@vub.be, alex.housen@vub.be, magali.paquot@uclouvain.be

Linguistic complexity has been investigated and measured at various dimensions of language (syntax, morphology, lexicon), but has not, until very recently, been considered at the linguistic interfaces. This research is a response to a recent call to widen the scope of L2 complexity research to the lexis-grammar interface (Paquot, 2018; Housen et al., 2018), building on Paquot (2018, 2019) who looked at the phraseological dimension of language in English as a Foreign Language (EFL) learner writing and found that measures of phraseological sophistication are better suited to index proficiency than measures of syntactic and lexical complexity, particularly at the B2 to C2 levels of the Common European Framework of Reference (CEFR). As a partial replication of Paquot (2018, 2019), we aim to establish the cross-linguistic validity of measures of phraseological complexity by assessing their effectiveness as indices of L2 Dutch (writing) proficiency. The research questions that we aim to answer are:

1. How effective are measures of phraseological complexity in indexing L2 Dutch (writing) proficiency?
2. How do measures of phraseological complexity compare to traditional measures of syntactic and lexical complexity in L2 Dutch written productions?
3. How does the development of phraseological complexity in L2 Dutch compare to the results observed for L2 English in Paquot (2018, 2019)?

This investigation expands on the metrics of lexical and syntactic complexity seen throughout L2 research by applying phraseological measures of diversity and sophistication outlined in Paquot (2018, 2019). The corpus used in this study is compiled from more than 2,700 written extracts (ca. 700,000 words) of the Dutch certification exam CNaVT (Certificaat Nederlands als Vreemde Taal), produced by learners of Dutch from multiple L1 backgrounds at the B1-C1 CEFR levels. Quantitative measures of phraseological diversity (root type token ratio—RTTR) and sophistication (MI-based) are computed for the target phraseological units, and are then compared with traditional measures of syntactic and lexical complexity. Finally, a mixed-effects regression analysis is carried out to determine which indices of complexity best contribute to an explanation of the proficiency levels assigned to the texts.

In line with Paquot's findings for L2 English, we expect to observe a significant increase in measures of phraseological complexity, but no such increase in measures of syntactic or lexical complexity, particularly as proficiency level increases from intermediate to advanced levels. Such results would serve to strengthen the cross-linguistic validity of measures of phraseological complexity as indices of L2 proficiency, as well as to support Paquot's (2018) argument for the inclusion of phraseological competence in foreign and second language education and assessment. Furthermore, the results of this investigation will contribute to our understanding of the development of linguistic complexity and phraseological competence in L2 Dutch, and their role in the assessment of L2 Dutch writing quality, filling several gaps in the L2 Dutch literature.

References

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2018). Multiple approaches to L2 complexity. *Second Language Research,* 35(1), 3–21.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research,* 35(1), 121–145.

Paquot, M. (2018). Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly,* 15(1), 29–43.

Vague language in the Lithuanian Learner Corpus

Jūratė Ruzaitė
Vytautas Magnus University
jurate.ruzaite@vdu.lt

The present paper has a twofold aim: to introduce the Lithuanian Learner Corpus (LLC), which is currently under construction, and to present a preliminary investigation of the use of vague language (hereinafter VL) in Lithuanian learner language. The study focuses on three main categories of VL: general extenders (GEs), e.g. *ir taip toliau* ('and so on'), vague quantifiers, e.g. *keletas* ('several'), and approximated quantities, e.g. *apie dvidešimt* ('around twenty'). Recent research on VL has shown that it constitutes an important part of pragmatic language competence and thus should be addressed in language teaching in a systematic way (e.g. Buysse, 2014; Lin 2013; Fernández, 2015). However, so far, research on learner language has mostly focused on linking words, writer positioning, hedging, and multi-word clusters, and VL has not been addressed yet either in widely used languages or lesser used ones, such as Lithuanian.

The research questions that the current study aims to answer are as follows: (1) How extensively and in what contexts do Lithuanian learners use VL?; and (2) How do the frequency results in learner Lithuanian compare to those in the corpus of native Lithuanian speakers? On the basis of the quantitative frequency results and a qualitative analysis of the concordance lines, some tentative implications of the results are discussed suggesting what could be taken into account when teaching Lithuanian as a foreign language.

This study is based primarily on the LLC, which currently consists of 140,442 words and contains written and spoken texts produced by students from various language backgrounds learning Lithuanian as a foreign language. The LLC contains texts written by beginning, pre-intermediate, intermediate and upper-intermediate learners of Lithuanian. It comprises a large variety of text types (essays, narratives, argumentative texts, letters, emails, postcards, interviews, etc.). The corpus uses the TEITOK programme developed by Maarten Janssen (2014-, http://www.teitok.org/), which is "a web-based framework for corpus creation, annotation, and distribution, that combines textual and linguistic annotation within a single TEI based XML document" (Janssen, 2016: 4037).

To contextualise the findings obtained from the LLC, learner corpus data is compared to native speaker language by resorting to the general reference corpus *The Corpus of the Contemporary Lithuanian Language* (CCLL; tekstynas.vdu.lt). This corpus is over 140 mln words and represents different types of written and spoken discourse.

The preliminary data analysis has shown that in the LLC, VL items appear 712 times (50.7, f/10,000). There are some major differences in the frequency of the three VL categories under investigation. Vague quantifiers are the most frequent category (632 occurrences, or 45 per 10,000); general extenders are the least frequent category (21 occurrences, or 1.5 per 10,000); and approximators appear 59 times, or 4.2 times per 10,000. These findings are to some extent similar to the VL use in native Lithuanian. Quantifiers, similarly to learner Lithuanian, are the most numerous category in the CCLL (96.5 per 10,000). Approximators are the smallest category (9.3 occurrences per 10,000), and general extenders are the second largest category (45.1 times per 10,000 words). The total number of VL items in native Lithuanian make up 150.9 occurrences per 10,000 words, which is approximately three times as frequent as in non-native Lithuanian.

In terms of VL frequency in different levels of language proficiency, the findings tentatively suggest that the higher the proficiency, the more frequently VL is used. However, some items (e.g. *šiek tiek*, 'a little bit') are used most extensively in the beginner level, which can be directly related to the course curriculum, where learners are taught, for example, to say that they speak Lithuanian a little bit.

Concerning the variation of form of VL items in the LLC, it is evident that VL items are restricted to a very limited number of types (e.g. there are only two types of general extenders), whereas in native Lithuanian there exists a much a higher variety (e.g. more than 20 types of general extenders).

To generalise, the low frequency and variety of VL types in learner Lithuanian point to the underuse of VL in Lithuanian learner language. On the one hand, the differences in frequency and form could be influenced by the nature of the learner corpus, which is differently balanced and is considerably smaller in size if compared to the

CCLL. On the other hand, it can be assumed that VL is not sufficiently represented in language teaching curricula.

References

Buysse, L. (2014). 'We went to the restroom or something': General extenders and stuff in the speech of Dutch learners of English. In J. Romero-Trillo (Ed.), *The Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, pp. 213–237. http://link.springer.com/chapter/10.1007/978-3-319-06007-1_10; accessed 08.10.2015.

Janssen, M. 2016. *TEITOK: Text-Faithful Annotated Corpora*. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia, pp. 4037–4043.

Fernández, J. (2015). General extender use in spoken Peninsular Spanish: metapragmatic awareness and pedagogical implications. *Journal of Spanish Language Teaching*, 2 (1), 1–17.

Lin, Y. L. (2013). "Vague language and interpersonal communication: an analysis of adolescent intercultural conversation". *International Journal of Society, Culture and Language*, 69–81.

The longitudinal development of clausal and noun-phrasal complexity in German intermediate learners of English

Steffen Schaub
University of Bremen
steffen.schaub@uni-bremen.de

Recent research into the syntactic complexity of learner English has increasingly taken into account measures of noun-phrasal complexity, both for academic learners (Parkinson & Musgrave 2014) and for pre-academic learners (Bulté & Housen 2018, Kreyer & Schaub 2018). The impetus for much of this interest can be tracked to Biber, Gray & Ponpoon (2011), who call for a more diversified understanding of syntactic complexity in the study of learner language, i.e. one which takes into account phrasal complexity features (in addition to clausal measures). The underlying assumption is that beginning learners produce complexity at the clausal level, influenced by a conversational writing style, whereas advanced learners increasingly rely on phrasal complexity, aiming at an academic writing style. It can thus be hypothesized that the development from clausal to phrasal complexity already sets in during the intermediate stage of the language learning process.

The present study is a longitudinal, corpus-based analysis of clausal and noun-phrasal complexity in the written exams of German EFL learners (age: 13-18). Syntactic complexity is operationalized as the number and range of dependent structures (dependent clauses, NP modifiers) per independent structure (main clause, NP). The study seeks to answer two research questions:

- RQ1: How do clausal and phrasal complexity measures develop in the written language of German intermediate EFL learners over a long period of time?
- RQ2: How do the learners incorporate linguistically complex input from the task material into their own writing?

The data is drawn from the Marburg corpus of Intermediate Learner English (MILE) (Kreyer 2015), a longitudinal corpus of written exam texts produced by a cohort of German secondary school pupils over a 4-year period. The study is based on a multi-layer annotated subsample of ten pupils with German as their L1. The semi-automatic annotation process involves i) part-of-speech tagging, ii) automatic segmentation into clauses and phrases (with manual correction), iii) manual categorization of clause and noun phrase types, iv) manual annotation of clausal and phrasal modification, and v) annotation of extra-linguistic features (age, task type, etc.).

Preliminary results for RQ1 confirm previous findings that group measures of complexity generally tend to increase over time, while learner-individual trajectories vary considerably (cf. Bulté & Housen 2018). RQ2 is approached by means of a qualitative analysis of complex structures in the task material (task descriptions, supplementary material) and the students' writing. The results show that students incorporate complex structures from the input material into their own writing.

Overall, the findings of the study contribute to our understanding of the emergence of syntactic complexity in the writing of intermediate learners, and highlight the variable, idiosyncratic nature of interlanguage systems.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly* 45(1), 5–35.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics* 28(1), 1–18.

Kreyer, R., & Schaub, S. (2018). *The development of phrasal complexity in German intermediate learners of English* 4(1). 82–111.

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies, & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* (pp. 13–34). Amsterdam: John Benjamins.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes* 14, 48–59.

Introducing language teachers to learner corpora: The development of online tutorials for pedagogic uses of the MuSSeL corpus

Erin Schnur, Fernando Rubio, Jane Hacking
The University of Utah
erin.schnur@utah.edu, fernando.rubio@utah.edu, j.hacking@utah.edu

This presentation will describe the process of designing a series of online tutorials on pedagogical uses of the Multilingual Spoken Second Language Speech (MuSSeL) learner corpus (https://l2trec.utah.edu/database/). The tutorials are aimed at elementary and secondary foreign language teachers in Utah's Dual Language Immersion (DLI) program, but available to foreign language instructors in other contexts.

Much of the research exploring the direct use of corpora in language instruction/classrooms focuses on the use of native speaker (NS) corpora as a representation of target language use (Chambers, 2015). Still, regardless of corpus type (NS or learner), and despite the potential of corpus linguistic tools and methods for direct classroom application, few language teachers utilize corpora in their classrooms or to inform their instruction (Boulton, 2011; Ebrahimi & Faghih, 2017). Discussions of this gap between potential and practice have largely focused on the need for training for both pre- and in-service teachers. Recent studies on such training have shown that, when teachers are introduced to corpus linguistics and provided adequate training, it is effective (Heather & Helt, 2012), with most teachers believing that corpora and corpus tools can benefit language teachers (Mukherjee, 2004).

In creating online tutorials, this project has two goals: 1) to contribute to the discussion of how learner corpora, and specifically the MuSSeL corpus, can be a valuable resource to inform instruction, and 2) to offer free, online training for teachers who wish to incorporate corpus linguistics into their pedagogical tool set.

While the tutorials introduce general corpus linguistic principles and techniques using AntConc concordancing software, they focus specifically on using the MuSSeL corpus. The corpus draws from oral proficiency exam data and comprises texts from L2 speakers of six languages: Chinese, French, German, Portuguese, Russian and Spanish, with speech samples collected in both the 3rd and 5th grades.

This presentation will discuss the process of planning and creating the tutorials, as well as attempts, at each step, to facilitate audience buy-in. These steps include: identifying the information and technical skills that should be included, deciding upon the mode of delivery, creating how-to documentation for specific corpus methods, choosing example research questions and problem sets, and linking those examples to typical instructional and/or curricular goals and objectives. While the creation of the tutorials is ongoing, we will share the first tutorial and accompanying feedback received from a focus group of foreign language instructors in the DLI program, followed by discussion of how this feedback has impacted future direction of the project.

References

Boulton, A. (2011) Bringing corpora to the masses: Free and easy tools for language learning. In N. Kübler, (Ed.), *Corpora, language, teaching, and resources: From theory to practice*. Bern: Peter Lang, 69–96.

Chambers, A. (2015). The learner corpus as a pedagogic corpus. *Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press, 445–464.

Ebrahimi, A., & Faghih, E. (2017). Integrating corpus linguistics into online language teacher education programs. *ReCALL*, *29*(1), 120–135.

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

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. *Language and Computers 52*, 239–250.

Graph-based modeling of Lexicosyntactic Coselection Constraints in Learner German

Anna Shadrova
Humboldt University of Berlin
shadrova@hu-berlin.de

Based on the ideas of Sinclair's 'idiom principle' (Sinclair 1991) and a number of studies from the wider field of usage-based linguistics that reflect the idea of forces of attraction and repulsion between lexical and lexicosyntactic elements (Herbst 2015, Bartsch & Evert 2014 , Faulhaber 2011, Gries & Stefanowitsch 2004 among many others), this study looks into the acquisition of such lexical constraints on the coselection of verbs and their arguments at different stages of acquisition in learners of German. The talk will discuss two main hypotheses, a), that learners combine more freely compared to native speakers even at late stages of acquisition, and b), that learners show a u-shaped learning trajectory in coselectional constraints rather than a linear development. Both will be confirmed for the Kobalt corpus (Zinsmeister et al. 2012), a medium-sized, strictly controlled corpus of overall 151 essays written by learners of German from universities in China and Belarus and a German-L1 control corpus of 20 essays.

For this, a new, graph-based approach is presented that models lexemes as nodes and syntactic dependency as edges of a network graph upon which graph metrics for community detection and overall connectivity of the graph are computed (Louvain modularity, Blondel et al. 2008) and compared across several splits of the data based on language test score ranges in a cross-sectional or 'quasi-longitudinal' design. Until recently, graphs have mostly been used in the humanities for visualizing existing analyses or for exploratory purposes. The talk will discuss how using graph metrics allows for quantitative analysis of relatively abstract, structural properties and the interrelations of all lexemes within a corpus rather than individual words and their combinations as is the case in statistical computations of co-occurrence measures.

Results indicate that learners do indeed show fewer lexical constraints than native speakers. An initial lexical diversification followed by a differentiation can be viewed as a u-shaped learning curve or acquisition trajectory as predicted by the hypotheses, with expectably different results for varying verb argument structures. L1-specific effects do appear, where learners with L1 Russian/Belarussian show a higher flexibility than the Chinese-L1 learners and the German native speakers, which might be explained typologically and/or from a language teaching perspective. Methodological aspects such as the verification of results through sampling, the acceptability of the 'quasi-longitudinal' perspective and the specific advantages of graph metrics for small to medium-sized corpora will be discussed.

References

Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern* 2, 48–61.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008(10), P10008.

Faulhaber, S. (2011). Verb Valency Patterns: A challenge for semantics-based accounts. *Topics in English Linguistics* vol. 71. Berlin/New York.

Gries, S. & A. Stefanowitsch (2004). Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1), 97–129.

Herbst, T. (2015). Why Construction Grammar Catches the Worm and Corpus Data can Drive you Crazy: Accounting for Idiomatic and Non-Idiomatic Idiomaticity. *Journal of Social Sciences* 11 (3), 91–110.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., & Skiba, D. (2012). Das Wissenschaftliche Netzwerk „Kobalt-DaF". *Zeitschrift für germanistische Linguistik* 40(3), 457–458.

The effect of time and dimensions of collocational relationship on phraseological accuracy: a study on Chinese learners of Italian

Stefania Spina
University for Foreigners of Perugia
stefania.spina@unistrapg.it

A vast amount of work in the area of Learner Corpus Research (LCR) has been devoted to the analysis of phraseology in learner language, revealing the critical role of the phraseological dimension in the processing and production of learner language (e.g. Ellis et al., 2015; Nesselhauf, 2005; Wang, 2016). However, there are still at least two issues that LCR has to face in the analysis of phraseology. Firstly, much of this work has been done with data collected at one point in time. Secondly, phraseological units are often only analysed in terms of their frequency, strength of association (Bestgen & Granger, 2018) or comparison with those produced by native speakers, but not in terms of their accuracy in context (Thewissen, 2015). Investigations of phraseological errors are key to obtain information on the quality of the word combinations produced by learners, and on the different stages in their process of gaining accuracy. Again, much of the work on such errors has been done with learner data collected at one point in time: longitudinal studies on phraseological errors are still rare (Bartning & Forsberg, 2006, Crossley & Salsbury, 2011, Qi & Ding, 2011, and Spina, forthcoming).

This study tries to fill both the above-mentioned gaps, using an error-annotated sample of the *Longitudinal Corpus of Chinese Learners of Italian* (Spina & Siyanova-Chanturia, 2018) and mixed-effect models (Cunnings & Finlayson, 2015). The sample includes sixty essays, of which thirty from data-collection 1, written by ten learners for each of the A1, A2 and B1 CEFR levels, and thirty from data-collection 2 - six months later - written by the same thirty learners.

The present study analyses phraseological errors of beginner and pre-intermediate Chinese learners of Italian in the word combinations used within the adjectival modifier grammatical dependency (noun+adjective and adjective+noun) and the verb-direct object dependency, trying to verify: a) if time affects the accuracy of the selected word combinations, and b) if this effect varies for the different type of combinations, and for the three considered proficiency levels (A1, A2 and B1).

In addition, the effect on accuracy of the different dimensions of collocational relationship (Brezina et al., 2015), each one represented by a specific measure, is considered. This issue is key for a deeper understanding of the process of acquisition of formulaicity by L2 learners (Gablasova et al., 2017). To date, however, little empirical evidence is available about the degree to which some of these dimensions (the repetition, strength, directionality and type-token distribution, represented by the measures of frequency, mutual information, DeltaP and lexical gravity) affect accuracy in the use of word combinations.

Accordingly, the three research questions that this study seeks to answer are:

- Does time affect the accuracy of noun+adjective, adjective+noun, verb+$noun_{do}$ word combinations?
- Does this effect differ for the different types of combinations, and for the three proficiency levels A1, A2 and B1?
- Are different dimensions of collocational relationship (repetition, strength, directionality and type-token distribution), represented by the measures of frequency, mutual information, DeltaP and lexical gravity, significant predictors of phraseological accuracy?

Preliminary results of the mixed-effect models built with accuracy as a dependent variable show that the effect of time varies significantly across the three combination types, and that the effect of the different collocational relationships varies across the different measures, revealing that the developmental patterns of phraseological accuracy are basically slow and uneven (Bestgen & Granger, 2014).

References

Bartning, I., & Forsberg, F. (2006). Les séquences préfabriquées à travers les stades de développement en français L2. In *Actes du 16e congrès des romanistes scandinaves*. Department of Language and Culture, Roskilde University.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26, 28–41.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L.M. Dillmann (Eds.), *Corpora and* Lexis (pp. 277–301). Leiden & Boston: Brill.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2), 139–173.

Crossley, S., & Salsbury, T.L. (2011). The Development of Lexical Bundle Accuracy and Production in English Second Language Speakers. *International Review of Applied Linguistics in Language Teaching* (IRAL), 49(1), 1–26.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York: Routledge.

Ellis, N., Simpson-Vlach, R., Römer, U., O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in SLA. In S. Granger, G. Gilquin, & F. Meunier (Eds), *The Cambridge Handbook of Learner Corpus Research* (pp. 357–378). Cambridge: Cambridge University Press.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning* 67(S1), 155–179.

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

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

Qi Y., & Ding Y. (2011). Use of formulaic sequences in monologues of Chinese EFL learners. *System 39*, 164–174.

Spina, S. (forthcoming). The development of phraseological errors in Chinese learner Italian: a longitudinal study. In *Proceedings of Learner Corpus Research Conference 2017*. Louvain: Presses Universitaires de Lovain.

Spina, S., & Siyanova-Chanturia, A. (2018). *The Longitudinal Corpus of Chinese Learners of Italian (LOCCLI).* Poster presented at the 13th Teaching and Language Corpora conference, University of Cambridge, UK.

Thewissen, J. (2015). *Accuracy across Proficiency Levels. A Learner Corpus Approach*. Louvain: Presses Universitaires de Louvain.

Wang, Y. (2016). *The Idiom Principle and L1 Influence. A contrastive learner-corpus study of delexical verb+noun collocations.* Amsterdam: John Benjamins.

An integrative approach to L2 accuracy and complexity development

Jennifer Thewissen, Alena Anishchanka
University of Antwerp
Jennifer.thewissen@uantwerpen.be, alena.anishchanka@uantwerpen.be

The complexity, accuracy, fluency (CAF) triad has been the object of thriving research activity, both in second language acquisition and learner corpus research (LCR). To-date, LCR has tended to investigate the CAF constructs in isolation of each other. While studying these dimensions one by one has helped refine the theoretical understanding and operationalisation of each construct separately, Norris and Ortega (2009) rightly claim that more attention needs to be paid to "CAF as a dynamic and interrelated set of constantly changing subsystems". Answering the call for a more "integrative approach" (Larsen-Freeman, 2009) to the study of CAF, this paper proposes to use learner corpus data to capture the dynamic interaction between accuracy and complexity across proficiency levels in EFL writing. The current paper is an exploratory study which addresses the following two research questions:

(1) Is it the case that accuracy helps discriminate between certain proficiency levels and complexity between others?
(2) How do the two constructs interact across the B1 to C2 proficiency range? Do they operate in support of each other and/or do they show signs of trade-offs?

This study uses a sample of 223 EFL texts from the *International Corpus of Learner English* (Granger et al., 2009), amounting to c. 150,000 tokens. To allow for a developmental approach, each text from the sample was professionally rated as either B1, B2, C1 or C2. To capture accuracy, all 223 texts were manually coded for errors according to the Louvain error tagging taxonomy which includes 40 different error subtypes. Complexity was operationalised as syntactic and lexical complexity. The texts were analysed with the L2 Syntactic Complexity Analyzer (Lu, 2010) which provides 14 indices related to syntactic complexity and the Lexical Complexity Analyzer (Lu, 2012) which taps into lexical variation, sophistication and density. The 223 texts in the corpus sample thus come hand in hand with an individual profile which includes the proficiency level of each text, the number of errors per text, and the syntactic and lexical complexity measures. Inferential statistics were subsequently used to identify which accuracy and complexity measures significantly differentiate between the proficiency levels and explore the nature of the interaction between the two constructs.

With regard to accuracy, results show that global accuracy (i.e. the total errors) is a useful discriminator between all four proficiency levels. However, local accuracy (the number of errors in specific error categories) displays more moderate discriminatory power as local error types tend to progress between certain proficiency levels while stabilising between others. Complexity results are being added to those yielded for accuracy, with an aim to empirically showing that "not all traits of CAF will have an equally predictive value for all proficiency levels" (Norris & Ortega, 2009: 573).

References

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

Larsen-Freeman, D. (2009). Adjusting expectations: the study of complexity, accuracy and fluency in second language acquisition. *Applied Linguistics* 30(4), 579–589.

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 Language Journal* 96(2), 190–208.

Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics* 30(4), 555–578.

Restructuring of case system in non-standard Russian in Finland: Evidence from the Russian Learner Corpus

Ekaterina Vlasova, Maria Hokkanen
National Research University Higher School of Economics in Moscow, University of Helsinki
evlasova@hse.ru, maria.hokkanen@helsinki.fi

This paper examines restructuring of Russian case system in written production of advanced students whose dominant language is Finnish. The non-academic essays were collected at the University of Helsinki, where advanced Russian learners fall into two groups: a) Russian heritage speakers b) Finnish learners of Russian as a foreign language. The data is annotated by the means of the Russian Learner Corpus assembled with automatic part-of-speech annotation and manual error tags. This study seeks corpus-informed quantitative support for the following research questions related to the second language acquisition.

1. Do advanced learners with different backgrounds (heritage speakers vs foreign language learners) simplify the case system in the same way and make similar morphosyntactic errors?

2.What are the Russian case forms that undergo attrition and what are the default forms compensating the loss?

As the recent studies confirm, language learners tend to simplify the nominal paradigm and prefer the Nominative over oblique case forms, for the overview see (Polinsky, 2018). The error-focused corpus study of Russian learners, dominant in Finnish, provides new insights into the acquisition of morphosyntax and case paradigm, as the both languages have rich inflectional morphology. This paper analyses deviational prepositional phrases which contain inappropriate case forms, for example, the overuse of the Genitive or the Nominative instead of the Prepositional:

a. bukv na anket-ov.GEN.PL
bukv na anket-ah.PREP.PL
'letters on the forms'

b. Timo sidel v kresl-o.NOM.SG
Timo sidel v kresl-e.PREP.SG
'Timo was sitting in an armchair'

The proportional analysis of the detected deviational forms demonstrates a significant difference between learners with different backgrounds. The foreign language learners favour the Nominative (48% of errors) as an unmarked form over wide range of oblique cases in the prepositional phrase. In contrast to this common pattern, the heritage speakers overuse the Genitive (55% of errors) as a default oblique case, and this is a well-documented phenomenon in Russian vernaculars in Finland. One hypothesis (Leisiö, 2004) claims that Russian-Finnish bilinguals overuse the Genitive as a cross-linguistic transfer from Finnish, where adpositional phrases can only contain the Genitive and the Partitive, which are functionally close to the Genitive in Russian (Mustajoki, 1984).

The different morphosyntactic patterns of Russian advanced learners from Finland suggest a few implications. First, the attrition of oblique cases in favour of the Nominative is indeed a common phenomenon, but not universal. The heritage learners speaking Russian and Finnish from early childhood overuse the oblique Genitive form and create their own pattern of prepositional phrase induced by the dominant Finnish language. Second, the different patterns of morphosyntactic simplification demonstrate effects related to the learners age and background. Although the heritage speakers and the advanced foreign learners study Russian in the same classroom and speak the same dominant language, the cross-linguistic influence in morphosyntax only occur among heritage speakers, who have been learning Russian and Finnish from early age and in natural communication.

References

Leisiö, L. (2004). Sijamuotojen käytöstä suomenvenäjässä. In *Virittäjä* (2), 162–199.

Mustajoki, A. (1984). O raznyh stepenjah sootvetstvija glagol'nogo upravlenija v russkom i finskom jazykah. In *Studia Slavica Finlandensia* I (pp. 73–87). Helsinki: Neuvostoliittoinstituutti.

Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

L1-specific difficulties in L2 German: A learner corpus-based study on the use of prepositions by learners with typologically different first languages

Tassja Weber
University of Mannheim
tasweber@mail.uni-mannheim.de

Research has shown an influence of first language (L1) on the use of prepositions and prepositional phrases (PP) by learners of German as a foreign language (GFL) (Grießhaber, 2007; Turgay, 2010; Bryant, 2012). However, quantitative, corpus-based studies are still lacking. The present study[6] seeks to close this research gap. It presents a learner corpus-based contrastive interlanguage analysis (Granger 2015) on the effect of L1 on the use or prepositions, i.e. PP by GFL learners. 'Difficulties' are operationalized as accuracy and error type in prepositional use. The research questions are:

(1) Do accuracy rates and (2) error types in the use of prepositions differ according to the L1 and if so, how?

Data is taken from the German sub corpus of the learner corpus MERLIN (Abel et al., 2014) ($MERLIN_G$). $MERLIN_G$ contains 1033 texts from high-quality language tests written by GFL learners with different L1 backgrounds and rated according to the levels of the Common European Framework of Reference (CEFR). $MERLIN_G$ includes multi-layer (error) annotations as well as target hypotheses (Lüdeling, 2008).

Since the typological distance between L1 and L2 is considered to affect the acquisition of prepositions and PPs (Lütke, 2011: 111), I distinguish between typologically similar L1s (L1TS) and typologically different L1s (L1TD). *English* and *Italian* are considered L1TS since both languages know prepositions which – similar to German – can display lexicalized or grammaticized use (Huddleston/Pullum, 2006: 603; Schwarze, 2011: 292). *Turkish* and *Hungarian* are considered L1TD because both languages do not know prepositions. Equivalents of German prepositions, i.e. PP, are prototypically realized in the form of case suffixes or, less often, via postpositions (Kalkavan-Aydin, 2017: 376; Forgács, 2004: 235f.).

For the corpus study, all PP-contexts[7] and corresponding error annotations were extracted from the sub corpora L1TS and L1TD in $MERLIN_G$. Only CEFR levels that were represented sufficiently in both sub corpora were considered for the analysis. All in all, more than 1,000 PP-contexts from the CEFR levels A2, B1+ and B2 have been analyzed via logistic regression modelling (see Gries, 2015: 165ff.). I calculated a generalized (logistic) mixed-effects model using the *glmer* function from the package *lme4* (Bates et al., 2015) in *R* (R Core Team, 2018). In the model, I incorporated fixed factors[8] (L1 and CEFR level) and treated the learner as a random factor in order to control for individual effects of single learners (Tagliamonte/Baayen 2012: 143).

Overall results of the statistical analysis show that 1) the L1 has no effect on accuracy rates but 2) an effect on specific error types. Accuracy rates are similar for L1TS and L1TD and increase steadily independent of L1. With regard to error types in prepositional usage, however, the results indicate that the learners' L1 has a significant effect: on the one hand, learners with L1TD show a significant higher probability for errors in the realization of prepositions (β 1.38358=, SE= 0.42243, p= 0.00106), i.e. they leave out prepositions completely or use prepositions where no preposition is required. On the other hand, learners with L1TD show a significant lower probability for selecting the wrong prepositions compared to learners with L1TS (β = -1.38358, SE= 0.42243 p= 0.00106). Thus, the study shows that there are L1-specific difficulties in prepositional use. However, they do not become evident in accuracy rates but rather in specific error types.

References

Abel, A., Wisniewski, K., Nicolas, L. et al. (2014). A trilingual learner corpus illustrating European reference levels. *Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne* 2 (1), 111–126. http://www.ojs.unito.it/index.php/ricognizioni/article/view/702 (last accessed on 24 January, 2019).

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[6] This abstract presents partial results from a dissertation project.
[7] PP-contexts include instances of target like and not-target like use of prepositions..
[8] No interactions of fixed factors could be determined in the regression models.

Bryant, D. (2012). *Lokalisierungsausdrücke im Erst- und Zweitspracherwerb. Typologische, ontogenetische und kognitionspsychologische Überlegungen zur Sprachförderung in DaZ*. Baltmannsweiler: Schneider Verlag Hohengehren.

Forgács, T. (2004). *Ungarische Grammatik. 2*., verbesserte Auflage. Wien: EDITION PRAESENS. Verlag für Literatur- und Sprachwissenschaft.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7–24.

Grießhaber, W. (2007). "und wir faren in die andere seite" – Der Gebrauch lokaler Präpositionen durch türkische Grundschüler. K. Meng, & J. Rehbein (Eds.) *Kindliche Kommunikation - einsprachig und mehrsprachig*. Münster: Waxmann, 371–392.

Gries, S. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, F. Meunier (Eeds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge UP, 159–181.

Huddleston, R., Pullum, G. K. (et al.) (2006). *The Cambridge Grammar of the English Language*. 4th Edition. Cambridge: Cambridge/New York u.a: UP.

Kalkavan-Aydin, Z. (2017). Lokale Präpositionen im Deutschen und ihre Entsprechungen im Türkischen im Fokus kindlicher Spracherwerbsprozesse. In Y. Ekinci, E. Montanari, & L. Selmani (Eds). *Grammatik und Variation. Festschrift für Ludger Hoffmann zum 65. Geburtstag*. Heidelberg: Synchron, 375–387.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter, & P. Grommes (Eds.): *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitsprachenerwerbsforschung*. Tübingen: Niemeyer, 119–140.

Lütke, B**.** (2011)**.** *Deutsch als Zweitsprache in der Grundschule: Eine Untersuchung zum Erlernen lokaler Präpositionen*. Berlin/Boston: de Gruyter.

R Core Team (2018). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwarze, C. (2011). *Grammatik der italienischen Sprache*. 2., verbesserte Auflage. Berlin/Boston: de Gruyter.

Tagliamonte, S., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135–178.

Turgay, K. (2010). *Der Zweitspracherwerb der deutschen Präpositionalphrase: eine Studie zum Sprachentwicklungsstand von Kindern mit Migrationshintergrund*. Trier: WVT.

Linking CEFR levels to text quality indicators. An empirical investigation on the basis of the KOLIPSI learner corpus

Andrea Abel, Katrin Wisniewski
EURAC Bozen, University of Leipzig
andrea.abel@eurac.edu, katrin.wisniewski@uini-leipzig.de

In our exploratory paper, we try to link text quality indicators commonly used in L1 writing (development) research to CEFR-based proficiency ratings in an L2 and an L1 German learner corpus. Bridging these disciplines is challenging: In order to arrive at generalizable estimates of L2 proficiency, standardized criterion-oriented proficiency tests often use scales and levels such as those provided by the CEFR (2001). However, particularly in institutional educational settings, CEFR-based assessments are often considered vague, and the scales have been subject to much criticism. Furthermore, whereas language testing heavily builds on psychometric approaches and increasingly benefits from learner corpus research, writing research has traditionally focused on L1 competences and their development, rarely relying on language corpora.

In this paper, we will exploit both approaches to answer the following research questions (RQ):

- RQ1: Can selected text quality indicators such as text routines be found at specific CEFR levels and if so, which ones?
- RQ2: Does the use of selected text routines differ between L1 writers and L2 writers or is it a common challenge of general writing development?

Our data basis is a subcorpus of the KOLIPSI learner corpus. We use 150 texts (25000 tokens) in German as L2 produced on the basis of a standardized writing test task by high school students in South Tyrol (Italy) aged 17 years. Of these, 50 were rated A2, 50 were placed at B1 and 50 at a B2 level. These scores have reliably been linked to the CEFR. Second, we use 50 L1 responses to the same task (10000 tokens) written by students of the same age (Abel et al. 2012).

The corpus processing included automatic structural and linguistic annotations. For corpus querying NoSketch Engine was used (Kilgarriff et al. 2004). The operationalization of text quality features is based on both the relevant research literature and analyses of learner productions. Manual annotation indicators are:

- Introduction (a. yes/no, b. content yes/no/partly)
- Conclusion (a. yes/no, b. content yes/no/partly)
- Procedures of establishing the writer's point of view (a. means of realisation, b. correctness)

"Introductions" and "conclusions" refer to structural text aspects. They often represent a challenge for novice writers (e.g., Augst et al. 2007; Petersen 2013). In addition, the concept of "text routines" [Textroutinen] (Feilke 2012a, b, 2018; Feilke & Lehnen 2012) was explored (see Knopp et al. 2014 for a similar approach): We focused on procedures for establishing a point of view [Positionierungsprozeduren] such as "meiner Meinung nach" ("In my opinion") " (Gätje et al. 2012; Steinhoff 2007). In addition, statistical measures such as text length (Grabowski et al. 2014) and the number of connectors per sentence were included (Hancke 2013, Weiss 2017). Manual coding quality is reported by means of interrater agreement.

For qualitative data analysis descriptions of emerging features (RQ1, RQ2) and comparisons (RQ2) are presented. Quantitative analyses include logistic regressions (RQ1) (Knopp et al. 2014) and tests of significance (RQ2).

First data analyses suggest considerable differences between texts at different proficiency levels. Despite its limited sample size and its exploratory character, implications of the study might thus be of relevance for both the field of language assessment and language teaching. For the former, the study might serve to explore selected task-related and theoretically valid potential additional rating criteria. For the latter, the study might contribute to a better understanding of a potential link of L2 writing development to the often-criticized, yet ubiquitous CEFR levels.

References

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

Abel, A., Vettori, Ch., & Wisniewski, K. (Ed.) (2012). Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. *Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Volume 1. Bozen-Bolzano: Eurac.

Augst, G., Disselhoff, K., Henrich, A., Pohl, T., & Völzing, P.-L. (2007). *Text – Sorten – Kompetenz: eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter*. Frankfurt a.M. u. a.: P. Lang.

Bachmann, T., & Steinhoff, T. (2010). Schreibaufgaben situieren und profilieren. In T. Pohl, H. Günther, M. Becker-Mrotzek, U. Bredel, & T. Steinhoff (Ed.), *Textformen als Lernformen* (pp. 191–210). Duisburg: Gilles & Francke.

Feilke, H. (2012a). *Schreib- und Textroutinen. Theorie, Erwerb und didaktisch-mediale Modellierung*. Frankfurt a.M.: P. Lang.

Feilke, H. (2012b). Was sind Textroutinen? Zur Theorie und Methodik des Forschungsfeldes. In In H. Feilke (ed.), *Schreib- und Textroutinen. Theorie, Erwerb und didaktische Modellierung* (pp. 1–32). Frankfurt a.M.: P. Lang.

Feilke, H. (2018). Argumente für eine Didaktik der Textprozeduren. In T. Bachmann, & H. Feilke (Ed.), *Werkzeuge des Schreibens. Beiträge zu einer Didaktik der Textprozeduren* (pp. 11–34). Stuttgart: Klett.

Feilke, H., & Lehnen, K. (2012). *Schreib- und Textroutinen: Theorie, Erwerb und didaktisch-mediale Modellierung*. Frankfurt a.M.: P. Lang.

Gätje, O., Rezat, S., & Steinhoff, T. (2012). Positionierung. Zur Entwicklung des Gebrauchs modalisierender Prozeduren in argumentativen Texten von Schülern und Studenten. In H. Feilke (Ed.), *Schreib- und Textroutinen. Theorie, Erwerb und didaktische Modellierung* (pp. 125–154). Frankfurt a.M.: P. Lang.

Grabowski, J., Becker-Mrotzek, M., Knopp, M., Jost, J., & Weinzierl, C. (2014): Comparing and combining different approaches to the assessment of text quality. In D. Knorr, C. Heine, & J. Engberg (Ed.), *Methods in writing process research* (pp. 147–165). Frankfurt a.M.: P. Lang.

Hancke, J. (2013). *Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language*. Tübingen: Universität Tübingen. http://merlin-platform.eu/docs/MAThesis-Julia-Hancke.pdf

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams, & S. Vessier (Hrsg.), *Proceedings of the 11th EURALEX International Congress*. (pp. 105–116). Lorient: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Knopp, M., Jost, J., Linnemann, M., & Becker-Mrotzek, M. (2014). Textprozeduren als Indikatoren von Schreibkompetenz – ein empirischer Zugriff. In T. Bachmann, & H. Feilke (Ed.), *Werkzeuge des Schreibens. Beiträge zu einer Didaktik der Textprozeduren.* (pp. 111–128). Stuttgart: Fillibach.

Petersen, I. (2013). *Schreibfähigkeit und Mehrsprachigkeit*. Berlin, Boston: De Gruyter Mouton.

Steinhoff, T. (2007). *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*. Berlin u.a.: de Gruyter.

Weiss, Z. (2017): *Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects*. Tübingen**:** Eberhard Karls Universität Tübingen. http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/weiss2017-distr.pdf

Introducing the *Corpus of English as a Foreign Language* (COREFL): A bimodal, multi-task corpus for SLA research

Ana Díaz-Negrillo[1], Cristóbal Lozano[1] & Marcus Callies[2]
University of Granada (Spain)[1], University of Bremen (Germany)[2]
anadiaznegrillo@ugr.es, cristoballozano@ugr.es, callies@uni-bremen.de

Learner corpora are currently widely used for a variety of purposes in foreign language research including second language acquisition (SLA), language teaching, language testing, and computational linguistics (see Granger et al. 2015 for an overview). In this presentation, we will introduce the COREFL (*Corpus of English as a Foreign Language*) as a powerful new resource in SLA research. Its complex and comparable design follows the design principles of the CEDEL2 (*Corpus Escrito del Español como L2:* http://cedel2.learnercorpora.com) (Lozano & Mendikoetxea 2013), which opens up a wide range of comparative perspectives for SLA research within Contrastive Interlanguage Analysis (CIA; Granger 2015), and even beyond:

- *Two L1 backgrounds*: The corpus contains learner data from L1 Spanish and L1 German learners of English, which allows to explore cross-linguistic and typological factors;
- *Bimodal*: The corpus contains a sample of written and spoken data produced by the same learners in the same task, which allows to examine the effects of mode and processing constraints on the production of L2 narratives (see Ädel 2008 et al. 2015);
- *Bidirectional:* the COREFL (L1 Spanish/German – L2 English) can be compared with its equally-designed CEDEL2 counterpart (L1 English/German – L2 Spanish);
- *Multiple tasks:* It contains four different narrative tasks which allows to test the effect of task on L2 production (cf. Tracy-Ventura & Myles 2015);
- *Two control corpora:* While the standard in LCR is to use only one control corpus, i.e. data produced by native speakers of the learners' target language (English in the present case), COREFL also contains control corpora of the learners' L1, i.e. Spanish (with Peninsular and Latin American varieties) and German (currently being compiled). The use of these control corpora will provide insights into the likely sources of knowledge of L2 learners with respect to L1 transfer (L1 control corpora) and input (L2 control corpus);
- *Various proficiency levels:* It contains data from a range of EFL learner populations of a variety of proficiency levels and learning contexts. The COREFL covers the proficiency levels A1-C2, which allows for research into L2 development, as well as a variety of ages (from 12 years onwards) and educational levels (secondary education and university), which allows for research into the effects of age and acquisitional setting in L2 acquisition;
- *Learner variables*: It contains a wide range of SLA-relevant learner variables, which allows to test essential questions in L2 research (effects of age of onset, length of exposure to L2, language use patterns, etc.).
- *Different settings:* The L1 Spanish - L2 English subcorpus contains data from learners in different types of instructional settings: secondary school bilingual programmes (CLIL, *Content-and-Language-Integrated-Learning*) vs. mainstream EFL classrooms, i.e. university EMI (English as a Medium of Instruction) learners vs. university SMI (Spanish as a Medium of Instruction) learners. This will provide insights into the effect of these bilingual education/immersion programmes in L2 acquisition.

Even though the COREFL is still in a pre-release stage, around 1612 texts of written and spoken data have been collected so far and its compilers will continue to collect data over the next two years (cf. Table 1). This presentation aims to discuss the design features of the corpus, its development and its potential for SLA research with a view to gathering feedback before the corpus is released.

Table 1: Current holdings of the COREFL (as of 29/01/2019)

Subcorpus	**No. of texts (approx.)**
L1 English native texts	105
of which oral	23
L1 Spanish native texts	248 (+ 796 from CEDEL2 version 1.0)
of which oral	54
L1 German native texts	under compilation
L1 Spanish-L2 English (university)	624
of which oral	112
L1 Spanish-L2 English (secondary)	485
L1 German-L2 English (university)	150
TOTAL	*1612 (+ 796 from CEDEL2 v. 1.0)*

References

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking Up Contrastive and Learner Corpus Research* (pp. 35–53). Amsterdam: Rodopi.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

Granger, S., Gilquin, G., & Meunier, F. (Eds.) (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65–100). Amsterdam: John Benjamins.

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), 58–95.

Accuracy in spoken learner English at B2 and C1 levels (and future inclusion of A2 and B1 levels)

Tomáš Gráf, Lan-fen Huang
Charles University, Shih Chien University
tomas.graf@ff.cuni.cz, lanfen.huang@gmail.com

One of the many aspects of spoken learner language mentioned in the CEFR (Council of Europe, 2001; 2018) is accuracy. It characterizes B2 speakers as having "a relatively high degree of grammatical control" and no longer producing errors which cause misunderstandings. C1 speakers are characterized by "a high degree of grammatical accuracy" with rare, difficult to spot, and generally corrected errors. As this cursory description deals with only grammatical accuracy, the present study explores the accuracy of B2 and C1 spoken learner English more broadly, including grammatical, lexical and lexico-grammatical errors, and aiming to identify which types of errors characterize each level, and which errors either disappear or persist at level C1. Also considered is the effect of task design, and of the two typologically different L1s of the learners. This research is unique in that it exploits corpora which have been assessed for proficiency and thus provides an empirical basis for the understanding of CEFR B2 and C1 accuracy, and identifies particular areas of difficulty for speakers at these levels.

The data derives from the Czech (Gráf, 2015a) and Taiwanese (Huang, 2014) components of LINDSEI[9]. These two subcorpora have been rated for accuracy by professional IELTS examiners who also received a special CEFR rater-standardisation training (Huang et al., 2018). The size of the dataset and the distribution of levels is shown in Table 1. B1 and C2 speakers were excluded from the analysis. The transcriptions of the remaining 89 speakers' 15-minute interviews were error-tagged using the Louvain error-tagging manual (Dagneaux et al., 2008) extended by Gráf (2015) to include 59 error types at grammatical, lexical, lexico-grammatical, and syntactical levels and taking into account characteristic features of speech grammar and of the genres determined by the tasks.

	B1 (n)	**tokens**	**B2 (n)**	**Tokens**	**C1 (n)**	**tokens**	**C2 (n)**	**tokens**	**Total tokens**
Czech	0	0	13	24,165	35	66,305	2	5,499	**95,969**
Taiwanese	9	10,028	39	55,707	2	3,785	0	0	**69,520**
Total	**9**	**10,028**	**52**	**79,872**	**37**	**70,090**	**2**	**5,499**	**165,489**

Table 1. *Size of the dataset: numbers of speakers and of tokens the speakers at different proficiency levels produced.*

A total of 5,108 errors was identified. The comparison of error rates (errors per 100 words, henceforth phw) between the two levels of proficiency showed that B2 speakers produce errors at a higher frequency (6.7 errors phw at B2 and 1.9 errors phw at C1), and this was similar for the performance in the three different tasks (monologue, dialogue and picture description). Grammatical errors are the most frequent, followed by errors of lexical nature. Other types are much less frequent (see Table 2).

	Morphological errors	**Grammar errors**	**Lexical errors**	**Lexico-grammatical errors**	**Word order errors**	**Infelicities**
B2	7 (0.2%)	3,023 (72.5%)	701 (16.8%)	152 (3.7%)	196 (4.7%)	76 (1.8%)
C1	3 (0.3%)	558 (59.4%)	277 (29.5%)	55 (5.9%)	34 (3.6%)	13 (1.4%)

Table 2. *Representation of error types at B2 and C1 levels.*

[9] Louvain International Database of Spoken English Interlanguage (Gilquin et al., 2010)

As for grammar, article and verb tense errors are the most frequent for both groups. At C1, these are less frequent and also some types of errors common at B2 do not appear here. Persistent errors typically involve the use of articles, of the present perfect, and in tense agreements. Differences also occur between Czech and Taiwanese speakers. The latter produce errors at a higher frequency and are more prone to commit errors in areas which the English Vocabulary Profile classifies as B1. Lexical errors involve mostly those affecting single lexemes (esp. prepositions).

Qualitative analysis of a selection of these errors reveals that none of these errors impact intelligibility. While the CEFR claims that at C1 errors are generally corrected, this is not the case here: most of the errors still seem to be errors of competence rather than performance and the speakers do not appear to be aware of them and are thus suitable targets for pedagogical intervention.

References

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with New Descriptors*. Strasbourg Cedex: Council of Europe.

Dagneaux, E., Denness, S., Granger, S., Meumier, F., Neff, J., & Thewissen, J. (2008). *The Louvain Error Tagging Manual Version 1.3*. Centre for English Corpus Linguistics, Université catholique de Louvain. Louvain-la-Neuve.

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

Gráf, T. (2015a). *Accuracy and fluency in the speech of the advanced learner of English.* (PhD thesis), Charles University, Prague. Retrieved from https://is.cuni.cz/webapps/zzp/detail/151663

Gráf, T. (2015b). *Korpus LINDSEI_CZ*. Prague: Charles University.

Huang, L.-f. (2014). Constructing the Taiwanese component of the Louvain International Database of Spoken English Interlanguage (LINDSEI). *Taiwan Journal of TESOL*, 11(1), 31–74.

Huang, L.-f., Kubelec, S., Keng, N., & Hsu, L.-h. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, 8(14), 1–17.

Adverbial –*ing* clauses in L2 learner English

Rita Juknevičienė
Vilnius University
rita.jukneviciene@flf.vu.lt

Owing to their ambiguous grammar and semantics, the English –*ing* words pose many challenges to learners of English as a foreign language (EFL). On the one hand, –*ing* forms could be viewed as borderline cases in terms of word class reference and function as nouns, verbs or adjectives (Biber et al., 1999; De Smet, 2010). As a result, teaching and learning such forms involves a considerable difficulty to EFL learners, especially when non-congruence of grammatical categories in L1 and L2 is an important factor. On the other hand, –*ing* clauses are characteristic of written academic English and thus a targeted feature of writing classes. For example, adverbial –*ing* clauses allow for more condensed ways of expression and, among other aspects, are viewed as evidence of a higher level of linguistic competence in L2 written English (Hawkins & Filipovic, 2012). Yet semantic ambiguity of certain types of –*ing* clauses makes their use problematic for non-native users. Earlier studies of adverbial –*ing* clauses in L2 English show a varied picture of over- and underuse which to some extent might be accounted for by mother tongue influence (cf. Cosme, 2008; Granger, 1997; Grigaliūnienė & Juknevičienė, 2012; Springer, 2012).

This study was set up to identify factors which significantly influence the occurrence of –*ing* clauses in L2 learner writing in cases where there is an alternative finite clause option. More specifically, it focuses on one type of adverbial clauses, namely, time clauses, which also have a finite alternative, and aims to identify factors that affect L2 learners' choice of the finite or non-finite type of adverbial time clause.

The data for this study was extracted from the ICLE corpus (Granger et al., 2009), and it represents EFL learners of ten mother tongue backgrounds. The corpus search tool #LancsBox (Brezina et al., 2015) was used to generate concordances of subordinators of time clauses, i.e. *after*, *before*, *when* and *while*, from which a random set of finite and non-finite time clauses was selected. Each instance was manually coded for a number of contextual variables, for instance, position of the clause in relation to the matrix clause, clause length in words, semantics of the predicate, L1, etc. The test of binary logistic regression was run on the data to identify which variables have significant effects on the use of finite vs non-finite clauses. The statistical tests were computed using R (R Development Core Team 2008). It is hypothesized that the importance of individual contextual variables will vary for speakers of different L1 languages.

References

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

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139–173.

Cosme, Ch. (2008). Participle clauses in learner English: the role of transfer. In G. Gilquin, Sz. Papp & M. B. Díez-Bedmar (Eds). Linking up contrastive and learner corpus research (pp. 177–198). Amsterdam: Rodopi.

De Smet, H. (2010). English –ing-clauses and their problems: the structure of grammatical categories. Linguistics, 48(6), 1153–1193.

Granger, S. (1997). On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared. In J. Aarts, I. de Monnick & H. Wekker (Eds) Studies in English Language and Teaching – In honour of Flor Aarts (pp. 185–198). Amsterdam: Rodopi.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds). (2009). International Corpus of Learner English (ICLE). V. 2. Louvain-la-Neuve: Presses Universitaires de Louvain.

Grigaliūnienė, J. & Juknevičienė, R. (2012). Corpus-based learner language research: contrasting speech and writing. Darbai ir dienos, 58, 137–152.

Hawkins, J. A. & Filipovic. L. (2012). Criterial features in L2 English: Specifying the reference levels of the Common European Framework. English Profile (vol. 1). Cambridge: Cambridge University Press.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Springer, P. E. (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. Amsterdam: Vrije Universiteit.

A corpus-based text-analytic tool for novice writers of Academic Russian

Mikhail Kopotev[1], Olesya Kisselev[2], Mariia Fedorova[3], Alexandr Klimov[3],
Anna Dmitrieva[3], Anastasiia Baranchikova[3]
[1] University of Helsinki, [2] University of Texas at San Antonio, [3]
Higher School of Economics in Moscow
mihail.kopotev@helsinki.fi, olesya.kisselev@utsa.edu

The study of English academic discourse has benefited greatly from the application of corpus-based tools and analysis devoted to it in the past few decades (Ackerman & Chen, 2013; Biber et al. 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013). Similar studies of academic registers of languages other than English have been lagging behind. The project, titled CAT&kittens, described in this paper intends to address this gap, as well as to contribute to a general exploration of (semi)automated tools available to the learning of academic genres. The service is intended to help a user with two tasks: to highlight fragments, which differs from a reference corpus, and to offer, when possible, a substitute that better serves in a given context.

The central part of the project involves the development of the comprehensive representative Russian Corpus of Academic Texts (CAT). Following well-established corpus development procedures (e.g., BAWE). Texts in the CAT corpus are sourced from six general disciplinary fields: social studies, political science and international relations, law, linguistics, economics, psychology and education science. The discipline sub-corpora consist of about 370 to 480 thousands tokens, amounting to approximately 2 mln. tokens in the corpus in general. Texts entered in CAT are supplied with metalinguistic, morphological and syntactic annotations, carried out with the help of the Universal Dependencies pipeline (Straka et al. 2017).

CAT is outfitted with built-in data processing tools, which allows for evaluation of texts written by novice writers of Academic Russian, both native and non-native:

- *General statistics* of an analyzed novice text: a readability test, average length of words, sentences, and paragraphs, and TTR.
- *Lexical analysis* highlights terminology that are unattested in the discipline domain, and suggests alternatives.
- *Collocational analysis*. Based on n-gram frequencies, all non-attested word choice selections in the novice texts will be identified; attested collocational alternatives extracted from CAT will be provided.
- *Grammar check*. Unlike available spell-checkers, the tool is focused on detecting deviations that feature in academic writing, e.g. genitive chains, mixtures of synthetic and analytical comparatives etc.

Tools like the one described above are routinely evaluated in terms of recall and precision, when both measures are taken equally important. We believe, however, that for many CALL tools, precision is more instrumental. While a complete automatic correction of every error is an impossible task, focusing on a precise improvement based on a shortlist is likely to be realistic (if challenging) and pedagogically useful.

References

Ackerman, K., & Chen, Y-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12(4), 235–247.

BAWE (The British Academic Written English), available at:
http://www2.warwick.ac.uk/fac/soc/al/research/collections/bawe/how_to_cite_bawe. Last retrieved April, 15, 2019.

Biber, D., Conrad., & Cortes, V. (2004). If you look at …: Lexical bundles in University teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1), 58–72.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109–136.

Straka, M., and Straková J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Lexical diversity and lexical transfer in a longitudinal English learner corpus

Eliane Lorenz, Sharareh Rahbari, Peter Siemund
University of Hamburg
eliane.lorenz@uni-hamburg.de, sharareh.rahbari@uni-hamburg.de, peter.siemund@uni-hamburg.de

Lexical diversity, or lexical richness, in a language is associated with proficiency. In foreign language production, we expect to find more infrequent words and more overall lexical variety in language use with increasing competence or proficiency (Milton, 2009). In addition, foreign language learners frequently use words from their previously acquired language(s) in their foreign language production (lexical transfer), especially if these languages are related (Ringbom, 2001).

In this light, we investigate how lexical diversity and lexical transfer develop in monolingual and bilingual learners of English: (i) can we see a developmental progress within a time span of two years of studying English, (ii) do bilingual leaners show more or less lexical transfer than their monolingual peers?, and (iii) do the bilingual learners use lexical borrowings from both their background languages or just from one?

This study uses a pilot version of a longitudinal English learner corpus. The corpus consists of written data from a longitudinal project that investigates the multilingual development of children living in Germany (Multilingual development: a longitudinal perspective (MEZ)). Two cohorts, students in school year 7 and 9, of three different language groups (monolingual German, bilingual Russian-German, and bilingual Turkish-German) participated in four measuring points between Spring 2016 and Summer 2018. For the present study, we analyse the English written performance at these four measuring points (the same picture description task for all students), yielding four English texts per student. In total, the pilot version of the MEZ CORPUS consists of approximately 63,000 word tokens.

First results reveal that, on average, the younger cohort of the German monolinguals produced slightly more words per text than their bilingual peers. The older German monolingual cohort, however, produced fewer words per text than the bilinguals across all four measuring points. This may indicate a difference in L2 and L3 learners. Regarding the first two measuring points, both cohorts of the German monolinguals have higher type-token ratios than the bilingual participants. In the latter two measuring points, this result is less clear and exhibits a varied picture. This suggests that lexical diversity does not develop in a linear manner and that it is affected by additional variables, for example the topic of the writing task. In addition, lexical transfer seems to come exclusively from German and not from Russian or Turkish, which could be explained by the typological similarity between English and German, and because German is the academic language of all participants. A more detailed analysis reveals distributional differences between the bilinguals and monolinguals, in that we find considerably fewer lexical borrowings in the bilingual data than in the monolingual learner data. In addition, we observe a developmental difference across less proficient and more proficient learners. In the second part of the analysis, we use the comprehensive meta-data (i.e. socio-economic status, type of secondary school, language use at home) to interpret the results.

References

MEZ – Mehrsprachigkeitsentwicklung im Zeitverlauf. (2014–2019). Projektkoordination: Prof. Dr. Dr. h. c. Ingrid Gogolin, Universität Hamburg; © MEZ 2017.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Ringbom, H. (2001). Lexical Transfer in L3 Production. In J. Cenoz, B. Hufeisen, U. Jessner (Eds.), *Cross-linguistic Influence in Third Language Acquisition* (pp. 59–68). Clevedon: Multilingual Matters.

Multi-L1 learner corpus design for SLA research purposes: CEDEL2 (Corpus Escrito del Español L2, version 2.0)

Cristóbal Lozano, Nobuo Ignacio López-Sako
Universidad de Granada, Universidad de Granada
cristoballozano@ugr.es, nilsako@ugr.es

The last decade has witnessed an upsurge of new learner corpora, mainly L2 English corpora (Granger et al. 2002, 2009, 2015). Recent interest in L2 Spanish research has brought about the creation of large corpora like CEDEL2 (Lozano 2009, Lozano & Mendikoetxea 2013), CAES (Rojo & Palacios 2015) and SPLLOC (Mitchell et al. 2008).

We will discuss the importance of SLA-informed learner corpus design by focusing on CEDEL2 (http://cedel2.learnercorpora.com). It follows the design criteria proposed for normative corpora (Sinclair 2005) (content selection, representativeness, contrast, structural criteria, annotation, sample size, documentation, balance, topic, homogeneity). These were adapted to collect SLA-relevant variables (Lozano & Mendikoetxea 2013, Granger 2008): learner and task variables (e.g., L1, age of onset, length of exposure, language use patterns, self-reported proficiency, placement test score, task conditions and timing). Such design can inform SLA researchers about key factors (effects of: proficiency, L1, age, input, task, modality, learning environment, etc).

CEDEL2 (v. 1.0) contains 750,000 words written by 2,578 participants. Data come mostly from native speakers of English who are learning Spanish (L1 English-L2 Spanish) in a variety of countries and learning environments (USA, UK, Canada, Spain, etc.) at all levels of language proficiency, plus an equally-designed control subcorpus of Spanish natives (Peninsular and Latin American varieties). CEDEL2 (v. 2.0) is progressively being expanded to incorporate a wider variety of L1s (see list below) and, following Myles' (2015) recommendations, oral data. Crucially, the small oral data come from the same task(s) and participant(s).

The inclusion of Japanese, a topic-drop language, is justified because learner corpora containing Japanese native data come from L2 English corpora (the written ICLE v. 2.0, NICT JLE, SILS corpus, ICNALE, and the spoken LINDSEI). The L1 Japanese-L2 Spanish combination is rare in both L2 acquisition research and learner corpora (cf. Campillos Llanos 2014 for an exception, albeit very limited in size, comprising just 4 participants). Ours represents a potential contribution to the understanding of L1 Japanese effects on L2 Spanish acquisition (see Tono 2005 for L1 Japanese effects on L2 English in the NICT JLE corpus).

We will discuss the design criteria which are common to all subcorpora in CEDEL2 v. 2.0. This will allow SLA researchers to make cross-linguistic contrasts under the framework of Contrastive Interlanguage Analysis (Granger 2015) and beyond: (i) learner vs. native varieties; (ii) same-proficiency-level learners of different L1s; (iii) different-proficiency-level learners of the same L1; (iv) learners of typologically related languages (Germanic: English vs German vs Dutch; Romance: Portuguese vs Italian) and (v) learners of typologically related vs. distant vs. unrelated languages (Germanic language(s) vs. other Indoeuropean languages vs Japanese/Chinese); (vi) spoken vs written effects while keeping the participant constant; (vii) L1 effects vs L2 input effects on certain learner subcorpora since, following (Hendriks 2003) controlling for the learners' native language (L1 native Japanese subcorpus in this presentation), as well as for the target language (L1 native Spanish), allows to check likely L1-transfer vs. L2-input effects on a given corpus (L1 Japanese-L2 Spanish).

Finally, we will show the project main objectives regarding online data collection and will make a call for participation and collaboration from universities and other institutions that might be interested in enriching our international CEDEL2 corpus by providing data from the following or other L1 scenarios (see data-collection online forms in http://learnercorpora.com):

- Native control corpora: Spanish, English, Japanese, Portuguese
- Learner corpora with different L1s: Japanese, Chinese, English, German, Dutch, Italian, Portuguese, Greek, Russian.

References

Campillos Llanos, L. (2014). A Spanish oral learner corpus for computer-aided error analysis. *Corpora*, 9(2), 207–238. Corpus available at: http://cartago.lllf.uam.es/corele/index.html

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1), 7–24.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-neuve: Presses universitaires de Louvain.

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

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

Hendriks, H. (2003). Using nouns for reference maintenance: A seeming contradiction in L2 discourse. In A. Giacalone Ramat (ed.), *Typology and Second Language Acquisition* (pp. 291–326). Berlin: Mouton de Gruyter.

ICLE v. 2.0 (International Corpus of Learner English): https://uclouvain.be/en/research-institutes/ilc/cecl/iclev2.html

ICNALE corpus (The International Corpus Network of Asian Learners of English): http://language.sakura.ne.jp/icnale/

LINDSEI corpus (Louvain International Database of Spoken English Interlanguage): https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español como L2. In C. M. Bretones, & et al (Eds.), *Applied Linguistics Now: Understanding Language and Mind* (pp. 197–212). Almería: Universidad de Almería. Online versión 1.0: http://cedel2.learnercorpora.com

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 65–100). Amsterdam: John Benjamins. Online version1.0: http://cedel2.learnercorpora.com

Mitchell, R., Domínguez, L., Arche, M., Myles, F., & Marsden, E. (2008). SPLLOC: A new corpus for Spanish second language acquisition research. In L. Roberts, F. Myles, & A. David (Eds.), *EUROSLA Yearbook 8* (pp. 287–304). Amsterdam: John Benjamins. Online 1 and 2 versions: http://www.splloc.soton.ac.uk/

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Guilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research* (pp. 309–311). Cambridge: Cambridge University Press.

NICT Japanese Learner English (JLE) Corpus: https://alaginrc.nict.go.jp/nict_jle/index_E.html

Rojo, G., & Palacios, I. (2015). CAES (Corpus de Aprendices de Español) (v. 1.0). Available in http://galvan.usc.es/caes

SILS Learner Corpus of English: http://www.f.waseda.jp/vicky/learner/

Sinclair, J. (2005). How to build a corpus. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 79–83). Oxford: Oxbow books.

Tono, Y. (2005). Corpus-based SLA Research: State of the art of Learner Corpus Studies. In M. Minami, H. Kobayashi, M. Nakayama, & H. Sirai (Eds.). *Studies in Language Sciences* (4): Papers from the Fourth Annual Conference of the Japanese Society for Language Sciences (pp. 45–77). Tokio: Kurosio Publishers.

Comparing L1 and advanced learner English academic writing: the case of *–ly* adverbs

Tomáš Novotný, Markéta Malá
Charles University
tomizilek@gmail.com, Marketa.Mala@ff.cuni.cz

The paper compares English texts written by two types of novice academic writers – L1 university students and advanced Czech learners. We focus on the use of *–ly* adverbs, the most frequent form of adverbs in academic prose (Biber et al., 1999).

We aim to answer two research questions: first, we are interested in the extent and the ways in which linguistic patterns of adverb use differ in English academic texts written by L1 students and Czech advanced learners. The second question relates to the material used. The analysis is based on VESPA-CZ, a member of the VESPA family of advanced learners' corpora. Following a similar investigation by Hasselgård (2015), we try to assess the comparability of the VESPA-CZ to other VESPA corpora and the possibilities of exploring the characteristics of academic English written by advanced learners with different first languages.

The analysis relies on two corpora of academic student writing – VESPA-CZ and BAWE. BAWE comprises L1 university students' assignments (Alsop & Nesi, 2009); VESPA-CZ essays written by Czech advanced learners of English. Thematically close texts, English literature essays, were selected from both corpora (hence BAWE-EL). An additional corpus compiled from papers published in English literary academic journals (AP) serves as a yardstick against which the students' essays are compared. The two L1 corpora are of approximately the same size (235 000 tokens); VESPA-CZ is half the size yet (106 600 tokens).

In this pilot study, we examine the frequency of *–ly* adverbs in the three corpora, their syntactic and semantic functions, and the specific lexical choices made by the writers. The analysis revealed 3 types of difference among the corpora:

1. The findings are arranged along a scale which goes from VESPA-CZ to the AP corpus. This type of difference can be illustrated by modifiers of adverbs and adjectives, with the relative frequency in AP almost twice that in VESPA-CZ, and the BAWE-EL in between (closer to AP). Due to the high number of modifiers embedded in phrases, journal articles in AP display the phrasal complexity typical of structurally 'compressed' academic writing (Biber & Gray, 2010). At the same time, the repertoire of modifiers in VESPA-CZ is more limited than in BAWE-EL, including degree adverbs which are not peculiar to academic discourse (*completely*, *absolutely*) (Granger, 1998; Biber, 2006). The same applies to particularizers (*mainly*, *mostly*, *particularly*).
2. Both novice academic writers' corpora display similar features, which differ from the academic journals' corpus. The relative number of *–ly* adverbs is similar in VESPA-CZ and BAWE-EL (1341 and 1340 per 100,000 words, respectively), while AP displays a higher relative frequency of *–ly* adverbs (1584).

A closer look at the results of the semantic analysis shows that two types of disjuncts appear to characterize novice academic writing: modal disjuncts (*possibly*, *probably*, *obviously*), perhaps signalling uncertainty, often in combination with modal verbs (Aijmer, 2002); and the stance disjunct *interestingly*. These adverbs are underrepresented in AP.

3. Czech novice academic writers differ from L1 writers, both novice and experienced, whose writing displays similar features. Czech learners overuse conjuncts, compared to both L1 corpora. They, however, tend to stick to several conjuncts, which they use frequently (*finally*, *firstly*, *consequently*).

The preliminary results of the analysis largely correspond to the findings for Norwegian learners based on VESPA-NO (Hasselgård, 2015). This suggests that VESPA-CZ will make a useful contribution to the VESPA family of learner corpora, enhancing the possibility to compare the features of academic writing by advanced learners of English with a variety of L1 backgrounds.

References

Aijmer, K. (2002). Modality in advanced learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, (pp. 55–76). Amsterdam/Philadelphia: John Benjamins.

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

Biber, D. (2006). *University Language*. Amsterdam/Philadelphia: John Benjamins.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9, 2–20.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

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

Hasselgård, H. (2015). Lexicogrammatical features of adverbs in advanced learner English. *International Journal of Applied Linguistics* 166(1), 163–189.

Corpora

BAWE – The British Academic Written English Corpus. https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

VESPA-CZ – Varieties of English for Specific Purposes dAtabase (Czech learners – The Czech Science Foundation Project No. 19-05180S). https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html

AP – articles from English Literary Renaissance (https://onlinelibrary.wiley.com/journal/14756757), Renaissance Studies (https://onlinelibrary.wiley.com/journal/14774658), Shakespeare Quarterly (https://academic.oup.com/sq)

Accounting for the effects of learner engagement in a corpus of computer-mediated communication

Tim Marchand
Gakushuin University
tim.marchand@gakushuin.ac.jp

Engaging L2 learners to become active, autonomous participants in the learning process has become a common aim of many language classrooms, and numerous educational institutions have turned to computer-mediated communication (CMC) and Web 2.0 platforms to encourage and facilitate this step (Peeters 2018). Recent research in computer assisted language learning (CALL) has provided case study evidence to support this pedagogical development, including writing gains in terms of fluency, lexical richness and syntactic complexity being experienced by learners who actively interact on classroom blogs (Dizon & Thanyawatpoki 2018). Some have argued that these results are in line with Astin's (1993) engagement theory, which holds that "development is proportional to quality and quantity of involvement" as one of its central tenets (Akbari et al. 2016).

This paper is a work-in-progress report of an investigation into the relationship between learner involvement and writing development in one particular CALL setting: an English language course for first year university students in Japan. In this course, lesson materials and tasks were provided online through a dedicated news-based blog. Each week, students wrote their reactions to news stories on the blog, and these comments then form the basis of a learner corpus. Over the course of one academic year, learners submit 20-40 written reactions to the news articles, producing texts that demonstrate an engagement with the various news topics, and an interaction with each other in the form of CMC.

The paper addresses the following research questions:

1) To what extent does learner engagement with topic affect written output in terms of fluency and complexity?
2) What are the effects on written output of task variables (with lesson support or not), and learner variables (profile data, interactive behaviour)?
3) To what extent does L2 CMC writing develop over time?

To answer the research questions, the paper first addresses the operationalisation of learner engagement with topic. This was done by reference to a questionnaire in which learners were asked to rate their level of interest and knowledge of each news topic; and by identifying certain common traits of learner behaviour as they applied themselves in the CALL tasks. Correlating these measures as predictors of the CMC produced, the paper explores whether increased engagement with lesson materials had any discernible effects on the nature of the written responses.

The paper concludes by suggesting that the early results support the dynamic systems framework to understanding language development (Larsen-Freeman 2006; Verspoor et al. 2012) in that several features of language do not develop in a linear fashion.

References

Akbari, E., Naderi, A., Simons, R.-J., & Pilot, A. (2016). Student engagement and foreign language learning through online social networks. *Asian-Pacific Journal of Second and Foreign Language Education* 1(4), 1–22.

Astin, AW (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

Dizon, G., & Thanyawatpokin, B. (2018). Web 2.0 tools in the EFL classroom: Comparing the effects of Facebook and blogs on L2 writing and interaction. *The EuroCALL Review* 26(1), 29-42.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics* 27.4, 590–619.

Peeters, W. (2018). Applying the networking power of Web 2.0 to the foreign language classroom: a taxonomy of the online peer interaction process, *Computer Assisted Language Learning* 31:8, 905-931.

Verspoor, M., M. S. Schmid & X. Xu (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing* 21.3, 239–263.

Non-native multi-word expressions in a corpus of spoken English: A study of errors in content and function words for FL pedagogy

Alisa Mitchel Masiejczyk
Warsaw University
a.masiejczyk@uw.edu.pl

Native language corpora inform foreign language pedagogy by verifying descriptions and perceptions once 'informed' mainly by intuition about language; through valuable data on frequency, they offer teachers and learners chances to discover co-text – including phrase level phenomena that might be easily missed in regular classroom input. This has clear advantages in the context of materials development, such as those used in teacher and translator training programs. However, without looking into learner corpora as well, we stand to miss important indicators of problem areas. Knowing the likely distribution of foreign language (target) items cannot, in itself, provide predictive information as to what difficulties foreign language learners may encounter in learning and later retrieving that item (cf. Granger, et al. 2002, Tsui in Sinclair, 1996).

This project explores non-native speakers' errors based on contrastive interlanguage analysis using a spoken language corpus of 26 adult native speakers of English (American and British varieties) and 20 Polish non-native adult speakers of English (students of English philology at the University of Warsaw, Poland). The speakers in both groups performed the same story-telling task based on a picture board; the content words chosen in the NS and NNS performances overlap to a degree though the NNS corpus contains numerous examples of constructions which may point to transfer, or 'creative' use of language. Problem areas are often signalled through the use of meta expressions, though many multi-word expressions (sequences including compounds, phrasal verbs, idioms, fixed phrases, prefabricated routines, cf. Moon, 1997) seem to have been arrived at through analytic processing (cf. the dual mode hypothesis in Sinclair, 1991 and Wray, 2002). These effortful, rules-and-lexis-based constructions often rely upon meaning-bearing content words (cf. Gozdawa Gołębiowski 2008).

In the analysis of the spoken language corpus, the present study tags content- and function-word-related errors according to the researcher's taxonomy (cf. Mitchel Masiejczyk, 2011) to examine the proportion of phrase-level errors in terms of components – content words (by instances of incorrect selections, added words, or omitted words) or function words (also by instances of incorrect selections, added words, or those omitted). In the data set, function words emerged as significantly more vulnerable to idiosyncratic errors than content words, in terms of underuse and overuse (where items have been added in by the user, in multiword expressions).

The presentation of content items (primary in the FL classroom) without a clear focus upon their likely accompaniment (function words), may lead learners to construct phrase-level utterances via a series of ad hoc decisions about grammar and lexis, which takes place at all levels of competence. In the context of foreign language teaching and advanced translation workshops for adults with knowledge of multiple languages, highlighting the essential role of 'small' system markers (function words) should go in hand in hand with the teaching of new vocabulary and rules of usage. It is argued that this simplified approach to viewing error frequencies allows FL learners and their teachers to go beyond word-by-word treatments of texts and notice the properties of the co-text which strings islands of meaning – very often content words – together.

References

Gozdawa Gołębiowski, R. (2008). Grammar and Formulaicity in Foreign Language Teaching. In *Glottodidactica* XXXIV (2008), 75-86.

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

Mitchel Masiejczyk, A. (2011). Formulaically speaking: error and the adult foreign language learner. *Kwartalnik Neofilologiczny* LVIII, I/2011, 46–68.

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

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

Tsui, A. (2004). What teachers have always wanted to know. In J. Sinclair *How to Use Corpora in Language*

Teaching. Amsterdam: John Benjamins, 40–61.
Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

The Role of Restricted Collocations and Learner and Course Variables in Determining Writing Quality in Assignments from a First Year Composition Programme

Lee McCallum
University of Exeter
lm489@exeter.ac.uk

The quantitative measurement of writing quality has a long history via corpus and computational tools. Within this paradigm, researchers and practitioners alike aim to determine which linguistic features correlate with writing grade scores in often large-scale international first and second language assessments (Hunt, 1970, Kyle & Crossley, 2015, 2016,2018). This aim has led to a plethora of studies examining grade relationships with syntactic structures such as noun phrases, different types of clauses, and lexico-grammatical features that include counts of nouns, verbs, and features of metadiscourse (Biber et al., 2011; Ortega, 2015; Eckstein & Ferris, 2017). However, the phenomenon of collocation continues to lag behind this emphasis on individual structures and words (Paquot, 2017). While previous studies have studied collocation diversity and sophistication through the use of learner frequency counts and external benchmarks to reference corpora (Bestgen & Granger, 2014; Granger & Bestgen, 2014), few studies have used an extensive range of measures or factored in interactions between rater and context variables (Paquot, 2018).

Overlooking these variables has several important implications for our understandings of this relationship. Firstly, previous studies have used 'nested' corpora which contain essays written by the same writer or batches of essays graded by the same individual rater. The corpus structure may also contain essays written on essay tasks, topics and be written by individuals from different language backgrounds. On the surface, overlooking these variables limits our understanding of the above relationship and over simplifies the grading process. However, at the same time, the results of such analyses lack validity because the subsequent correlation and regression analyses violate the assumption that the feature counts and grades are independent observations, in other words, the counts of collocations and the grades emanate from individually different writers and raters. Therefore, there is a need for an approach that takes this nesting into account so that we can gain a more holistic understanding of relationships between features and grades as well as develop this understanding in a way that does not statistically invalidate the results.

Given these issues, this paper presentation contributes to this research area by examining sophisticated restricted lexical collocation use in assignments written by L1 and L2 first year university students enrolled in a large public university in the U.S. The presentation centres around answering the following research questions:

RQ1: What is the relationship between measures of lexical collocation diversity and writing quality grades in a corpus of FYC projects?
RQ 2: What is the relationship between measures of lexical collocation sophistication and writing quality grades in a corpus of FYC projects?
RQ 3: To what extent can these relationships be used to explain grade variation in a multi-level model when fixed and random contextual and learner variables are taken into account?

In answering these questions, the presentation reports preliminary findings from a multi-level model that looks at the relationship between collocations and essay grades when several fixed and random effects are taken into consideration. The fixed effects include the essay task, language status and grading scale type. The random effects include individual raters and student writers. A discussion of how findings influence current assessment practice in the FYC (First Year Composition) programme for the assessment of L1 and L2 writers concludes the presentation.

Textual borrowing from tasks and proficiency levels in assessment-related learner corpora: An exploratory study for DiSKo and MERLIN

Elisabeth Muntschick, Annette Portmann, Katrin Wisniewski
University of Leipzig
elisabeth.muntschick@uni-leipzig.de, annette@portmann.org, katrin.wisniewski@uni-leipzig.de

This work-in-progress contribution focuses on textual borrowing in learner corpora and is located at the interface of language testing and assessment and LCR (Callies & Götz, 2015; Barker et al., 2015; Wisniewski, 2017). Its aim is to better understand learners' use of borrowed structures (i.e. text copied from input) and to reflect methodological implications across research fields, as borrowing potentially threatens the validity of both (automatized) ratings and learner corpus-based studies.

While different task types are known to elicit different types of learner texts (Alexopoulou et al., 2017), there is little research on the influence of linguistic input on learner texts in terms of borrowing. Most corpus studies do not make it clear how borrowed text is treated. Previous studies mainly focus on academic referencing in summary writing (Shi, 2004; Keck, 2006, 2014) or integrated tasks (Plakans & Gebril, 2013; Weigle & Parker, 2012, 2014), showing proficiency-related differences in borrowing behavior. Shi (2012) revealed different perceptions of acceptable citation practices across disciplines. Montee (2017) was the first to analyze input-rich tasks in the context of optimizing test prompts.

This contribution addresses textual borrowing exploratorily by focusing on data from two learner corpora that contain language test texts, treating the research questions (RQ) below:

RQ1: How frequently does borrowing occur across CEFR levels?

RQ2: Do the types of textual borrowing differ across CEFR levels?

RQ3: Which linguistic structures are borrowed and in what way are they changed?

First, we focus on textual borrowing in DiSKo (*Deutsch im Studium: Lernerkorpus – German at the University: Learner Corpus*), which is currently being compiled. DiSKo uses a writing task from TestDaF (a university admission language test) performed by international students (currently, N=130) and an L1 control group (currently, N=14). All texts were rated B2-C1, and transcribed and annotated in EXMARaLDA (Dulko) (https://bitbucket.org). The corpus will be accessible via ANNIS (Krause & Zeldes, 2016) by 2020. Secondly, texts rated A2 (N=28) from the German section (N=1,033) of the CEFR-related MERLIN corpus are used (Abel et al., 2014; https://merlin-platform.eu) to cover a broader proficiency range.

An in-depth *annotation* of borrowed structures was carried out. A structure was defined "borrowed" if ≥3 words included in a defined syntactical range of the task appeared in a defined syntactical range of the learner text. We distinguished "copied" from "modified" borrowing: If the text was borrowed with no change of words or word order, it was tagged "copied". If ≥ 1 words were changed, omitted, or added, it was tagged "modified".

Data analysis involves quantitative and qualitative methodology. For RQ1 and RQ2, descriptive statistics is applied to examine normalized frequencies for borrowing as well as the occurrence of various types of borrowing at different CEFR levels. For between-group comparisons (L1/L2; CEFR levels), non-parametric tests of significance are run (Kruskal Wallis), and effect sizes are provided. The qualitative analysis (RQ3) has an inductive, data-based approach. Borrowed structures are examined for syntactic modifications.

Despite its limited sample size and its exploratory character, the methodological implications of this work-in-progress study might be considerable, touching questions of rating validity and the definitions of boundaries of "productivity" in SLA or "criterial feature" studies (Harrison & Barker, 2015). On a corpus linguistic level, the annotation of borrowing or the provision of input texts might turn out important issues.

References

Abel, A., Wisniewski, K., Nicolas, L., Boyd, A., Hana, J., & Meurers, D. (2014). A Trilingual Learner Corpus illustrating European Reference Levels. *Ricognizioni – Rivista Di Lingue, Letterature E Culture Moderne* 2, 111–126.

Alexopoulou, T., Michel, M. C., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: a large-scale learner corpus analysis employing Natural Language Processing techniques. *Language Learning* 67, 180–208.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: developing language assessments and justifying their use in the real world. *Oxford Applied Linguistics*. Oxford u.a.: Oxford Univ. Press.

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge Handbooks in Language and Linguistics. The Cambridge Handbook of Learner Corpus Research* (pp. 511–534). Cambridge: Cambridge University Press.

Callies, M., & Götz, S. (Eds.). (2015). *Studies in corpus linguistics. Learner corpora in language testing and assessment*. Amsterdam: John Bejamins.

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

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *Cambridge Handbooks in Language and Linguistics. The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Harrison, J. & Barker, F. 2015. *English Profile in Practice*. Cambridge: Cambridge University Press.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing* 15, 261–278.

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing* 25, 4–22.

Krause, T. & Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 2016 (31), 118–139.

Montee, M. (2017). *Input-rich Writing Tasks and Student Writing on an English Language Proficiency Test* (Dissertation). Georgia State University, Atlanta.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing* 22, 217–230.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., . . . Andreas, T. (2012). *Das Falko-Handbuch: Korpusaufbau und Annotationen Version 2.01* (Technical report). Humboldt University, Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/at_download/file

Shi, L. (2004). Textual Borrowing in Second-Language Writing. *Written Communication* 21, 171–200.

Shi, L. (2012). Rewriting and paraphrasing source texts in second language writing. *Journal of Second Language Writing* 21, 134–148.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing* 21, 118–133.

Wisniewski, K. (2017). Das Potenzial von Lernerkorpora im Sprachtestbereich. *Deusch Als Fremdsprache* 54, 33–40.

Automatic identification of unacquired linguistic concepts underlying grammatical errors in English learner writing

Mick O'Donnell
Universidad Autónoma de Madrid
michael.odonnell@uam.es

A lot of attention has been given to the automatic tagging of grammatical errors in learner productions in English (e.g., Menzel & Schroeder, 1999; Feng et al 2016; Lo et al 2018). However, the identification of a grammatical error does not by itself give sufficient information to be useful to a language learner. Each surface error can have a number of explanations. For example, learners of English commonly produce errors in Subject-Finite agreement. However, it is not usually the case that they don't understand the idea that Subject and Finite need to agree in number. It is more commonly a secondary concept that they have not acquired, for instance, Spanish learners of English often assume that "people" in English is singular because the corresponding word in Spanish, 'gente', is singular. Another cause is the case of learners producing existential clauses ("there are…") who do not understand that the Subject is actually the noun phrase that follows the verb. Yet a third explanation comes from the learner not understanding that noun phrases with "either" and "neither" are syntactically singular ("Either parent is…"). So, what is usually classed as a single error in error tagging systems often corresponds to several underlying concepts which need to be acquired separately. Just telling a student that the Subject and the Finite need to agree in number is not sufficient to correct the student's underlying misunderstanding.

We are currently developing an online grammar learning system for Spanish learners of English. The system identifies the grammatical rules which the user has not yet fully acquired, and focuses their study on those which are within their Zone of Proximal Development (Vygotsky, 1978).

Currently we are using sentence correctness probes to determine the learner's state of development for each concept, but we are moving towards augmenting the learner model using the student's written productions. The system analyses the learner's text, identifying cases in the text where the learner has applied grammatical rules incorrectly, and modifying the learner model to lower the estimated level of acquisition of the associated concept. Equally so, the student's correct application of a rule is meaningful, so the system also detects correct applications, which increase the recorded level of acquisition (cf. work by McCoy et al 1996, although that work assumes each surface error to have one cause only).

Learning a language involves acquiring a vast multitude of linguistic concepts, not all of which can be addressed by any learning system. To this end, our project analysed a 700,000 word corpus of learner English (in a Spanish context), and identified 16,000 errors. From these, we identified the 20 most frequent grammatical surface errors. We are in the process of coding each of these 20 errors more deeply in terms of the underlying broken concepts. This has provided us so far with around 100 linguistic concepts which we consider critical for our learner-base to master. Our automatic tagging is thus limited to identifying these 20 surface errors, and the 100 underlying concepts. Our system syntactically parses sentences using UDPipe (Straka & Straková, 2017), and applies a rule-based approach to identify errors. Not all errors are easily accessible to automatic tagging. Because of this, we will only address a subset of the 100 critical concepts that are most open to automatic identification.

This paper will report on our efforts towards automatic identification of correct and incorrect uses of grammatical concepts, and the application of this to build learner profiles.

References

Feng, H.H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016) Automated Error Detection for Developing Grammar Proficiency of ESL Learners. *CALICO* Vol. 33, No. 1 (2016), 49–70.

Lo, Y. C., Chen, J. J., Yang, C., & Chang, J. (2018) Cool English: A Grammatical Error Correction System Based on Large Learner Corpora. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 82–85).

Menzel, W. & Schroeder, I. (1999) Error diagnosis for language learning systems. *Special edition of the ReCALL journal*, 20–30.

Straka, M., Straková, J. (2017) Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, August 2017.

Vygotsky, L.S. (1980) *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

Learner Translator Corpus (LTC) as didactic material in translation classes

Joacyr Oliveira
Universidade de São Paulo/UNICAMP
joacyr@outlook.com

Due to management constraints, students in translation programs in Brazilian universities are taught in large groups, varying from 30 to 60 learners in one lecture room. Delivering feedback becomes mission impossible because not only desks are placed one behind the other as in traditional classroom settings, but also calling on students individually to read out loud their suggested translation to each segment of the source text may turn a once interesting class into a tedious and dreary task. How can a Learner Translator Corpus (LTC) be used as a powerful tool to transform such classes into interesting lessons that will cater for all students in large groups?

This presentation aims at suggesting a relatively pragmatic teaching method developed and systematized as a result of our extensive experience teaching translation practice courses in Brazil. By compiling an LTC, we were able to find ground work for class preparation, creation of translation awareness class activities and in-class individual feedback.

The teaching method proposed comprises some steps. First we need a Google Form which will be used by the students to work on their translation. This step is crucial for the corpus compilation. As translations of the source text are sent out, Google Forms engine automatically creates a spreadsheet in the instructor's Google Drive. The source text is added once to line one and the translations are added to the subsequent lines. At the end, each column contains a segment, the top line in the column contains the source text and all the lines underneath, the translations. Next, by using the Filter Tools available on the electronic spreadsheet, the instructor is able to analyze the target texts. If concordancing tools are available, the translations to each segment can be further analyzed when saved in *txt* format. The results deriving from the analyses provide input to lesson plans, which guide instructors preparing classes that will cater for students' real needs. Our experience in the classroom has shown that the same text translated by different groups generates distinct results. Consequently, each group requires a specific lesson plan. Finally the instructor uses the LTC as classroom material. With the aid of a computer, a projector and a screen, the spreadsheet created by Google Forms is projected to the whole group. The instructor stops at each segment and goes over all the students' translations, pointing out the strategies used and/or calling students' attention to any recurrent problems that might have occurred. Filters and other grouping and sorting resources can be used to pinpoint some important issues previously identified during lesson preparation. Together the whole class discusses the options and suggests improvements. At the end, a "final" translation is proposed based on everyone's collaboration.

As shown above, there are numerous advantages in using an LTC displayed on a spreadsheet. In addition to teaching a class that was prepared based on students' real needs, the corpus as a classroom material gives students the opportunity to read every suggested translation to each source language segment. As both instructor and students evaluate each translation, learners develop a sense of criticism and learn how to evaluate their own texts. Besides discussing translation strategies, corpus analyzes offer ground for reflections on grammar and meaning. Finally, a more down to earth consequence is that students work harder on their task, since their translation will be read by other classmates, and, as a result, become more involved in the class.

In this session, the presenter will initially go over the use of Google Forms to compile a learner corpus; next, he will use a real spreadsheet created with his students' translations to demonstrate how discussions can be led in class.

References

Bernardini, S. (2002). Educating translators for the challenges of the new millennium: The potential of parallel bi-directional corpora. In J. Haller, B. Maia, & M. Ulrych (Eds.), *Training the language services provider for the new millennium* (pp. 173–186). Porto: Universidade do Porto.

Hurtado Albir, A. (1999). *Enseñar a traducir: metodología en la formación de traductores e intérpretes - teoría y fichas prácticas*. Madrid: Edelsa.

Hurtado Albir, A. (2015). *Aprender a traducir del frances al español: Competencias y tareas para la iniciación a la traduccion*. Castelló de la Plana: Universitat Jaume.

Santos, D. (2014). (http://www.linguateca.pt/Diana/download/KK.pdf). Ser, estar, ficar, haver e ter contra ha, bli e være: quem disse que era fácil traduzir sentimentos e sensações? In S. Ebeling, A. Grønn, K. R. Hauge, & D. Santos, *Corpus-based Studies in Contrastive Linguistics*. Oslo: OSLa, Oslo Studies in Language.

Quantitative analysis of errors in the COPLE2 corpus

Iria del Río
CLUL, University of Lisbon
igayo@letras.ulisboa.pt

This work presents the first quantitative analysis of errors in the COPLE2 corpus (Mendes et al., 2016). We describe the data and the annotation system, and we present the first conclusions of our study.

Among other types of annotation, error tagging constitutes an important step since it helps to identify problematic areas in the learning process (Granger, 2004). In COPLE2, error annotation is in progress. So far, we have annotated a 62% of the corpus, and all the texts for four L1: Chinese, English, Italian and Spanish. The annotation schema, described in (del Río & Mendes, 2018), has three main classes: spelling, grammar and lexis. In this first analysis, we were interested in a general description of the errors we found, as well as in the interaction of these errors with the following variables: error category, L1, proficiency level[10] and number of tokens in the text.

In our data, 17% of the tokens show an error. Among the three error categories, the most common is grammar, followed by spelling and lexis. A Wilcoxon Rank-Sum test[11] showed that the difference in frequency between the three categories is significant. Concerning standard deviation, lexis shows the lowest value, and grammar the highest. To check the dispersion of the different types of errors, we used the DP measure (Gries, to appear). We got the highest DP for lexical errors (0.31), and the lowest for grammatical ones (0.24). This fact shows that grammatical errors are more equally distributed (without considering other factors as L1) than lexical or spelling ones.

Since the L1 is crucial in learning a second language, we were also interested in the way this variable interacts with the types of errors annotated. We found that the normalized frequencies of errors are different among the four L1. These frequencies, in fact, seem to indicate that there are two different groups of L1 in our data: Spanish and Italian, on the one hand, and Chinese and English, on the other. This result makes sense if we consider the linguistic distance between Portuguese and the four L1s under analysis. For Chinese and English, the most frequent type of error is grammar, followed by spelling and lexis. For Italian and Spanish, the most common type of error is spelling, followed by grammar and lexis. The lowest frequency of the lexis category is, therefore, a constant. If we consider a third variable, proficiency level, the general picture is the same, but we see specific differences between L1. For Chinese and English, grammar errors have a similar frequency in the three proficiency levels, but spelling errors are much more frequent in English than in Chinese. For the four languages and for all proficiency levels, lexis errors are pretty similar and remain constant.

The dispersion of errors by L1 shows that the lexical type is again the most dispersed. For grammar and spelling, we have now different results considering the L1: for Chinese and English, grammar is the less dispersed type, while for Italian and Spanish, it is spelling.

Finally, we performed an analysis of the correlation between the number of tokens and the number of errors per text, considering only the error types and the error types+L1. In both scenarios, we found that the correlation is always positive with a medium value. This result shows that, for further analysis, it would be good to consider the length of the text as a variable.

Our preliminary results show that the distribution of error categories is not balanced in the corpus, and it is connected with variables like text length, L1 or proficiency. As immediate future work, we plan to test statistically the impact of variables like L1, proficiency, text length or type of text in the frequency of errors.

References

Del Río, I., & Mendes, A. (2018). Error annotation in the COPLE2 corpus. *Revista Da Associação Portuguesa De Linguística*, (4), 225–239.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor, & T. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 123–145). Amsterdam & Atlanta: Rodopi.

[10] We consider only the proficiency levels documented for the four languages: A2, B1 and B2.
[11] The distribution of errors by category was not normal.

Gries, Stefan Th. (To appear). Analyzing dispersion. In M. Paquot, & S.Th. Gries (Eds.). *Practical handbook of corpus linguistics*. Berlin & New York: Springer.

Mendes, A., Antunes, S., Janssen, M., & Gonçalves, A. (2016). The COPLE2 Corpus: A Learner Corpus for Portuguese. In *Proceedings of LREC 2016*, Portorož, Slovenia.

Five Key Lexemes in German and Danish Academic Language

Irene Simonsen
University of Southern Denmark
simonsen@sdu.dk

The focus of this study is on the academic writing competences of L2-Danish BA students with German as a native tongue (CEFR B2-C1). The study compares the academic writing of these students to on the one hand, L1-Danish novice and expert standard, and on the other hand, L1-German expert standard.

Factors like input, exposure to the genre in question, and cultural and social integration (Bondi/Lorés-Sanz 2014; Paquot/Granger 2012; Henriksen 2013) may influence the language of L2 learners. Hüttner (2007) furthermore argues that the L2 learners' academic level of "apprenticeship" (Hüttner 2007) plays a role. Durrant and Mathews-Aydınlı (2011) have moreover shown that there are quantitative differences between which features are realized in the texts of students and the texts of experts respectively, which, they conclude, may be due to different expectations in respectively novice and expert text (Durrant&Mathews-Aydınlı 2011: 71). However, the lexical choices made by L2 learners may also be signs of L1 transfer and reflect general variation in the use of the lexemes in the L2 language and L2 learner's native language. Both erroneous and deviating under- and over-use of L2 units are justified by transfer from their L1 language (Paquot&Granger 2012: 140f., Nesselhauf 2005).

On the basis of this, one may pose the question: may we learn more about the choices made by L2-Danish learners by looking at comparisons of, on the one hand, L2-Danish and L1-Danish learner texts, and, on the other hand, German and Danish experts writing in their mother tongues?

The study examines the use of some very common academic terms, namely the 5 lexemes *analysis*, *investigation* (dan. *undersøgelse*, germ. *Untersuchung*), *method*, *theory*, and *empirical* in 4 corpora: one L2 Danish corpus (bachelor's level), two L1 Danish corpora (master's level, PhD level), and one L1 German corpus (PhD). Only texts from the humanities are examined. First, the frequency and distribution of the five lexemes in the corpora are measured, and secondly, the co-occurrences of the lexemes in the following collocations are compared: as verbs in noun + verb, verb + noun collocations, as nouns in noun + verb, verb + noun, noun + noun, adjective + noun, preposition + noun, and as adjectives in adjective + noun collocations.

In conclusion, there seem to be differences with regard to frequency and the use in collocations. While the three former lexemes are partly, to a large extent, over-represented in L2 Danish in comparison to the L1 Danish norm, the latter (*theory*, *empirical*) are proportionally under-represented in the L2 Danish learner texts. Furthermore, there is a difference in the use of the lexemes in collocations, especially as nominals and as adjectives. And finally, frequency difference can be observed relative to the functionally use in the sense that for instance a frequent use of the lexemes in prepositional phrases distinguishes the L2 learners from the Danish NSs.

References

Bondi, M., Lorés-Sanz, R. (Ed.) (2014): *Abstracts in Academic Discourse. Variation and Change.* Frankfurt am Main: Peter Lang.

Durrant, P., Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. In *English for Specific Purposes* 30, 58–72.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development – a progress report. In *Eurosla Monographs Series 2, L2 vocabulary acquisition, knowledge and use*: 29–56.

Hüttner, J.I. (2007). *Academic Writing in a Foreign Language. An Extended Genre Analysis of Student Texts*, Frankfurt a.M.: Peter Lang.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Paquot, M., Granger, S. (2012). Formulaic Language in Learner Corpora. In *Annual Review of Applied Linguistics*, 32, 130–149.

Phraseological Complexity as an Index of L2 French Writing Proficiency

Nathan Vandeweerd, Alex Housen, Magali Paquot
Université catholique de Louvain, Vrije Universiteit Brussel
nathan.vandeweerd@uclouvain.be, alex.housen@vub.be, magali.paquot@uclouvain.be

This study aims to expand the construct of L2 complexity beyond purely syntactic or lexical measures to include the ways in which words combine to form meaningful word combinations. We define phraseological complexity as the "range of phraseological units that surface in learner production and the degree of sophistication of such phraseological units" (Paquot, 2019, p. 124). For L2 English, previous research has shown that measures of phraseological complexity can predict proficiency level better than measures which target solely syntactic or lexical characteristics of a text (Paquot, 2019). Although research in L2 French has shown that learners use more phraseological units overall as they increase in proficiency (Forsberg & Bartning, 2010), thus far no study has compared phraseological complexity (i.e. diversity and sophistication of phraseological units) across proficiency levels in L2 French. The current study attempts to fill this gap by providing cross-linguistic validation of the results of Paquot (2019) for L2 written French.

The data for this study come from the *Leerdercorpus Frans*, a 100,000 word corpus of argumentative essays written by L1 Dutch parser. Phraseological sophistication is operationalized as the mutual information score of those units. In addition to phraseological complexity, we also calculate several measures of syntactic and lexical complexity. Measures of syntactic complexity include measures of length (mean length of clause), subordination (clauses per T-unit, dependent clauses per T-unit, dependent clauses per clause) and part-of-speech based structures (verb phrases per T-unit, complex nominals per T-unit and complex nominals per clause). Measures of lexical complexity include measures of diversity (transformations of type-token-ratio) and sophistication (proportions of low-frequency words). Mixed effect regression analyses are used to determine which of the phraseological, syntactic and lexical complexity measures best account for human raters' overall proficiency assessment (in terms of CEFR levels) of the same texts.

In line with Paquot (2019), we expect to find that measures of phraseological complexity will better predict L2 proficiency level than syntactic and lexical complexity measures, especially for highly advanced learners. However, the burden of learning a relatively richer inflectional system may mean that our learners of L2 French do not exhibit the same degree of phraseological complexity as learners of L2 English (cf. Stengers, Boers, Housen, & Eyckmans, 2011).

References

Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing* (Vol. 1, pp. 133–157). European Second Language Association.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research [Special issue on linguistic complexity]. *Second Language Research*, *35*(1), 121–145.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *IRAL - International Review of Applied Linguistics in Language Teaching*, *49*, 321–343.

Chaos is merely order waiting to be deciphered: Corpus-based study of word order errors of Russian learners of English

Olga Vinogradova, Elizaveta Ershova, Aleksandr Sergienko, Darya Overnikova, Anton Buzanov
National Research University Higher School of Economics in Moscow
olgavinogr@gmail.com, eoershova@edu.hse.ru, alser99@yandex.ru, darya.snailstep@gmail.com, anton.buzanov.00@gmail.com

Researchers of word order errors have written on difficulties of diagnosing them (cf. Boyd & Meurers 2008) and about the interplay of a learner's native free word order language with the English word order standards (cf. Harves 1998, Hoffman 1996, Rankin 2009, Rebuschat & Williams 2009). The specific areas in focus of the previous research have been verb-subject structure, or postverbal subjects (Oshita 2004; Rankin 2009; Lozano & Mendikoetxea 2010); "breakability" of some close-knit constructions; the order of multiple premodifiers (Koliopoulou 2019, Wulff 2003, Zielinska 2007, Toldova & Mukhina 2017); the position of modifiers; word order in negative constructions (cf. Pitts 2005, Fuentes 2008). The goals of the present study were the following:

- annotate word-order errors in written essays in the corpus of Russian Error-Annotated Learner Corpus of English (http://realec.org)
- choose classes of word order errors that can be diagnosed using lexically-anchored patterns (Metcalf and Meurers 2006)
- analyse regularity in wrong phrase sequences
- evaluate possibility and effectiveness of regular expressions for identifying word order errors
- analyse possible interference with Russian as L1.

We identify 7 classes of most frequently occurring learner errors:

1) post-verbal subjects in cases of unnecessary focalization (such as *Second result shows Sweden, and after that USA).* As it is a prevailing word order in sentences of this type in Russian, Russian students familiar with the possibility of fronting focalized group overuse it;

2) absence of inversion in direct questions and unnecessary inversion in indirect questions (sentences like *It has been widely discussed in recent years whether is it relevant to reduce the number of air flights or not* and *What you would see on the streets?*) -interference with Russian is quite evident here: Russian word order in this case is literally the following: BE-interrogative particle-relevant..., while the direct questions in Russian do not have any change in word order at all;

3) position of the direct object (cf. fronting as in *Such a scenario South Park creators Trey Parker and Matt Stone predict,* or adverbial modifier in front of direct object, such as *As a result, people define differently the happiness*); the correlation with focalisation of the indirect object or modifier in the Russian equivalent is often attested;

4) positions of multiple modifiers in a syntactic group (as in *The consumption of electricity gradually decreases to reach its lowest point around 9 o'clock of a bit more than 10000 units* or *There was a sharp growth in 2030 by 25%*);

5) positioning of discourse-navigating adverbial modifiers - *also, for example, enough, too* (as in *Also it is possible to reach their, for example, workplaces not using private cars,* or *We also can notice that death rate is low)*;

6) positioning negation (*not)* within infinitival, attributive and adverbial phrases instead of the verbal group. (*However, others feel not the same way; It is wrong to not say...* are prototypical examples);

7) construction within an attributive group placed before the head noun (such as *the percentage of 65 and over aged people).*

The procedures in the research included the following stages: wrong patterns were identified by annotators; sequences of parts of speech (tagged by TreeTagger) with lexical anchors were constructed; they were translated into regular expressions; sentences found by regular expressions were evaluated with regard to True/False positives; models were updated accordingly. The initial results of applying three such models to student texts are precision - 83%, recall - 42%.

References

Boyd, A., & Meurers, D. (2008). On Diagnosing Word Order Errors. Accessed at https://www.researchgate.net/publication/242239417_On_Diagnosing_Word_Order_Errors

Fuentes, A.G.(2008). The Use of English Negation by Spanish Students of English: a Learner Corpus-Based Study. In M. J. Lorenzo Modia (Ed.), *Proceedings 31st AEDEAN Conference*, pp. 315–327. A Coruña: Universidade.

Harves, S. (1998). The Syntax of Negated Prepositional Phrases in Slavic. In Z. Boskovic et al. (Eds.), *FASL 6, The Connecticut Meeting*. Ann Arbor, MI: Michigan Slavic Publications, pp. 166–186.

Hoffman, B. (1996). Translating into free word order languages. *COLING'96 Proceedings of the 16th conference on Computational Linguistics* Vol. 1, pp. 556–561.

Koliopoulou, M. (2019) *Compounds and multi-word expressions in Greek*. De Gruyter. Accessed at https://www.academia.edu/38132304/Compounds_and_multi-word_expressions_in_Greek.pdf

Lozano, C., & Mendikoetxea, A. (2010). Interface conditions on post-verbal subjects: A corpus study of L2 English. *Bilingualism: Language and Cognition* 13 (04), 475–497.

Metcalf, V. and Meurers, D. (2006). *Exploring Interfaces and Issues in ICALL* https://www.researchgate.net/profile/Detmar_Meurers/publication/237706513_Exploring_Interfaces_and_Issues_in_Intelligent_Computer-Aided_Language_Learning_What_the_OSU_ICALL_group_is_up_to/links/0f31753372a57556ff000000/Exploring-Interfaces-and-Issues-in-Intelligent-Computer-Aided-Language-Learning-What-the-OSU-ICALL-group-is-up-to.pdf

Oshito, Y. (2004) Is there anything there when there is not there? Null expletives and second language data. *Second Language Research* 20(2), pp. 95–130.

Pitts, A. (2005). *The International Corpus of English (GB). A Study of Negation*. Accessed at http://ucrel.lancs.ac.uk/publications/CL2007/paper/61_Paper.pdf

Rankin, T. (2009). Verb Second in Advanced L2 English: A Learner Corpus Study. In M. Bowles et al. (Eds.), *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference*, pp. 46–59. Somerville, MA: Cascadilla Proceedings Project.

Rebuschat, P., & Williams, J. N. (2009). Implicit learning of word order. In N. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, pp. 425–430.

Toldova, S. & Mukhina, R. (2017). Multiple prenominal adjectives ordering in Russian: A corpus study. In *Proceedings of 24th conference Computational Linguistics and Intellectual Technologies* Vol. 2, pp. 429–441 (in Russian).

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. In *International Journal of Corpus Linguistics* 8(2), pp. 245–282.

Zielinska, D. (2007). A Polish-English Contrastive Study of the Order of Premodifying Adjectives: A Procedural Model Account. In M. Davies, P. Rayson, S. Hunston, & P. Danielsson (Eds.), *Proceedings of the Corpus Linguistics Conference CL2007*, article #5.

Collecting Learner Language Data through Crowdsourcing

Elżbieta Gajek
University of Warsaw
e.gajek@uw.edu.pl

There are various methods of collecting learner data. In the first one unified groups of learners write texts for analysis. In the second one a crowd of learners, having unknown profile, produces texts even being unaware of the use of their material as a source of research.

The former is called explicit crowdsourcing that is collecting data with the help of teachers has been the method used for many years in learner corpus studies. The teachers set the conditions of the study and control the students' output. They know the sociolinguistic parameters of the providers of the data.

The latter, that is the implicit crowdsourcing, is a voluntary learner contribution to the development of language materials usually for other reasons than collecting learners' language data. The technique seems to be very promising as it triggers active learning, that is learners learn while using and creating a crowdsourcing activity. However, the conditions and boundaries of the processes of collecting data are less controlled. If the crowdsourcing activity is open to a wide public it is difficult to get sociolinguistic parameters of the contributors. Many of them get discouraged and withdraw from the activity when they are asked about basic personal data so it is almost impossible to verify such data. What is more, the motivational factors of the contributors may vary and usually they are unknown to the researcher. Another issue is the quality of data and their comparability. This seems to be a critical issue for learners corpora researchers.

Although there are highly successful examples of a wide response to crowdsourcing activities as means for language learning, such as Duolingo, Busuu, Memrise, which get millions of contributors. There are also studies which demonstrate reluctance to participation in such activities. This approach may strongly affect specialists and experts, who potentially may provide high quality data.

Both techniques of collecting data will be presented, basing on the results gained in enetCollect project: Combining Language Learning with Crowdsourcing Techniques that is COST CA16105 action from January 2017 to December 2019. The project focuses on enhancing the production of learning material in order to cope with the increasing demand for language learning and the striking diversification of learner profiles due to the intensified migration flows motivated by educational, professional/economic or geopolitical circumstances. In the presentation its potential for collecting learner language data is discussed. Over 100 researchers from 35 countries participate in this project so it covers various dimensions of the topic.

Thus the author will focus on her own contribution to collecting data. One set of data include 98 short (ca 130 words) essays written by learners of English in a Polish secondary schools and collected via explicit crowdsourcing

The most effective approach to promote the use of crowdsourcing techniques among language learners is to encourage language teachers to introduce crowdsourcing platform in class. Thus, the other set of data is collected via implicit crowdsourcing from language teachers on their knowledge and readiness to involve crowdsourcing activities in their teaching. The response of teachers was lower than expected which illustrates lower effectiveness of collecting data among experts via implicit crowdsourcing.

The exact data, what users do, are available to the owners of the crowdsourcing platforms as implicit crowdsourcing activities take place in informal settings when contributors are even not aware that they participate in crowdsourcing. Thus, it is hardly to assess their effectiveness and impact on learning as there is hardly possible to measure the progress the learners make with the use of rigorous scientific methods. The studies carried out on selected users refer to their opinions and satisfaction more than the effectiveness of learning. However, the fact that millions of language learners use and contribute to their development should focus attention of teachers, and researchers in the field.

References

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence* 14 (1): 75–90.

Doan, A., Ramakrishnan, R. and Halevy, A. Y. (2011). Crowdsourcing Systems on the World-wide Web. *Communications of the ACM*, 54 (4): 86–96.

Garcia, I. (2013). Learning a Language for Free While Translating the Web. Does Duolingo Work? I*nternational Journal of English Linguistics*; Vol. 3, No. 1, 19–25.

Lyding, V., Nicolas, L. Bédi, B. and Fort, K. (2018). Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing Techniques (enetCollect). *Future-proof CALL: Language Learning as Exploration and Encounters – Short Papers from EUROCALL 2018*, 176.

Munro, R. (2013). Crowdsourcing and the Crisis-Affected Community: Lessons Learned and Looking forward from Mission 4636. *Journal of Information Retrieval*, 16, 21–266.

Odo, D. M. (2016). Crowdsourced Language Learning: Lessons for TESOL from Online Language-Learning Enthusiasts. *English Teaching Forum* 54 (4): 14–23. ERIC. Retrieved from https://files.eric.ed.gov/fulltext/EJ1123197.pdf.

von Ahn, L. (2013). Duolingo: Learn a Language for Free While Helping to Translate the Web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (pp. 1–2). New York, NY, USA: ACM. https://doi.org/10.1145/2449396.2449398

Vesselinov, R., & Grego, J. (2012). *Duolingo Effectiveness Study*. DuoLingo. Retrieved from https://www.duolingo.com/effectiveness-study

Bridging further comparison-based and detection-based arguments for crosslinguistic influences

Ilmari Ivaska
University of Turku
ilmari.ivaska@utu.fi

In his seminal work, Jarvis (2000; 2010) postulated 4 necessary conditions for empirically verifying crosslinguistic influences (CLI): 1) similarities between a given L1 and the studied L2, 2) differences in other L1s and the studied L2, 3) similarities in L2 behavior among people with a shared L1, and 4) differences in L2 behavior between people with different L1. Fulfilling these conditions constitutes a comparison-based argument for CLI. A detection-based argument for CLI (e.g. Jarvis 2010) in turn is data-driven and only takes into account the product-related premises: 1) similarities in L2 behavior among learners with a shared L1 and 2) differences in L2 behavior between learners with different L1. Comparison-based argument leads to reliable results, but it may omit more elusive forms of CLI. Detection-based approach identifies such cases better – but the results may be difficult to interpret. When the two approaches have been brought together, the interpretations have typically focused on single differences detected in a corpus-driven analysis or on the overall classification accuracy of either pre-defined or corpus-driven features (Jarvis & Crossley 2012). The present paper aims at narrowing this gap by means of a 2-phase methodological process: a bottom-up detection of consistent inter-L1 differences, followed by a dimension reduction to group the found features and to interpret them in terms of independently documented typological differences. The approach is tested parallelly in two typologically diverging languages: L2-English (L1s: Czech, German, Finnish) and L2-Finnish (L1s: Czech, German, English).

Data and Methods

Our data come from ICLE for L2-English, and ICLFI, LAS2 and YKI for L2-Finnish. All included data are argumentative texts of advanced proficiency. We annotated the texts with universal dependencies (Straka & Straková 2017). Our feature set consists of the POS bigrams defined by their dependency relations (figure 1). Such bigrams provide information on POS, syntactic functions as well as word order. We used Boruta variable selection (Kursa & Rudnicki 2010), an implementation of random forests, to detect the bigrams that contribute to distinguishing enL2-L1fi and fiL2-L1en data from their respective comparison data. We then explored the correlations between these features using Exploratory Factor Analysis (EFA). The resulting factors – the sets of inter-correlated bigrams – were linked to correspondingly grouped typological differences documented in the World Atlas of Language Structures (Dryer & Haspelmath 2013) to construct a comparison-based argument for CLI.

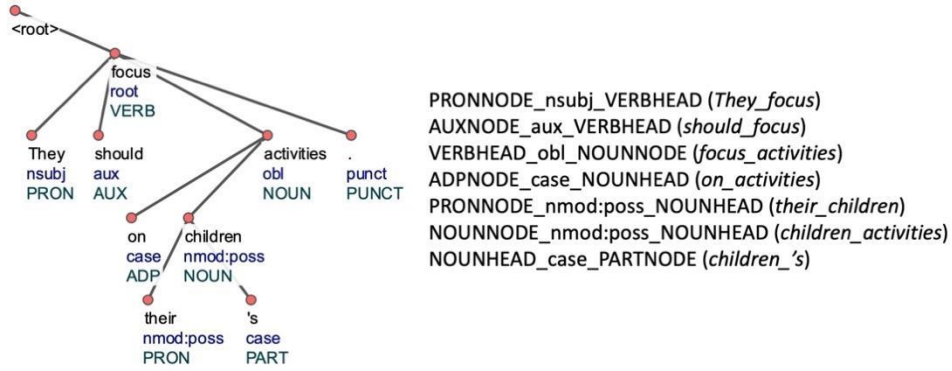


Figure 1. Example of dependency-defined POS bigrams.

Preliminary Results and Discussion

We extracted frequencies of 695 bigrams in the English data and of 437 bigrams in the Finnish data (threshold: 5 occ. / subcorpus). Boruta identified for both datasets 15 bigrams that contributed to distinguishing the contrasted data. Based on these variables, the EFA suggested a solution with 4 factors for both English and Finnish data. For this abstract, we highlight factor 1 in English data, but the rest of the factors have been explored in a similar fashion. Factor 1 consists of the bigrams NOUNNODE_nmod:poss_NOUNHEAD and

NOUNHEAD_case_PARTNODE. Both reflect the use of *s*-genitive, which are more common in enL2-L1fi than in enL2-L1other (enL2-L1fi: 2.1 / 1000 words vs. enL2-L1other 1.3 / 1000 words). The difference can be related to structural differences in genitive phrases (Dryer 2013). Contrary to Czech and German, in Finnish the genitive NP is structurally very similar to *s*-genitive, with a corresponding word order and morphologically marked possessor.

Our results support the applicability of the method in linking detection-based and comparison-based arguments. It provides with statistically identified inter-correlated linguistic features that characterize the L1-specific datasets and that can be related to potential typological differences. Verifying a comparison-based argument for CLI still requires closer study of contexts of use, but the results can be used to target the enquiry to meaningful directions and to both corroborate earlier findings and reveal previously unearthed CLI.

References for Corpora

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

ICLFI = International Corpus of Learner Finnish. Jantunen, J. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. Lähivõrdlusi. *Lähivertailuja* 21, 86–105.

LAS2 = The Corpus of Advanced Learner Finnish. Ivaska, I. 2014. The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples – Journal of Applied Language Studies* 8(3), 21–38.

YKI = *The National Certificates of Language Proficiency Corpus*. 2019. Centre for Applied Language Studies, University of Jyväskylä.

References

Dryer, M. (2013). Order of Genitive and Noun. In M. Dryer, & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

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. *EUROSLA Yearbook* 10, 169–192.

Jarvis, S., & Crossley, S. (Eds.). (2012). *Approaching Language Transfer through Text Classification: Explorations in the detection-based approach*. Bristol, Buffalo, Toronto: Multilingual Matters.

Kursa, M., & Rudnicki, W. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, Articles* 36(11), 1–13.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston: Northwestern University.

Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task*, 88–99. Vancouver: ACL.

Dynamic changes in the development of L2 inflectional morphology

Lucie Jiránková, Luca Cilibrasi
Charles University
Lucie.Jirankova@ff.cuni.cz, Luca.Cilibrasi@ff.cuni.cz

The study of learner language by means of psycholinguistics is an established, if a rather novel research method (Durrant&Siyanova-Chanturia, 2015: 57-76). Integrating corpus data with experimental methods and using their complementary perspectives on linguistic phenomena, combined with the use of new statistical approaches (e.g. mixed-effects modelling), enables us to gather information about different aspects of learner language and thus yield a more complete picture. The aim of this study is to examine the dynamic changes in the development of inflectional morphology in second-language learners of English with Czech as L1 and compare this development with native speakers.

Since the advent of psycholinguistic research, there has been a major debate about a speaker's ability to produce novel morphologically inflected forms. Two opposing accounts of morphological productivity have been proposed: one attributing the productivity to the application of rules (Prasada&Pinker, 1993) and one attributing it to analogy based on stored exemplars (Bybee&Slobin, 1982). With its relatively clear distinction between regular and irregular patterns of inflection, English past-tense morphology provides a particularly suitable framework to decide between these two approaches. Consequently, two models have been proposed: the single-route model (e.g. Bybee&Moder, 1983), which posits that both regular and irregular past-tense forms are generated by analogy across stored exemplars (e.g. heal/healed - steal/*stealed, see (ii)), and the dual-route model (e.g. Prasada&Pinker, 1993), which posits that regular forms are generated via the application of a default rule (–ed) and irregulars are generated by analogy (see (i)).

This study builds on previous work by Albright and Hayes (2003) and Blything et al. (2018) and uses an elicited production paradigm to investigate which of the two models best accounts for L2 learners' morphological productivity. 88 adult English L2 learners with L1 Czech at A1-C1 proficiency levels and a control group of 9 native speakers heard sentences with someone performing a novel action described with a nonword (e.g. *The baby likes to bize. Look, there he is bizing. Everyday he bizes.*). Past-tense forms were then elicited by prompting the participant to describe what the agent "did yesterday." Produced forms were recorded and analysed with a binomial linear mixed-effects model in the R environment.

The results showed that different language levels perform differently. For native speakers, the likelihood of a verb being produced in regular past-tense form was positively associated with its phonological similarity to existing regular verbs (in line with the single-route model and the findings of Albright and Hayes (2003) and Blything et al. (2018)). However, L2 learners showed lesser dependence on the verbal similarity to regulars. The A1-, A2-, and B1-level participants did not rely on the nonword's similarity to existing regulars or irregulars to produce the inflected form. In contrast, a main effect of similarity-to-regulars and similarity-to-irregulars was observed with the B2-level and C1-level participants, respectively.

The results indicate that the L2 acquisition of the English past-tense is characterized by a progressive development from the mastery of mechanistic rules (~ the dual-route mechanism at the A1, A2, and B1 levels) to the refinement of their application by spotting analogical patterns of existing verbs (~ the single-route analogical mechanism at the B2 and C1 levels, ~ native speakers). The second-language speakers thus show dynamic changes in the development of inflectional morphology that come closer to native speakers with the higher proficiency of B2 and C1 levels.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90. 119–161.

Blything, R. P. et al. (2018) Children's Acquisition of the English Past-Tense: Evidence for a Single-Route Account from Novel Verb Production Data. *Cognitive Science*. 1–19.

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language* 59. 251–270.

Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language* 58. 265–289.

Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (Cambridge Handbooks in Language and Linguistics, pp. 57–78). Cambridge: Cambridge University Press.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes* 8. 1–56.

Specifics of the acquisition of a closely related language in a corpus of Czech produced by Polish learners

Elżbieta Kaczmarska, Gabriela Gawrońska
University of Warsaw
e.h.kaczmarska@uw.edu.pl, g.gawronska@student.uw.edu.pl

After presenting our project we show possibilities of analysing learner language using a corpus compiled from Czech texts, produced by students who are native speakers of Polish.

With Polish (L1) as a language closely related to Czech (L2), a strong L1 interference is observed at all levels – pronunciation, morphology, syntax, lexicon, including phraseology (false friends), and metalinguistic communication. To make teaching (and learning) more efficient, we need to focus on specific weaknesses and strengths of the learner on any level. To identify them, both incorrect and correct use of Czech by the learners should be studied. For this purpose, we build a corpus of Czech texts produced by Polish students by extending the L1 Polish – L2 Czech subcorpus of CzeSL (Czech as a Second Language), a learner corpus built at Charles University in Prague.

Before the start of our project, the Polish–Czech subcorpus of CzeSL was quite small (77 texts, 15 thousand words). Currently the Polish–Czech subcorpus of CzeSL is significantly larger (200 texts, 60 thousand words). However, it still requires not only collecting new texts and cooperation with the CzeSL team, but also applying some subtle changes to the annotation system of CzeSL, considering the common mistakes made by Polish learners of Czech, such as the use of a single Polish equivalent z for the two Czech prepositions s and z:

cz	Tom	pracuje	s	Michalem.
pl	Tom	pracuje	z	Michałem.
en	Tom	works	with	Michal.
cz	Robert	je	z	Polska.
pl	Robert	jest	z	Polski.
en	Robert	is	from	Poland.

Another typical mistake is the use of the Polish genitive of negation instead of accusative in Czech:

cz	(já)	Nemám	čas.
pl	(ja)	Nie mam	czasu.
en	I	do not have	time.

In the first step, the collected texts are annotated automatically using a toolchain tailored to Czech learner texts: (i) a standard Czech tagger to lemmatise and tag the original, (ii) a context-sensitive spell checker to correct most errors in spelling and morphosyntax, (iii) the same tagger to retag the result and (iv) a purpose-built tool to compare all corrected word tokens with their original forms, suggesting appropriate error labels (Jelínek 2017). After revision of the automatic annotation, the texts are manually annotated by error tags from the modified and extended CzeSL tagset, specifying more sophisticated error types beyond the reach of the automatic tool. The next step is the analysis of the annotated errors, a list of the most common error types, and a statistical summary of the results, taking into account the learner metadata such as proficiency level. Currently, we are identifying mistakes typical for Polish speaking students and analysing error types. At the conference we will present the results, including proposals for preventive and corrective exercises.

The results will be used by educators to produce teaching materials such as drills and other exercises, finely tuned to target the most disturbing errors, including grave failures in the achievement of a communication goal (Gawrońska 2018). We also hope to motivate experts in teaching a closely related language to conduct research and analyses using the annotated corpus data, and to encourage teachers to use the corpus and the research results in their teaching practice. For Polish speakers it would mean a significant improvement in the quality of learning Czech as a foreign language (Kaczmarska 2017).

References

Gawrońska, G. (2018). Směry vývoje studentského korpusu češtiny pro polsky mluvící studenty. *Konference mladých slavistů* XIV. 1.–2.11.2018, Charles University, Prague, Czech Republic.

Jelínek, T. (2017). Errors in inflection in Czech as a second language and their automatic classification. In K. Ekštein, & V. Matoušek, (Eds.) "Text, Speech, and Dialogue 20th International Conference", TSD 2017,

Prague, Czech Republic, August 27–31, 2017, *Lecture Notes in Artificial Intelligence* series, pp. 263–271. Springer International Publishing.

Kaczmarska, E. (2017). Towards a learner corpus of Czech for Polish speaking students. *Workshop on Interoperability of Second Language Resources and Tools*. 6–8.12.2017, University of Gothenburg, Sweden.

Rosen, A. (2016). Building and using corpora of non-native Czech. *ITAT 2016 Proceedings, CEUR Workshop Proceedings*, Vol. 1649, pp. 80–87.

Šebesta, K. (2010). Korpusy češtiny a osvojování jazyka (Corpora of Czech and language acquisition). *Studie z aplikované lingvistiky / Studies in Applied Linguistics*, pp. 11–34.

A Brazilian corpus of spoken learner English calibrated to the CEFR: From corpus design to data collection challenges

Mateus Miranda
University of Limerick
mateus.desouza@mic.ul.ie

Corpus linguistics (CL) as a language analysis and description tool has greatly contributed to language learning and teaching. Through learner corpora analyses, studies have been done on real language in use that contribute to the understanding of learners' difficulties in second language acquisition. Despite the many benefits to studying learners' interlanguage, many available corpora are not calibrated to the Common European Framework (CEFR; Council of Europe, 2001; Companion Volume, 2018) levels that identify the specific forms of any given language (words, grammar, etc.) at each of the six reference levels. These levels can be set as objectives for learning or can be used to establish whether or not a user has attained the level of proficiency in question (EnglishProfile, 2011). By not having a corpus according to the CEFR, it is difficult to establish learners' competence that has to be analysed as a whole. Furthermore, learners would classify their English level based on their years of experience, and many times a student would specify an unattained level of proficiency. Research regarding learner corpora in Brazil is commonly done through contrastive interlanguage analysis (CIA) and is mainly focused on written corpora (Berber Sardinha 2001; Shepherd 2009; Shepherd et. al., 2012). There are interlanguage written corpora such as the International Corpus of Learner English (ICLE) and the Louvain International Data Base (LINDSEI) (Granger et al., 2009 & 2010) under construction, but only one Brazilian spoken interlanguage corpus within the LINDSEI project has been compiled (LINDSEI-BR; Mello et al., 2012). Although much has been done on written corpora calibrated to CEFR (Capel, 2010; Buttery & Caines, 2012; Harrison & Barker, 2015; O'Keeffe & Mark, 2017; Díez-Bedmar, 2018), research on calibrated spoken corpora is in its nascence (Trinity Lancaster Corpus, 2010; Jones et al., 2017).

Taking into account the aforementioned considerations, the aim of this presentation is to offer a starting point to the creation of a national Brazilian corpus of spoken learner English, containing one million words, to make it available to the scientific community for research on interlanguage. The criteria for being a participant in this project are being a university student, participating in the English without Border (EwB) National Programme in Brazil, and holding an international proficiency certificate. A future step in this research is to investigate learner`s spoken grammar and how pragmatic markers (PMs) are developed across the CEFR levels. According to Carter & McCarthy (2001), the need for the investigation of spoken grammar within the greater language education community is urgent. Knowing how people use spoken grammar on a day-to-day basis can strongly benefit communicative approaches. Within spoken grammar, PMs are a class of items that operate outside the clause by encoding speakers' intentions and interpersonal meanings. Therefore, this study falls within the framework of Corpus Pragmatics (CP), which is a recent combination of Pragmatics and CL methods. This poster will detail the planning phase of designing the corpus matrix calibrated to the CEFR, such as ethics requirements, metadata information, hours of recorded speech and size, the recruitment of learners sorted according to the CEFR levels, and the task variables for the recording sessions. A pilot sub-corpus compilation was conducted in order to identify issues related to the transposition of speech to written form and to establish the conventions to represent speaker turns, pauses, hesitations, overlapping occurrence, among other spoken features. Moreover, this initial material will also be used at training sessions for transcribers in order to validate transcription.

References:

Berber, A. P. (2001). O corpus de aprendiz Br-ICLE. In *Intercâmbio* 10, 227–239.

Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam/ Philadelphia: John Benjamins, 187–204.

Carter, R., & McCarthy, M. R. (2001). 'Ten criteria for a spoken grammar.' In E. Hinkel, & S. Fotos (Eds.) *New Perspectives on Grammar Teaching in Second Language Classrooms*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Capel, A. (2010). A1 – B2 Vocabulary: Insights and issues arising from the English Profile Wordlists project. In *English Grammar Profile Journal* 1(1), 1–11.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: CUP.

Council of Europe. (2017). *The European Framework of Language Learning, Teaching and Assessment Companion Volume with New Descriptors*.

Díez-Bedmar, M. B. (2018). Fine-tuning descriptors for CEFR B1 level: insights from learner corpora, *ELT Journal* Volume 72, Issue 2, pp. 199–209.

English Profile: The CEFR for English. (2011). *English Profile Information Booklet*. [online] Available at: http://www.englishprofile.org/resources/information-booklet [accessed 10 April 2018].

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English* 1. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

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

Harrison, J., & Barker, F. (Eds.) (2015). English Profile in Practice. In *English Profile Studies* 5. Cambridge: Cambridge University Press.

Jones, C., Byrne, S., & Halenko, N. (2017). *Successful spoken English: Findings from learner corpora*. doi:10.4324/9781315101712.

Mello, H., Ávila, L., Orfano, B. M., & Neder Neto, T. (2012). LINDSEI-BR: an oral English interlanguage corpus. *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, pp. 85–86. Firenze University Press.

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: methodology and finding. *International Journal of Corpus Linguistics* 22(4), 457–489.

Shepherd, T. (2009). Corpora de aprendiz de língua estrangeira: um estudo de n-gramas. *Veredas* 2, 100–116.

Shepherd, T. M. G., Sardinha, T. B., Veirano, M. (Eds.). (2012). *Caminhos da Linguística de Corpus*. Campinas: Mercado das Letras.

Trinity College London. (2010). *Graded examinations in spoken English—Syllabus from 1 February 2010*. Trinity College London.

More on criteria for measuring text complexity

Irina Panteleeva, Olga Lyashevskaya, Olga Vinogradova
National Research University Higher School of Economics in Moscow
impanteleyeva@gmail.com, olesar@yandex.ru, olgavinogr@gmail.com

The study of the criteria for measuring the complexity of academic texts is considered an integral part of the research in the field of language learning. The term 'text complexity' refers to sophistication and variability of the text, written or oral. One of the main questions in this area concerns the choice of text complexity metrics as reliable indices for reflecting language proficiency. All criteria could be divided into groups that analyse different levels of language, namely, morphological, lexical, syntactic and discourse-oriented.

The criteria of morphological complexity include derivational and inflectional features; the former inspect the use of prefixes and suffixes (Bauer & Nation, 1993), while the latter look at inflections, more often - verbal inflection (Bulté & Housen, 2012). Lexical complexity reflects density, sophistication and diversity of the text, cf. TTR and Verb Variation metric suggested in (Lu, 2012; Torruella & Capsada, 2013; among others), as well as academic word lists (Xue & Nation, 1989). Discourse criteria are based on the number of connecting phrases, for example, linking words, discourse-organizing nouns (Tåqvist, 2016), and n-grams that are involved in text organization. The syntactic metrics refer to how complicated the structures of the text are, for example, the depth of a syntactic tree reflects the number of dependent clauses (Lu, 2012).

Moreover, recent years have seen significant advances in applying multi-feature regression, classification and multi-dimensional scaling techniques to disentangle the possible impact of various coarse-grained and fine-grained text features on holistic human judgements of L2 writing proficiency. In (Batinic et al., 2017, Grigonytė et al., 2018, among others) authors proposed Machine Learning models such as kNN, SVM, Naive Bayes, Logistic Regression and Decision Tree Classifier as a way to check how well the selected criteria distinguish low-level and high-level essays and to identify the best. In (Crossley et al., 2019), a multi-dimensional scaling was used to examine the style of students' academic essays.

In our study, we use classification models to identify criteria that can be used in evaluating essays written by Russian learners of English. All experiments were conducted on the basis of publicly available corpus REALEC (Vinogradova et al., 2017, http://realec.org). Our dataset consists of 3442 English examination essays written by Russian students. They were divided into two groups - "best" and "not best" (384 and 3058 respectively) - in accordance with the scores assigned by experts. We trained several machine learning models using scikit-learn package (Hackeling, 2017) taking into account 65 features (lexical, morphological, syntactic, and discourse). In the preprocessing stage UDPipe dependency parser (Straka, 2017) was applied.

The best result in 10-fold cross-validation was demonstrated by the Logistic Regression model: precision 0.86, recall 0.76, f1-score 0.81. Using the Random Forest Classifier, we ranged all criteria basing on their importance. The main idea was to estimate how well a criterion classifies the sample into classes. Figure 1 illustrates 11 most important features:

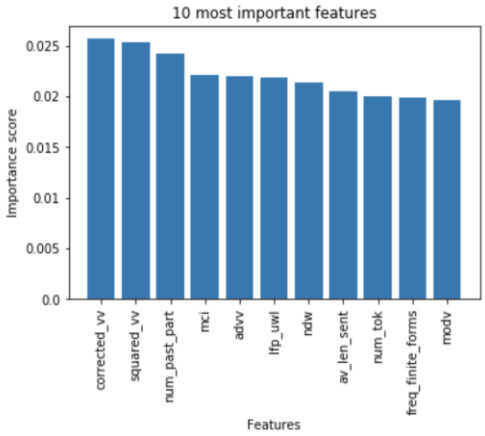


Figure 1. 11 most important features (left to right): corrected verb variation, squared verb variation, number of past participles, average inflectional diversity, adverbial variation, Lexical Frequency Profile (University Word

List), number of lemmas, average sentence length, number of tokens, frequency of tense (finite) forms, modifier variation.

Nine out of 11 most important features belong to the lexical and morphological levels of language. For example, the more different verb lemmas are used in the text, the higher the likelihood of getting a high mark (vv), or the more modifiers are included in the essay, the better it is (modv).

We argue that criteria that provide the absolute value (e.g. the number of past participles in the text) classify the essays better than relative frequencies. It is likely that when evaluating holistically, people often take into account the total number of different complex traits in the text.

References

Batinic, D., Birzer, S., & Zinsmeister, H. (2017). Automatic Classification of Russian text for didactic purposes. *Proceedings of Corpus Linguistics* 2017. St. Petersburg, 9–15.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography* 6(4), 253–279.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, I. Vedder. (Ed.) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. *Language Learning & Language Teaching* 32. Amsterdam: John Benjamins, 21–46.

Crossley, S. A., Kyle, K., & Römer, U. (2019). Examining lexical and cohesion differences in discipline specific writing using MDA. In *Multi-dimensional Analysis: Research Methods and Current Issues*, Bloomsbury Academic, 189–216.

Grigonytė, G., Kovalevskaitė, J., & Rimkutė, E. (2018). Linguistically-Motivated Automatic Classification of Lithuanian Texts for Didactic Purposes. In *Human Language Technologies – The Baltic Perspective*, Proceedings of Baltic HLT 8, 38–46.

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *Modern Language Journal* 96(2), 190–208.

Straka, M. (2017). CoNLL 2017 shared task – UDPipe baseline models and supplementary materials. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Tåqvist, M.–K. (2016). *"Another thing": Discourse-organising nouns in advanced learner English*. (Doctoral dissertation), Karlstad University, Karlstad.

Torruella, J., Capsada, R. (2013). Lexical Statistics and Typological Structures: A Measure of Lexical Richness. In *Proceedings of CILC 2013. Procedia – Social and Behavioral Sciences* 95, 25, October 2013, 447–454.

Vinogradova, O. I., Lyashevskaya, O. N., & Panteleeva, I. M. (2017). Multi-level Student Essay Feedback in a Learner Corpus. In *Proceedings of Dialogue 2017*, Moscow.

Xue, G., & Nation, I. S P. (1984). A university word list. *Language Learning and Communication* 3/2, 215–229.

Communication breakdowns in intercultural communication and implications for the foreign language classroom

Michał B. Paradowski, Elżbieta Pawlas
University of Warsaw
m.b.paradowski@uw.edu.pl, elzbieta.pawlas@student.uw.edu.pl

Communication breakdowns have deservedly been attracting the interest of researchers, as they constitute important factors influencing the process of linguistic interaction and language acquisition. Not only do they affect the process of communication *per se*, but also have other, often serious, consequences. Particular interest should be accorded to the process of achieving—and failing to achieve—understanding when English is spoken as a vehicular language.

We will present the results of the first comprehensive analysis of the complete conversations subcorpus of the Vienna-Oxford International Corpus of English (VOICE), focusing on the i) possible causes of communication breakdowns, and ii) strategies employed by speakers in order to both prevent and overcome such failures. We categorise and show the distribution of the sources of 122 detected breakdowns as well as the compensatory strategies employed by interlocutors to successfully avert and solve communication problems.

The VOICE contains transcripts representing naturally-occurring face-to-face ELF interactions, whose participants come from different cultural and linguistic backgrounds. For the purpose of this study we selected all speech events tagged as 'conversation' (the analyses hence do not consider other speech event types, such as seminar discussions or interviews). After the selection, the reduced corpus comprised 36 speech events (158,071 words), corresponding to approx. 15 hours of spoken interactions. The speakers come from different, mostly European, countries, have different L1s and occupations. Their ages vary from 17 to more than 50. The relations between them are fairly symmetrical. The conversations belong to different thematic domains: 21 of them are tagged as 'leisure', 7 as 'professional research/science', 4 as 'educational', 3 as 'professional organizational' and 1 as 'professional business'.

The entire material was first analysed in search of characteristic features and communication breakdowns. These were then analysed again in detail with regard to what caused the failures and how they were resolved, or at least how the speakers tried to resolve them. The list of identified causes covered unintelligible speech, simultaneous talk, overlap, pause, lack of topic shift signalling, lack of explicitness, wrong anaphoric or deictic reference reconstruction, faulty semantic reconstruction, code-switching, lack of shared cultural/world knowledge, misinterpretation of proper names, lack of shared lexical knowledge, wrong use of an existing word, wrong word order/tenses, and wrong/unfulfilled listener presupposition. Similar causes and similar strategies were then grouped together and tallied. Finally, the remaining data were again scrutinised in search of preventative strategies. These included enhancing explicitness, paraphrase, repetition, metadiscursive devices, completion of earlier utterance, dividing utterance into smaller parts, requesting assistance from other interlocutors, translating code-switches into English, and code-switch into language other than English.

The paper concludes with pedagogical recommendations.

Key references

Ädel, A., & Mauranen, A. (2010). Metadiscourse: Diverse and divided perspectives. *The Nordic Journal of English Studies*, 9(2), 1–12.

Kaur, J. (2009). Pre-empting problems of understanding in English as a Lingua Franca. In A. Mauranen & E. Ranta (Eds.) *English as a Lingua Franca: Studies and Findings* (pp. 107–123). Newcastle upon Tyne: Cambridge Scholars.

Kurhila, S. (2003). *Co-constructing Understanding in Second Language Conversation*. Helsinki: Univ of Helsinki.

Liddicoat, A. J. (2011). *An Introduction to Conversation Analysis*. London: Continuum.

Mauranen, A. (2006). Signalling and preventing misunderstandings in English as a lingua franca communication. *International Journal of Sociology of Language* 177, 123–150.

Meierkord, C. (2000). Interpreting successful lingua franca interaction. An analysis of non-native-/non-native small talk conversations in English. *Linguistik Online* 5, 1/00.

Pitzl, M. B. (2005). Non-understanding in English as a lingua franca: Examples from a business context. *VIenna English Working PaperS*, 14(2), 50–71.

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca.* Oxford: Oxford University Press.

Syntactic complexity across text genres. Findings from a learner corpus of written Danish

Mikołaj Sobkowiak
Adam Mickiewicz University in Poznań
miksobko@amu.edu.pl

Studying the acquisition of Danish by Poles has never been as relevant as it is at present due to the growing numbers of young Polish adults learning Danish as a second or foreign language in and outside of Denmark. With the exception of a few investigations, however, the acquisition of Danish by Poles remains by and large unchartered territory. This gap can be filled through corpus-based acquisition studies.

Linguistic complexity is commonly viewed as "a valid and basic descriptor of L2 performance, as an indicator of proficiency and as an index of language development and progress" (Bulté & Housen, 2014:43). Its subtype, syntactic complexity, is often considered "an important measure of second language (L2) writing proficiency" (Kyle & Crossley, 2018:1).

Various sets of complexity measures have been observed to capture genre differences in L2 writing (e.g. Polio & Yoon, 2018). According to a number of recent studies, both topic and genre can have an impact on a learner text's syntactic complexity (Bulté & Housen, 2018:150). Following a pilot study, the author investigates the interplay between text genre/topic and syntactic complexity in short texts written by Polish learners of Danish. The overall goal is to explore the impact of learner and task variables on texts produced by adult language learners.

The study is a cross-sectional one as the analyzed material consists of exam papers written by 6 different classes of students of Danish philology after having learned Danish for one (academic) year. The authors of the analyzed texts constitute a fairly homogeneous group as far as learner variables are concerned and the text sets written by the respective classes resemble one another in terms of most task-related variables as well (cf. Granger, 2008:264). What they differ in, however, are variables such as text type/genre as the analyzed material comprises creative narratives (n=30), factual narratives (n=11), short expository essays (n=11) and expository/argumentative essays on more abstract subjects (n=20).

All the texts have been digitalized and tagged using software developed specifically for this research. The set of complexity measures operationalized for this investigation is based on the one used by Bulté & Housen (2018) and supplemented by measures particularly relevant for Danish.

The preliminary results suggest that the analyzed texts do differ in terms of syntactic complexity based on topic and genre, some of the differences being statistically significant. This is despite the fact that the average proficiency level in the investigated population was relatively low (approximately A2/B1 according to CEFR), which as such leaves less room for potential variation.

References

Bulté, A., & Housen, B. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26, 42–65. DOI: 10.1016/j.jslw.2014.09.005.

Bulté, A., & Housen, B. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics* 28, 147–164. DOI: 10.1111/ijal.12196

Granger, S. (2008). Learner Corpora. In A. Lüdeling, M. Kytö (Eds.). *Corpus Linguistics. An International Handbook* (p. 259–275). Walter de Gruyter: Berlin and New York.

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal* 102, 333–349. doi:10.1111/modl.12468

Polio, C., & Yoon, H-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*. 2018; 28, 165–188. DOI: 10.1111/ijal.12200

What's in a comma: Corpus study of punctuation errors made by Russian Learners of English

Olga Vinogradova, Ksenia Pospelova, Anna Viklova, Veronika Smilga
Research University Higher School of Economics in Moscow
olgavinogr@gmail.com, pospelova1990@gmail.com, annakisseleva@gmail.com, smilgaveronika@gmail.com

The research was carried out over examination essays written in English by university students with Russian as L1. The essays are from REALEC – a learner corpus (http://realec.org) in which written texts are manually annotated for errors.

We extracted over 6,000 sentences with commas, 1407 with colon, 540 with hyphen, and 287 with semicolon. 1595 sentences were randomly chosen as the experimental dataset. Four expert annotators (English professors) approved of the use of punctuation in 1062 cases and identified punctuation errors in 491 sentences - 31% (cf. Company 2012; Alamin, Ahmed 2016). It is also almost twice the number of errors spotted by student annotators in the same sentences (256). It shows that not only authors of essays at the level of B2 to C1 have difficulty with English punctuation, but also that student annotators with reference materials at their disposal often fail to see a punctuation error (cf. Van Rooy 2015).

The highest numbers of errors were attested for the following 7 classes (listed in diminishing order of occurrences, the latter given as a figure in bold after the number of the class)[12]:

1. **104** A redundant comma after the main clause and before the subordinate clause when the latter is introduced with the conjunction other than the one that requires the left boundary (cf. rule [3] of the asymmetry of the marking boundaries on p. 1736):

(1) Moreover, according to the research made by american scientists in 2009, educational process is rather more productive, when people of different genders and even cultures study all together.

2. **81** Confusion of punctuation in different types of relative clauses (cf. comment #14 on p. 1745):

(2) For example, you won't pay much money to go to a concert of a musician, whose art you don't know really well, but…

3. **44** No delimiting commas (left or right or neither) for parenthetical construction (cf. p. 1744):

(3) All in all the graph shows, that the amount of people aged 65 and over is not static between 1940 and 2040…

4**. 38** Use of a colon instead of a dash or a comma (cf. pp. 1738 & 1741):

(5) From 1940 to 1960 the number rose steadily in both countries: in the USA and in Sweden, while…

5. **33** No delimiting comma between the two clauses with *and* as coordinator (cf. p. 1740 about possible misreading and the higher probability of this comma than in subclausal coordination with *and*):

(4) First of all, in modern world every person should have the same rights as others and any discrimination is prohibited.

6**. 16** A redundant comma before participial construction after the head noun:

(6) As for me, I fully agree with the first group and would like to provide several arguments, supporting my point of view.

7. **14** Insufficient punctuation in cases of supplementation (on complementation cases cf. p. 1740-1741):

(7) That's why for such a long time Russia, country with a biggest size of fields didn't do planting …

Russian learners of English are rarely taught punctuation conventions systematically, and this is consistent with reports on learners of English with other native languages (cf. Jiu, Yan 2016, Hirvela, Nussbaum & Pierson 2012; Moore 2016). However, unlike the situation in many other countries, punctuation is very rigidly taught in classes of native Russian language, and we can suspect interference with Russian punctuation conventions in the errors of classes 1, 2, and 6, because ALL subordinate clauses, ALL relative clauses and ALL participial constructions in any position have delimiting commas in Russian (Proshina, Eddy 2016). However, other types of errors cannot be explained by Russian punctuation model.

[12] For terminology and as a source of variation in punctuation we used (Huddleston and Pullum, 2002), and references to the specific pages in this book are given in the description of classes.

References

Alamin, A., & Ahmed, S. (2012). Syntactical and Punctuation Errors: An Analysis of Technical Writing of University Students Science College, Taif University, KSA. *English Language Teaching* 5. 10.5539/elt.v5n5p2.

Company, M. T. (2012). Error Frequencies Among ESL Writers: A Resource Guide. *All Theses and Dissertations* 3420. https://scholarsarchive.byu.edu/etd/3420

Jiu-Gen Xiao, Yan-Min Li (2016). Punctuation in language art – referring to the absence of punctuation in Chinese teaching. In *3d International Conference on Applied Social Science Research* (ICASSR 2015), Atlantis Press. doi:10.2991/icassr-15.2016.129

Hirvela, A., Nussbaum, A. & Pierson, H. (2012). ESL students' attitudes toward punctuation. System, 40(1).

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

Moore, N. (2016). What's the point? the role of punctuation in realising information structure in written English. *Functional Linguistics*, 3(1).

Proshina, Z., Eddy, A. (eds.) (2016). Russian English. History, Functions, and Features. Cambridge : Cambridge University Press

Van Rooy, B. (2015). Annotating learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (Cambridge Handbooks in Language and Linguistics, pp. 79–106). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.005

Discourse structure in German argumentative essays:
a comparison of L1 German and Chinese learners of German

Shujun Wan, Anke Lüdeling
Humboldt-Universität zu Berlin
shujun.wan@hu-berlin.de, anke.luedeling@hu-berlin.de

Writing is constantly identified as a weak link (Shi and Shang, 2011) for learners of German as a Foreign Language (GFL) with L1 Chinese. On the one hand, rhetoric is language and culture specific (Kaplan, 1967); on the other hand, the knowledge of discourse structure often remains overlooked in German teaching in China (Qi, 2011). Previous studies in area of rhetoric mostly regard Chinese learners of English as a Foreign/Second Language (EFL/ESL) as the research object (Kirkpatrick, 1997; Chen, 2014). They found that English compositions by Chinese EFL/ESL learners have consistently shown evidence of using the traditional Chinese text structures (in particular *qǐ-chéng-zhuǎn-hé* and *bā-gǔ-wén*) (Chien, 2007).

Inspired by those observations, the present study aims to investigate the rhetorical structure in argumentative essays written by Chinese GFL learners and by German native speakers. Our specific research questions pertain to the following aspects: (1) How does the frequency of using certain discourse relations differ between the two groups? (2) What are the similarities and differences between the two groups in terms of argumentation style? By researching these questions, we also draw a parallel comparison, aiming to see whether there are similar findings between Chinese GFL learners and Chinese EFL/ESL learners. The answer to these questions could help to shed light on (a) the transfer of discourse structure and (b) the reference of teaching methods.

Annotation on the discourse level is an extremely time and resource consuming task due to the inevitable subjectivity, complexity and ambiguity of discourse (Gries and Berez, 2017). In order to study these questions, we built the *RST German Learner Treebank*, which contains 20 argumentative essays written by advanced Chinese GFL learners as well as 20 by native speakers. All the 40 essays are on the same topic. According to *Rhetorical Structure Theory* (Mann and Thompson, 1988), the argumentative essays are annotated by two professional linguists based on the guidelines of the Potsdam Commentary Corpus (Stede, 2017). The inter-annotator-agreement was automatically evaluated by using *RST-Tace* (Wan et al., 2019). With a kappa value of 0.707, the annotation indicates a substantial agreement.

In response to the research questions above, the results of this study are as follows:

(1) Chinese GFL learners use significantly less *list* and *contrast* but use more *elaboration* and *evaluation* than German native speakers. An explanation could be that Chinese GFL learners prefer to use fewer arguments but explain more about each argument. They tend to classify their arguments into three or four main points, whereas German native speaker are inclined to list their arguments separately without categorization. As the traditional Chinese text structure *bā-gǔ-wén* consists of four points, we consider that this writing style by Chinese GFL learners is likely influenced by their L1.

(2) German native speakers tend to express their opinions directly at the beginning of the essay, while Chinese GFL learners are inclined to begin the essay with common knowledge such as the economic development of China and rather point out their views in the end. This finding could also be confirmed by the extensive use of the relation *background* by Chinese GFL learners. Notably, Chinese EFL learners also predominantly make reference to common world knowledge when introducing topics (Callies, 2015). Apart from these observations, we also found evidence for Chinese GFL learners' preference for citing sayings, proverbs as well as set phrases to support their arguments, which is not considered convincing in German rhetoric convention. This may result in a lower quality of German writing for Chinese GFL learners.

References

Callies, M. (2015). Towards a function-driven approach to annotating learner corpora: The case of topic marking. In Learner Corpus Research (pp. 8–11).

Chen, P. (2014). The Comparison of Intermediate and Advanced Chinese Learners' use of English Adverbial Connectors in Academic Writing. International Journal on Studies in English Language and Literature, 2(8), 85–92.

Chien, S. (2007). The role of Chinese EFL learners' rhetorical strategy use in relation to their achievement in English writing. English Teaching: Practice and Critique, 6(1), 132–150.

Gries, S. T., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus linguistics. In Gries, S. Th. And Berez, A. Handbook of Linguistic Annotation (pp. 379–409). Dortrecht: Springer. https://doi.org/10.1007/978-94-024-0881-2_15

Kaplan, R. B. (1967). Seeing the world through language-colored glasses. TESOL Journal, 1(4), 10-16.

Kirkpatrick, A. (1997). Traditional Chinese Text Structures and Their Influence on the Writing in Chinese and English of Contemporary Mainland Chinese Students, 6(3), 223–244.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Qi, D. (2011). 中国德语专业大学生论证型篇章中的连贯关系 *(Coherence Relations in argumentative essays of Chinese students major in German studies)*. PhD thesis. Beijing foreign studies University.

Shi, X. and Shang, X. (2011). 大学德语听说读写强化训练丛书-写作 *(A training book for collage German learning - Writing)*. Tongji University Press.

Stede, M. (2017). Annotation Guidelines for Rhetorical Structure. https://www.sfu.ca/~mtaboada/docs/research/RST_Annotation_Guidelines.pdf

Wan, S., Kutschbach, T., Lüdeling, A., & Stede, M. (2019). A tool for automatic comparison and evaluation of RST trees. Proceedings of 7th Workshop on Rhetorical Structure Theory and Related Formalisms.

Lexical Features in Argumentative Writing across English Writers from Different Language Backgrounds

Xiaoli Yu
Middle East Technical University
xiaoli@metu.edu.tr

This corpus-based research analysed three lexical features (lexical diversity, lexical sophistication, and cohesion) in English argumentative writing and examined the potential differences in lexical performance 1) between native and nonnative English writers and 2) across all writers from seven language backgrounds.

Two major research questions guided the analyses in the current study:

1. Are there significant differences in lexical features between native and nonnative argumentative English writing, as measured by lexical diversity, lexical sophistication, and cohesion?
2. Are there significant differences in lexical features, as measured by lexical diversity, lexical sophistication, and cohesion, in argumentative English writing across all writers from various mother tongue backgrounds?

The target population of nonnative English writers was advanced English learners in non-English-speaking countries; the referential native speaking population was native English-speaking university students. The learner English corpora were six subcorpora selected from the International Corpus of Learner English v2 (ICLE v2), including the mother tongues of Chinese, German, Japanese, Russian, Spanish, and Turkish. The native-speaking corpus was the Louvain Corpus of Native English Essays (LOCNESS), which included essays written by both British and American undergraduate students. For the seven selected subcorpora, 100 argumentative essays were randomly selected from each subcorpus. A total of 700 texts have been analysed in the study. The total tokens were 424,363 words.

The findings revealed that nonnative English writers demonstrated significantly lower performance in lexical sophistication than did native English writers. The comparison between writers from different language backgrounds suggested statistically significant differences in all three aspects of lexical features. German, Japanese, and Turkish writers, in particular, revealed potential needs in obtaining supports regarding lexical diversity and lexical sophistication.

Pedagogical implications for vocabulary instruction in argumentative writing for nonnative writers are introduced, such as emphasizing the mastery of academic, low-frequency, and discipline-specific vocabulary. Additionally, improving non-native writers' vocabulary size and lexical diversity can offer these learners more options to build cohesion in academic writing at a deeper level. The results of this study also highlight the wide but often under-considered variability within any group of writers as learner differences come into play, thereby downplaying the idea that writers of any given group tend to perform homogeneously. Instructors should acknowledge the unique writing characteristics of different non-native writers and their varied learner needs. Thus, targeted instruction is essential to provide effective enhancement to non-native English writers' lexical performance in academic writing.

ICLEv3: An extended web-based version of the *International Corpus of Learner English*

Sylviane Granger, Maïté Dupont, Fanny Meunier, Magali Paquot
University of Louvain
sylviane.granger@uclouvain.be, maite.dupont@uclouvain.be, fanny.meunier@uclouvain.be,
magali.paquot@uclouvain.be

This software demonstration proposal aims to present the new features of the third version of the *International Corpus of Learner English* (ICLE). The ICLE is a corpus of argumentative essays written by higher intermediate to advanced learners of English as a foreign language from a wide range of language backgrounds. The corpus collection was initiated by the Centre for English Corpus Linguistics of the University of Louvain (UCLouvain). The first version, which appeared in 2002, contained 2.5 million words produced by learners from 11 mother tongue backgrounds. The second version, which appeared seven years later, was larger both in terms of words (3.7 million) and language backgrounds (16). Both versions have been used extensively as a basis for a large body of studies in second language acquisition, foreign language teaching and testing, and natural language processing.

Ten years on we are now about to release the third version of the corpus which will differ from the preceding versions in two major ways. First, it will be even larger than the preceding versions, both in number of words (c. 5 million) and mother tongue backgrounds (26). In addition to those already represented in ICLEv2 (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, Tswana), ICLEv3 will also contain the following L1s: Brazilian Portuguese, Greek, Hungarian, Korean, Lithuanian, Macedonian, Persian, Serbian, Urdu and Punjabi. The second major difference is that access to the corpus will now be exclusively web-based. It will be stored on the Corpor@ platform, designed by the natural language processing centre (Centre de Traitement Automatique du Langage - CENTAL), which will host all the corpora collected at UCLouvain. The new web-based interface allows both for easier and more flexible access and for regular inclusion of new subcorpora as they are completed. In our demo we will focus on the following functionalities of the web interface:

– newly enhanced compilation of sub-corpora on the basis of learner and task variables (e.g. mother tongue of the speakers, age, number of years of English, time spent in an English-speaking country, text length, topic, text type);

– improved sub-corpora download facilities: after the compilation of a sub-corpus, the generated zip file includes (1) all learner texts in a single .txt file, (2) a directory with all learner texts stored as separate files and (3) a metadata file available in .xls or .csv;

– simple and advanced search (search for word forms, lemmas, CLAWS POS-tags and simplified tags);

– breakdown of the search strings in terms of the many demographic and task variables recorded in the ICLE database frequency data;

– new concordance export options facilitating further analysis and statistical treatment of the data: concordance lines can be saved in .xls or .csv files together with their corresponding learner and task-related metadata.

With the addition of some new features, such as separate files for each learner text, case-by-variable-format to export concordance lines together with learner and text-related variables, we hope to help answer the call for more attention to intra-variability and individual differences in learner corpora.

The demo will feature concrete illustrations of all these functionalities.

References

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

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

Granger, S., Dupont, M., Meunier, F. & Paquot, M. (forthcoming). *International Corpus of Learner English Version 3 – Handbook*. Louvain-la-Neuve: Presses universitaires de Louvain.

SVALA: an Annotation tool for Learner Corpora generating word-aligned parallel texts

Elena Volodina[1], Arild Matsson[1], Dan Rosén[1], Mats Wirén[2]
[1]University of Gothenburg
[2]Stockholm University
elena.volodina@gu.se

Learner corpora are actively used for research on Language Acquisition and in Learner Corpus Research (LCR). The data is, however, very expensive to collect and manually annotate, and includes steps like anonymization, normalization, error annotation, linguistic annotation. In the past, projects often re-used tools from a number of different projects for the above steps. As a result, various input and output formats between the tools needed to be converted, which increased the complexity of the task.

In the present project, we are developing a tool that handles all of the above-mentioned steps in one environment maintaining a stable interpretable format between the steps. A distinguishing feature of the tool is that users work in a usual environment (plain text) while the tool visualizes all performed edits via a graph that links an original learner text with an edited one, token by token.

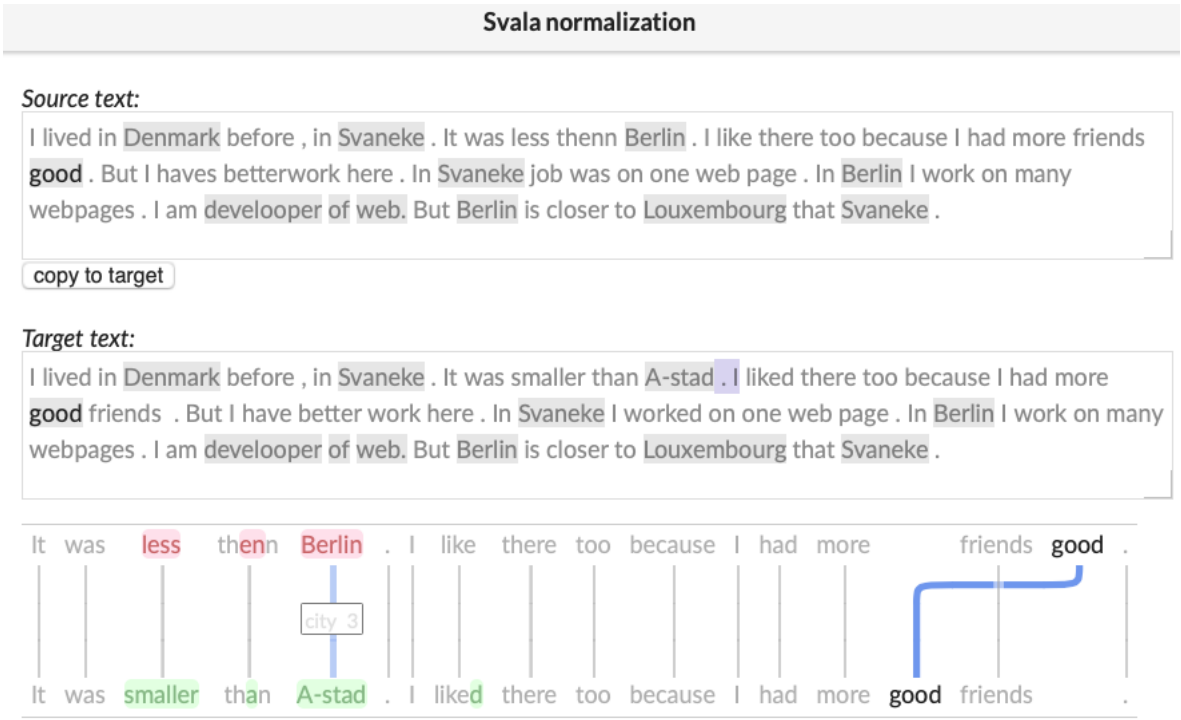


Figure 1. SVALA normalization view

Anonymization is a preprocessing step where categories are assigned to sensitive segments in the same way as correction labels, whereas the corresponding target text segments are automatically pseudonymized, e.g. *Berlin* vs *A-stad* (Fig.1).

During normalization, a user is working directly in the field "target text" editing a copy of an original text (Fig.1). All the while, the tokens in the original texts are automatically aligned with the tokens in the target text, a graph is incrementally updated and the result is visualized in a parallel view.

Correction labels are assigned to the links in the graph between the tokens that display difference between the original and the edited text, e.g. *thenn* and *than* (Fig. 2) and characterize the nature of the difference.

Alignments (i.e. links) in the graph are built automatically and can be manually corrected. It may especially be necessary when it comes to the word order changes (see *friends good* vs *good friends* in Fig. 2).

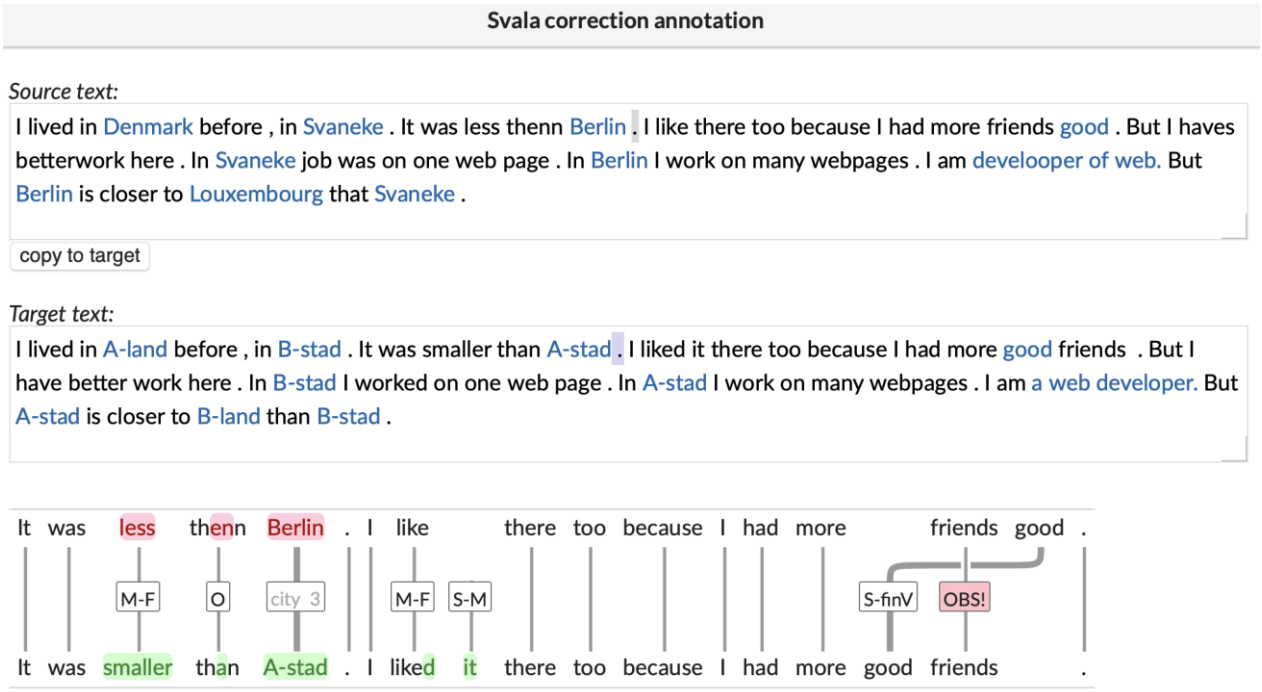


Figure 2. SVALA correction annotation view

The choice has been made in the project to support JSON format representation of the data as an alternative to a more universally accepted XML, since JSON ensures a relative light-weightedness of the tool and supports structuring data in an easily accessible way. Three data objects are created (Fig. 3) (see also Rosén et al. 2018):

1 to handle the original text

2. to handle the target text

3. to describe edges (links) with attached correction labels

Conversion to xml TEI format is a trivial step and in case of interest to SVALA outside the project, we can consider adding it to a format conversion tool, such as Pepper (Zipster & Romary 2010).

Adapting SVALA format to existing search environments may present challenges. We foresee wasting some of the information encoded in the present format when flattening it to a less expressive XML format. Those challenges and potential other consequences will be evaluated and addressed in the not so distant future.

Another choice in our project – a rather unusual one for LCR projects except a few cases (e.g. Boyd et al. 2014, Reznicek et al. 2012, Rosen et al. 2014) – is to separate normalization from correction labeling. Our belief is that rewriting a text to a more "target-language like" version is easier and more systematic when done in its entirety without being distracted by labeling the changes at the same time.

SVALA is a free software under the MIT license (https://github.com/spraakbanken/swell-editor).

```
{
  "source": [
    {"id": "s0", "text": "It "},
    {"id": "s1", "text": "was "},
    {"id": "s2", "text": "less "},
    {"id": "s3", "text": "thenn "},
    {"id": "s4", "text": "Berlin "},
    {"id": "s5", "text": ".  "}
  ],
  "target": [
    {"id": "t0", "text": "It "},
    {"id": "t1", "text": "was "},
    {"id": "t15", "text": "smaller "},
    {"id": "t19", "text": "than "},
    {"id": "t25", "text": "B-stad "},
    {"id": "t5", "text": ".  "}
  ],
  "edges": {
    "e-s4-t25": {
      "id": "e-s4-t25",
      "ids": ["s4", "t25"],
      "labels": ["1", "city"],
      "manual": true
    },
    "e-s0-t0": {
      "id": "e-s0-t0",
      "ids": ["s0", "t0"],
      "labels": [],
      "manual": false
    },
    "e-s1-t1": {
      "id": "e-s1-t1",
      "ids": ["s1", "t1"],
      "labels": [],
      "manual": false
    },
    "e-s2-t15": {
      "id": "e-s2-t15",
      "ids": ["s2", "t15"],
      "labels": ["L-W"],
      "manual": false
    },
    "e-s3-t19": {
      "id": "e-s3-t19",
      "ids": ["s3", "t19"],
      "labels": ["O"],
      "manual": false
    },
    "e-s5-t5": {
      "id": "e-s5-t5",
      "ids": ["s5", "t5"],
      "labels": [],
      "manual": false
    }
  }
}
```


Figure 3. SVALA format (*source*, *target* and *edges* objects in JSON format)

References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., Vettori, C. (2014). *The MERLIN corpus: Learner Language and the CEFR*. LREC 2014.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F. (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0*. Berlin: Humboldt-Universität zu Berlin.

Rosen, A., Hana, J., Stindlová, B., & Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* 48(1).

Rosén, D., Wirén, M., Volodina, E. (2018). *Error Coding of Second-Language Learner Texts Based on Mostly Automatic Alignment of Parallel Corpora*. Proceedings of CLARIN Annual Conference 2018.

Zipser, F., & Romary, L. (2010). *A model oriented approach to the mapping of annotation formats using standards*. Workshop on Language Resource & Language Technology Standards, LREC 2010.