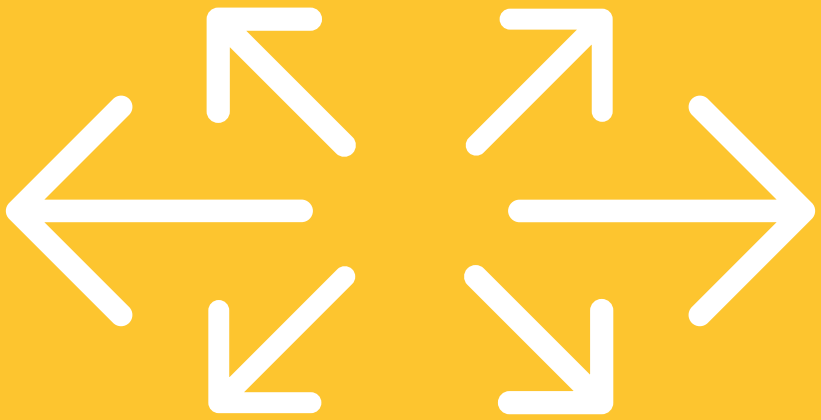


eurac
research

4th LEARNER CORPUS RESEARCH CONFERENCE

LCR 2017

Bolzano/Bozen
5-7 October 2017

Book of Abstracts

Sponsors

Città di Bolzano
Stadt Bozen

Contents

Welcome Message

Dear Participant,

On behalf of the organising committee, it is my pleasure to welcome you to the 4th Learner Corpus Research Conference (LCR 2017) themed “Widening the scope of Learner Corpus Research” at Eurac Research. LCR conferences, organised under the aegis of the Learner Corpus Association, bring together researchers and language teachers, software developers and others interested in Learner Corpus Research all over the world. This year LCR participants come from 23 countries!
In this booklet you will find the conference programme and social programme, together with some practical information that we hope you will find useful. The information desk will be open throughout the conference to assist you whenever required and inform you of any changes to the programme.
I wish you an inspiring conference and a good time in Bolzano/Bozen.

Andrea Abel
Chair
4th Learner Corpus Research Conference (LCR 2017)

Practical Information

Conference location
Eurac Research
Viale Druso 1
39100 Bolzano, Italy
http://lcr2017.eurac.edu/

Conference organization
Eurac Research Institute for Applied Linguistics
Eurac Research Meeting Management
39100 Bolzano/Italy
www.eurac.edu

On arrival
The registration desk and the info desk are located in the Eurac Research main entrance area.
▸ Opening hours: Mon-Sat: 8:00 – 18:00

Book/software exhibition
The book exhibition area is located in Foyer in lower floor.
The software exhibition will take place in Room 8.

Catering
The conference fee includes welcome reception on 5 October, coffee breaks, lunches on 5, 6 and 7 October and Conference Dinner at Maretsch Castle on 6 October.

Internet access
We offer open internet access to all participants. Please connect to the wireless network called “OpenAir”.
The password is wlan@eurac.edu

Transportation
For information about the local public transport as well as other trans- portation possibilities please contact the registration desk or go to
http://www.mobilcard.info/en/mobilcard.asp
Local taxi service: +39 0471 981 111

Water
We have good drinking water in Bolzano and South Tyrol. You can refill your bottles any time and nearly everywhere.

Smoking
Please note that smoking is not allowed at Eurac Reserach or in any public place, including restaurants and bars.

Exclusion of liability
The organizers decline all liability for any losses, accidents or damages that may occur for whatever reason to persons or goods.
Participants take part in the conference and in social events at their own responsibility.

Emergency numbers
Police: 113 or 112
Fire department: 115
Ambulance service: 118
Mountain rescue: 77171

Hospitals and doctors
Private clinic "Marienklinik"
Via Claudia de' Medici 2, 39100 Bolzano Tel: +39 0471 31 06 00

Bolzano General Hospital
Via Lorenz Böhler 5, 39100 Bolzano
Tel. +39 0471 908 111
Tel. +39 0471 908 330 (emergency calls)

Pharmacy
Farmacia Passazi Maria
Viale Druso 19, 39100 Bolzano
Tel. +39 0471 287559
Opening hours: Mon - Fri 8:30 – 12:30 and 15:00 – 19:00; Sat and Sun: closed

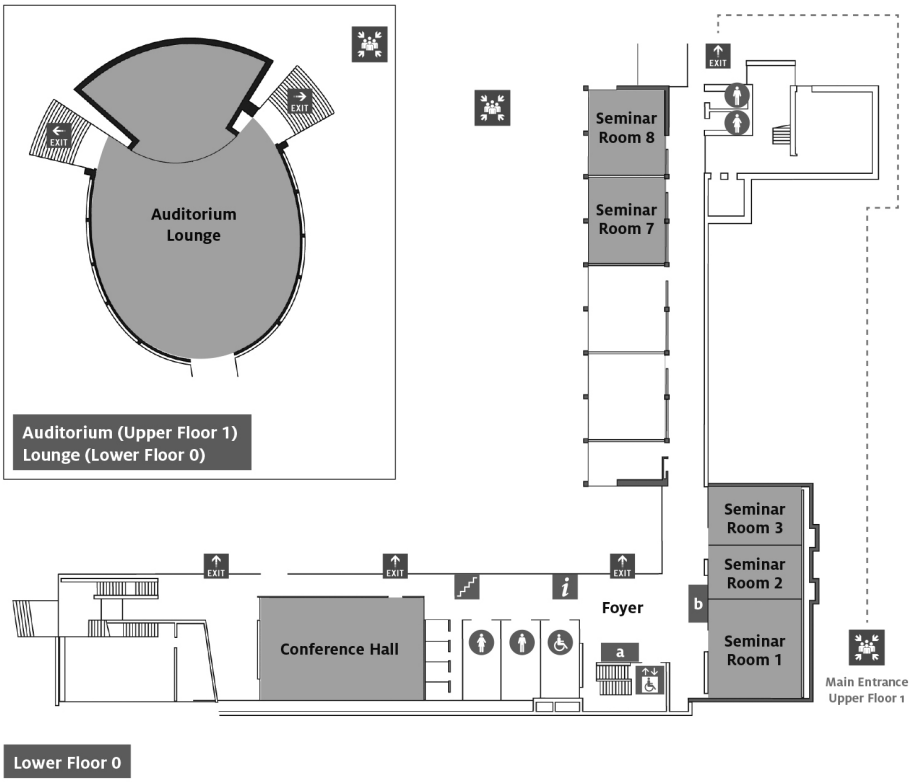
Bank and ATM
Südtiroler Sparkasse / Cassa di Risparmio di Bolzano Piazza Walther 26, 39100 Bolzano
+39 0471 231800

Post office
Piazza della Parrocchia 11, 39100 Bolzano
Opening hours: Mon – Fri 8:25 – 19:10, Sat 8:25 – 12:35

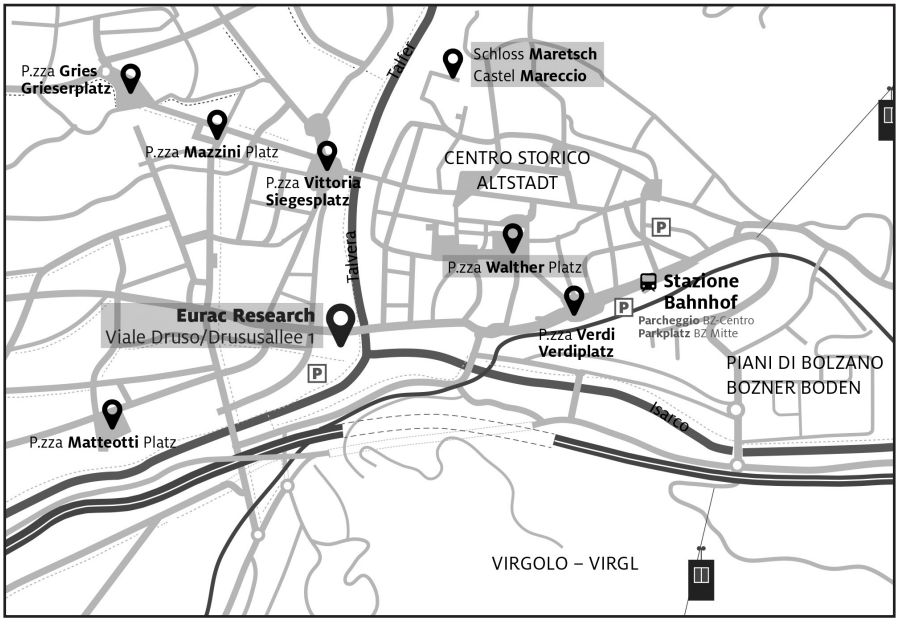
Downtown shops
Opening hours: Mon - Fri 9:30 – 19:00 (sometimes with a lunchtime break 13:00 – 15:00), Sat 9:30 – 19:00

Supermarket
Despar
Via Museo 2, 39100 Bolzano
Opening hours: Mon - Fri 8:30 – 19:15, Sat 8:00 – 18:00

Floor Plan


Auditorium
Lounge
EXIT
EXIT
Auditorium (Upper Floor 1)
Lounge (Lower Floor 0)
Seminar Room 8
Seminar Room 7
Seminar Room 3
Seminar Room 2
Seminar Room 1
Foyer
Conference Hall
a
b
i
Main Entrance
Upper Floor 1
Lower Floor 0

Social Programme



Guided City Tour, 6 October 2017
Meeting point:
Walther Square in front of the Tourist Information office
Time: 19:00

Meeting point for participants of the LCA General Assembly:
Eurac Research main entrance
Time: 19:00

Conference Dinner, 6 October 2017
Castle Maretsch/ Mareccio.
Via Claudia de' Medici 12
Time: 20:00
The location is at a 10 minutes' walk from Walther Square. We recommend to participate in the guided city tour which ends at Castle Maretsch/ Mareccio.

LCR 2017 Conference Schedule

time	04. Okt	05. Okt		06. Okt	07. Okt
08:00 - 08:30		Arrival + Registration			
08:30 - 09:00					
09:00 - 09:30		Opening LCR2017		Keynote 2	Keynote 3
09:30 - 10:00		Keynote 1			
10:00 - 10:30				Parallel Sessions - Full Papers	Parallel Sessions - Full Papers
10:30 - 11:00		Coffee		Coffee	Coffee
11:00 - 11:30		Parallel Sessions - Full Papers		Parallel Sessions - Full Papers	Parallel Sessions - Full Papers
11:30 - 12:00					
12:00 - 12:30					
12:30 - 13:00		Lunch		Lunch	Closing Session
13:00 - 13:30					Lunch + End of Conference
13:30 - 14:00					
14:00 - 14:30	Registration				
14:30 - 15:00		Parallel Sessions - Full Papers		Parallel Sessions - Full Papers	
15:00 - 15:30	Preconference Workshop				
15:30 - 16:00					
16:00 - 16:30		Coffee		Coffee	
16:30 - 17:00	Coffee	Parallel Sessions - WIP		Parallel Sessions - Full Papers	
17:00 - 17:30	Preconference Workshop				
17:30 - 18:00			Poster Session and Software Presentation	LCA Board Meeting	
18:00 - 18:30				LCA General Assembly	
18:30 - 19:00	Reception	LCA Board Meeting			
19:00 - 19:30				Social Walk	
19:30 - 20:00		Welcome Reception at Eurac Research			
20:00 - 20:30				Conference Dinner at Maretsch Castle	
20:30 - 21:00					
21:00 - 21:30					
21:30 - 22:00					
22:00 - 22:30					

Index of Presentations

Keynote Presentations

Full Paper Presentations

Keynote Presentations

Corpus research on the development of children's writing in L1 English

Philip Durrant
University of Exeter
P.L.Durrant@exeter.ac.uk

Since at least the 1940s, researchers have been interested in studying the development of first language writing through quantitative analysis of texts. In recent years, this has evolved into a small body of L1 written learner corpus research. The need for research of this kind has become pressing in England in recent years due to an increased curricular emphasis on explicit teaching of the linguistic features of writing. The current National Curriculum states that students should be taught to 'draft and write by: selecting appropriate grammar and vocabulary, understanding how such choices can change and enhance meaning' (DfE, 2013a) and specifies the ages at which children are expected master specific features of written grammar and vocabulary (DfE, 2013b). A convincing linguistic research base against which such policies can be evaluated does not yet, however, exist.

The Growth in Grammar project was developed in response to this need. It uses corpus methods to understand the linguistic development of English children's language throughout the course of their compulsory education. Our team at the University of Exeter is collecting educationally-authentic texts from children in schools across England from ages six to sixteen with the aims of understanding what distinguishes texts written at different ages, at different levels of attainment, and in different genres.

This presentation will describe what we think we know about how the language of children's writing develops based on the last six decades of research and discuss the Growth in Grammar project, focusing especially on methodological issues involved in creating and analysing a child learner corpus and on what initial results are telling us about written language development.

Quantitative methods in Learner Corpus Research: Goals and directions for the future and why some excuses don't count.

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Over the last 15-20 years, learner corpus research (LCR) has evolved from very small kind of niche discipline to a much larger, more visible, and more diversifed field. This evolution was fostered by an increasing number of different kinds of corpora of learner language, containing more and more diverse data, metadata, and annotation. This increase in data is welcome in all the obvious ways, but also raises the stakes when it comes to deciding how to analyze them: Seemingly trivially, "more data" is only better if one knows what to do with them. And yes, LCR has evolved: While even just 5-10 years ago, LCR studies were mostly just concerned with over-/underuse statistics in the form of normalized frequencies and chi-squared/log-likelihood statistics, studies with more sophisticated statistical analyses are now more frequently found. However, while linguistics as well as corpus linguistics, the larger disciplines of which LCR is a part, are now routinely employing sophisticated statistical methods, LCR appears to up the ante at a slower pace than is good for the discipline. With this talk, I will try to convince the audience that it is time to leave what still seems to be the current standard behind and move on to something more powerful than that. Specifically, I will make the case in point that much of current LCR - in fact most over-/underuse studies and most studies involving association measures - should in fact be reconceptualized as specific applications of regression modeling. I will try to show that regression modeling

- allows to analyze general over-/underuse both more legitimately and more insightfully than traditional methods;
- allows for a more informative and principled way of assocation (collocation/colligation/collostruction) in learner corpus data;
- affords the analyst to take many facets of LCR data into consideration that much traditional work neglects: individual variation, repeated measurements, corpus structure/homogeneity, and task effects.

To illustrate these points, I will end by discussing three case studies - two that involve (partial) reanalyses of published data and one that tries to bring nearly all of these aspects together in a single application.

Learner Corpus Research and the acquisition of Italian as a second language: the case of the Longitudinal Corpus of Chinese Learners of Italian (LoCCLI)

Stefania Spina
Università per Stranieri di Perugia
stefania.spina@unistrapg.it

According to the most recent data available from the Italian Ministry of Foreign Affairs (Ministero degli Affari Esteri, 2016), more than 2,200,000 million people were studying Italian as a foreign language in 2015 all over the world, and more than 42,000 were studying it as an L2 in Italy in the same year.
Italian can thus be considered, after English, French and Spanish, one of the most studied languages by second and foreign language learners.
Starting from the second half of the 1980s, a strong tradition of research has consistently developed in the area of the acquisition of Italian as an L2, producing a large body of studies, mainly focused on the acquisition of grammar and syntax (e.g. Giacalone Ramat, 1988; 2003).
In more recent years, a new and growing interest in learner corpus research (LCR) has emerged internationally, where English is by far the most widely investigated language. Although LCR shares with second language acquisition research "the objective of gaining a better understanding of the mechanisms of foreign or second language acquisition" (Granger, Gilquin & Meunier, 2015, p.3), it has its own, unique characteristics, such as its strong applied orientation, and, most importantly, its systematic use of large collections of digital learner data.
Despite the vast body of research produced in second language acquisition, Italian is currently a largely underrepresented language in the area of LCR, in terms of available learner corpora, application of quantitative and statistical methodologies to learner data, learner corpus design and annotation, NLP resources applied to learner language research, and language teaching applications stemming from LCR.
The aim of this presentation is to provide a comprehensive picture of the current state of LCR on Italian, describing corpora, resources, applications and the most represented areas of research. At the same time, it will try to outline its future prospects and the main challenges it will have to face, in the attempt to contribute to a deeper understanding of the acquisition of Italian as a second language.

References

Giacalone Ramat, A. (1988). L'italiano tra le altre lingue. Strategie di acquisizione. Bologna: Il Mulino.
Giacalone Ramat, A., ed., (2003). Verso l'italiano. Percorsi e strategie di acquisizione. Roma: Carocci.

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In Granger, S., Gilquin, G., & Meunier, F., Cambridge Handbook of Learner Corpus Research. Cambridge: CUP.
Ministero degli Affari Esteri e della Cooperazione Internazionale, Italiano Lingua Viva.
Stati Generali della Lingua Italiana nel Mondo, Firenze, 17-18 October 2016 (http://www.esteri.it/mae/resource/doc/2016/10/libro_bianco_stati_generali_2016.pdf).

Full Paper Presentations

Investigating frequency effects in learner corpus and experimental data: the case of the English catenative verb construction

Lina Baldus
University of Trier
baldus@uni-trier.de

Usage-based theories of second language acquisition claim that frequency has a considerable impact on our knowledge of language and is thus stated to be a "key determinant of acquisition" (Ellis 2002:144; cf. Bybee 2007; Goldberg 2006; Madlener 2015). Nevertheless, there has been only little research on the question which component parts of linguistic input need to be experienced with sufficient frequency by learners in order to form a native-like schema representation for a certain construction. This study addresses this issue with the help of mixed-effects models of learner corpus data and experimental data using the catenative verb construction as a testbed phenomenon.
The catenative verb construction is comprised of a so-called 'catenative verb' and a 'catenative complement', which is a non-finite subordinate clause functioning as a complement of the catenative verb. It prototypically occurs in the form of a to-infinitival complement as in (1) or a gerund-participial complement (cf. Huddleston & Pullum 2002) as in (2) below.

(1) She [decided to go to the cinema].
(2) His brother [finished reading his novel].

This construction is especially interesting because previous research (Gries & Wulff 2009; Martinez-Garcia & Wulff 2012) showed that foreign learners of English make choices of the catenative complement which are different from those of native speakers. This study explores how different frequency-based variables affect this variation, with the goal of finding out what component parts of the construction need to be experienced with sufficient frequency for the learner to build a native-like schema for this construction.
In order to address these questions, a pseudo-longitudinal corpus study with language data from German learners of different proficiency levels (A1-C2) was carried out using the *EF-Cambridge Open Language Database* (EFCAMDAT, Geertzen et al. 2013). In addition, two experimental studies, a sentence completion task and an acceptability judgment task with advanced German learners of English (C1 level), were conducted. In all cases, the focus was on a selection of verbs which are distinct for one of the two types of catenative complements mentioned above and which occur with different frequencies in the catenative verb construction, both determined on the basis of the *British National Corpus*. Each study was analysed with the help of a mixed-effects model (cf. Baayen et al. 2008) with the frequency of the matrix verb (i.e. the catenative verb in all of its different uses) and the frequency of the verb together with its distinct catenative complement among others as fixed effects as well as the participants as a random effect, in order to see to

what extent the different factors had an impact on the dependent variable, namely the type of catenative complement and whether this choice was target-like or not.
While in the experimental studies with the advanced learners of English a considerable number of non-target-like choices (e.g. *…*avoided to use*…) could be observed, the majority of catenative verb constructions produced by learners of different proficiency levels in the corpus data showed the opposite, namely a very high number of target-like choices. In all three studies, the respective mixed-effects model revealed that the frequency with which the catenative verb occurs with the respective complement type made a strong and significant prediction of the target-like choice of the complement type. However, the frequency of the matrix verb in general had an impact on the target-like choice of the complement type only in the corpus study but not in the experimental data. These findings provide an important insight into how frequency affects the mental representation of the catenative verb construction: it is essential for L2 learners to have experienced the catenative verb together with its target-like complement type rather than being familiar with the matrix verb alone. The catenative verb together with its complement type forms a processing unit (i.e. a construction) which needs to be experienced with sufficient frequency in order for the learner to build a native-like schema representation of the catenative verb construction.
Apart from the presentation of the most important research findings, this talk will also address methodological issues and will show how multivariate statistics be used to analyse corpus data and how corpus and experimental data can be successfully linked, despite their differences, to explore the acquisition of a syntactic construction.

References:

Baayen, R. H., Davidson D. J. & Bates D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Bybee, J. L. (2007). *Frequency of use and the organization of language*. Oxford, New York: Oxford University Press.

Davies, M. (2004-): BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

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(02). 143–188.

Geertzen, J., Alexopoulou T. & Korhonen A. (2013). The EF-Cambridge Open Language Database. https://corpus.mml.cam.ac.uk/efcamdat/ (20 October, 2015).

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

Gries, S. T. & Wulff S. (2005). Do foreign language learners also have constructions?: Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*(3). 182–200.

Huddleston, R. D. & Pullum G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK, New York: Cambridge University Press.

Levshina, N. (2015). *How to do Linguistics with R*: *Data exploration and statistical analysis*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Madlener, K. (2015). *Frequency Effects in Instructed Second Language Acquisition* (Applications of Cognitive Linguistics /ACL] 29). Berlin/Boston: De Gruyter Mouton.
Martinez-Garcia, M. T. & Wulff S. (2012). Not wrong, yet not quite right: Spanish ESL students' use of gerundial and infinitival complementation. *International Journal of Applied Linguistics* (22.2). 225–244.

A corpus-based evaluation of readability metrics as indices of syntactic complexity in EFL learners' written productions

Nicolas Ballier, Paula Lissón
Université Paris-Diderot (USPC)
nicolas.ballier@univ-paris-diderot.fr, paula.lisson@etu.univ-paris-diderot.fr

This paper deals with the lexical assessment and classification of learners through the implementation of readability metrics as indices of syntactical complexity. The aim of the paper is twofold: first, delimiting which of the 30 readability metrics used in the study shows the most appropriate values for classifying learners into different proficiency groups; and second, validating the possibility of using readability metrics with frequency lists of difficult words generated from the learner corpus analysed.

With the expansion of learner corpora, many studies dealing with the automatic assessment of learner's language complexity have tackled lexical and syntactic complexity (Cobb & Horst, 2015). For example, Lu (2010) creates a computational system for the analysis of syntactic complexity in second language writing with 14 built-in metrics. These metrics present a high degree of reliability when used, for instance, as an index of ESL learner's writing development (Lu, 2011). Similarly, Vajjala (2016) shows how lexical and syntactic metrics help assessing learners' production; and Ballier & Gaillat (2016) use these type of metrics in order to classify French learners of English into different proficiency groups.

However, the domain of readability in relation with Learner Corpus Research (LCR) remains slightly less explored. Broadly speaking, the role of readability measures in SLA has been used to establish the difficulty of texts in reading tasks (Kasule, 2011; Vajjala & Meurers, 2012). Readability measures are typically used so as to determine if a text is appropriate or not for learners of a particular level (François, 2011; Gala *et al.,* 2014). Few studies combine the use of readability and lexical/syntactical metrics, the Vajjala & Meurers (2012) study is an example of the interconnection between traditional readability measurements and SLA complexity metrics.

In this paper, we aim at changing the traditional point of view of readability metrics; we are not using readability in order to see how difficult a text might be for a given level of proficiency; but rather applying readability formulae to learners' productions so as to see if the metrics can be used to classify learners into different levels. In order to do so, we assess the validity of 35 of the readability metrics implemented in the {koRpus}(Michalke, 2016) package of R (R Core Team, 2016) by applying them to randomly chosen samples taken from NOCE (Díaz-Negrillo, 2007), a written corpus of Spanish university students of English. Replicating Lu (2012), we assess the strength of the correlations among the metrics using Spearman's '*ρ*' (see Table 1).

Some metrics (Spache, 1966; Bormuth, 1969; Chall & Dale, 1995) rely on the use/underuse of complicated words. These formulae rely on the implementation of lists of complex words which were originally compiled by and for native speakers of English, and its application to learner corpora might yield unsatisfactory results. Thus, the second aim of

this paper is to create a list of complex words according to their frequency in the NOCE corpus, and to implement it in the readability formulae, instead of using the original lists.

Table 1: Correlations among 3 metrics with their original lists implemented (*p.* < 0.001 in all the cases)

	Bormuth	**Dale.Chall**	**Spache**
Bormuth	1	0.824	-0.609
Dale.Chall	0.824	1	-0.834
Spache	-0.609	-0.834	1

By using a specific list generated from the corpus we are analysing, we can classify learners according to potentially more accurate criteria. Our contribution to *widening the scope of learner corpus research* is to suggest that we should design learner-based frequency lists to adequately describe learner data. Taking learner output as the baseline for linguistic analysis raises issues in terms of L2 attainment that we also discuss.

References:

Ballier, N., & Gaillat, T. (2016). Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique (Vol. 9, pp. 1–14). Presented at the JEP-TALN-RECITAL 2016.

Cobb, T., & Horst, M. (2015). Learner Copora and Lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Díaz-Negrillo, A. (2007). *A Fine-grained Error Tagger for Learner Corpora* (PhD Thesis). University of Jaen, Jaen.

François, T. (2011). Les apports du traitement automatique des langues à la lisibilité du français langue étrangère.

Gala, N., François, T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots (pp. 91–102). Presented at the Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014).

Kasule, D. (2011). Textbook readability and ESL learners. *Reading & Writing*, *2*(1), 63–76.

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

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

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208.

Michalke, M. (2016). koRpus: An R Package for Text Analysis (Version 0.06-5). Retrieved from http://reaktanz.de/?c=hacking&s=koRpus

R Core Team. (2016). R: A language and environment for statistical computing. (Version 3.3.1 (2016-06-21)). Vienna, Austria.: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Spache, G. D. (1966). *Good reading for poor readers* (Revised 9th edition). Champaign, Illinois: Garrard.

Vajjala, S. (2016). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *arXiv Preprint arXiv:1612.00729*. Retrieved from https://arxiv.org/abs/1612.00729

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition (pp. 163–173). Presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics.

On the relation between L1 and L2 speech rate

Michaela Banýrová, Lucie Jiránková, Martin Sedláček, Alena Novotná, Alena Kvítková
Charles University, Prague
banyrovam@gmail.com, lucie.jirankova@seznam.cz, sedlacek.martin256@gmail.com, novotna4alena@gmail.com, leiarevan@gmail.com

The presented research project contributes to the study of spoken learner language and focuses especially on the study of L2 fluency. It studies to what extent the speech habits in the mother tongue are projected into the learner language. The aim of the project is to supplement the LINDSEI_CZ corpus (i.e. the corpus of English spoken by proficient learners with Czech as their L1) with recordings and transcriptions of the same learners speaking their mother tongue. The purpose is to examine speech rate and disfluency features in both their L1 and L2, and thus contribute to the study of language transfer in speech fluency.

This field of study constitutes a research gap, as it is not very common to supplement learner corpora with recordings of mother language, and fluency is usually studied as a concept independent of learner's L1. Our project focuses on specific aspects of fluency such as speech rate, false beginnings, hesitation pauses, reduplications, time-buying strategies, etc. As regards speech rate, the only known comparison is Hincks' study (2008) which examined differences in the speech rate of the same speakers in Swedish and English. However, it was not a study based on a learner corpus.

The aim of this project is to establish whether L1 speech habits – and at this stage of our research project specifically whether L1 speech habits regarding speech rate and its variability – affect spontaneous spoken production of these speakers in a foreign language. To this end, we have started compiling a spoken corpus which is designed to supplement the existing learner corpus LINDSEI_CZ with the recordings of the same learners speaking their L1 (Czech), and compare the speech rates in the individual tasks. LINDSEI_CZ contains recording of Czech students at a C1-C2 levels of English (students of the 3rd and 4th grade of the bachelor programme English and American studies at Charles University in Prague). The structure of the new corpus is identical as in LINDSEI_CZ so that reliable comparisons can be made. It consists of three different tasks: first, a monologue on a chosen topic, next, a dialogue with the interviewer, and finally, a story reconstruction based on a set of pictures. We want to examine whether there is a correlation between L1 and L2 speech rates, and also to what extent L1 disfluency features are projected into English. This should help us establish whether disfluencies stem from inadequate automatization of learner's speech in the foreign language, or whether they might exist and have a similar character in the learner's mother tongue.

The methodology is based on the CAF model (complexity-accuracy-fluency) of language production and proficiency described for instance in Housen et al. (2012), Götz (2013) or Gráf (2015) but it will pave the way for comparisons of other LINDSEI subcorpora and contrastive interlanguage analysis (Granger, 2015). The actual data collection (i.e. the

recording and its transcription) will follow methodological recommendations published for example in the Routledge Handbook of Corpus Linguistics (2012) or the Cambridge Handbook of Learner Corpus Research (2015).

Thanks to its interdisciplinary overlap, the research results will have potential value even outside the context of Czech and English, and may serve as a starting point for other studies comparing more (foreign) languages, whether within LINDSEI or other learner corpora. It should also address the question of whether learner corpora ought to contain an L1 sample component.

The project is now in the stage of compilation of the L1 component of the corpus. We aim to complete the corpus in spring 2017 so that the results of our analyses are prepared for the LCR conference. However, the preliminary analyses indicate that there might exist a correlation between the L1 and L2 speech rates with the L2 speech rate being approximately 25% slower than in L1. The speech rate in L2 may indeed be affected by L1 speech rate and L1 speech habits.

References:

Götz, S. Fluency in Native and Nonnative English Speech. Studies in Corpus Linguistics, volume 53. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2013.

Gráf, Tomáš. 'Accuracy and Fluency in the Speech of the Advanced Learner of English'. PhD Thesis, Charles University, 2015. https://is.cuni.cz/webapps/zzp/detail/151663/.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 2015.

Hincks, R. 'Presenting in English or Swedish: Differences in Speaking Rate'. In Proceedings Fonetik 2008: The XXIst Swedish Phonetics Conference, edited by A. Eriksson and J. Lindh, 21–24. Gothenburg: Reprocentralen, Humanisten, University of Gothenburg, 2008. <http://www.ling.gu.se/konferenser/fonetik2008/papers/Proc_fonetik_2008.pdf>.

O'Keeffe, Anne, and Michael McCarthy (eds.) The Routledge Handbook of Corpus Linguistics. Routledge Handbooks in Applied Linguistics. Milton Park, Abingdon, Oxon ; New York: Routledge, 2012.

Normalization in Context: Facilitating Automatic Analysis of Learner Language

Adriane Boyd, Detmar Meurers
University of Tübingen
adriane@sfs.uni-tuebingen.de, dm@sfs.uni-tuebingen.de

Learner language frequently contains non-canonical orthography and morphosyntactic constructions that present difficulties for natural language processing tools developed for standard language. Since manually annotated learner corpora are often small and the high degree of variation in learner productions leads to data sparsity issues even for larger learner corpora, it is useful to consider tools that automatically normalize non-standard aspects of learner language. While normalization and applying standard language categories to learner language does not address the full spectrum of learner language analysis and fundamental concerns about analyzing learner language (cf. Meurers & Dickinson, 2017), it can facilitate access to learner language in applications such as corpus search tools and computer-aided language learning systems. In this work, we investigate the range of resources required to normalize learner language, in particular the extent to which an explicit task context can inform and improve normalization. We apply and evaluate these insights in an automatic normalization approach.

Normalizations such as the concept of a *minimal target hypothesis* from the FALKO German learner corpus (Reznicek et al., 2012) have been developed in order to provide a version of a learner production that can be systematically searched and is more appropriate for further automatic analysis. The minimal target hypothesis contains a minimal number of modifications that convert the learner sentence into a locally grammatical sentence. As it may not be possible to determine exactly what the learner intended to say in an open-ended task such as an essay task, what constitutes a minimal change is based on grammatical properties, e.g., preserving the provided verb and modifying its arguments rather modifying the verb itself.

In contrast to open-ended tasks, a more explicit task context can provide more information about the potential meaning of a learner production (Meurers, 2015, sec. 2.1). The task context thus can also provide additional information for normalization. One such corpus is the Corpus of Reading Comprehension Exercises in German (CREG, Ott, Ziai & Meurers, 2012), which includes reading texts, reading comprehension questions, target answers written by teachers, language learners' responses to the questions, and teachers' evaluations of those responses. Building on a subset of CREG, Keiper et al. (2016) present a manually annotated corpus containing two levels of normalizations and part-of-speech tags for approximately 1000 learner answers along with an automatic normalization approach focusing on misspellings related to words from the task context, which account for approximately 60% of their normalizations. In their part-of-speech tagging evaluation, a standard tagger's performance improves 0.9% (corresponding to a 12.5% reduction in error) on their automatic normalizations and a further 1.8% in accuracy on gold normalizations, showing the benefit of normalization and also the potential for further improvement.

We systematically explore the dependence of normalization on task context through a manual annotation study, focusing on responses marked as correct by teachers so that an explicit meaning-based target hypothesis is available. We find that the inter-annotator agreement for normalization of non-words increases as the amount of context presented to annotators increases: from 0.75 (Krippendorff's alpha) for the learner response in isolation to 0.79 for the response and question and further to 0.84 for the response, question, and reading text. We show that such a characterization supports the development of targeted automatic normalization approaches and discuss the linguistic resources needed to realize them. Some types of errors, such as German umlaut misspellings (e.g., *ueber* for *über*), require very little context and few linguistic resources to normalize while others such as *gelt*, which has a range of potential normalizations including *gilt/galt* 'is/was deemed' and *Geld* 'money', require more context, further resources, and deeper linguistic analysis. In addition to a direct evaluation of an automatic normalization approach integrating the task context, we will provide extrinsic evaluations for part-of-speech tagging, dependency parsing, and on the automatic scoring of learner answers to reading comprehension questions as an externally validated task.

References:

Keiper, L., A. Horbach & S. Thater (2016). Improving POS Tagging of German Learner Language in a Reading Comprehension Scenario. In *Proceedings of LREC 2016*. Portorož, Slovenia: European Language Resources Assocation.

Meurers, D. (2015). Learner Corpora and Natural Language Processing. In S. Granger, G. Gilquin & F. Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, 537–566.

Meurers, D. & M. Dickinson (2017). Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning* 67 (S1), 66–95.

Ott, N., R. Ziai & D. Meurers (2012). Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In T. Schmidt & K. Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam: Benjamins, Hamburg Studies in Multilingualism (HSM), 47–69.

Reznicek, M., A. Lüdeling, C. Krummes & F. Schwantuschke (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0*. http://purl.org/net/Falko-v2.pdf.

Comparing the "phrasicon" of teenagers in immersive and non-immersive settings: does input quantity impact range and accuracy?

Amélie Bulon, Fanny Meunier
Université catholique de Louvain
amelie.bulon@uclouvain.be, fanny.meunier@uclouvain.be

The present paper falls within the framework of an interdisciplinary project on Content and Language Integrated Learning (CLIL) in French speaking Belgium. The project aims to assess CLIL at the interface of linguistic, cognitive and educational perspectives (Hiligsmann et al. in preparation). In this paper we specifically focus on the acquisition of Dutch and English phraseological units by French-speaking CLIL (immersive setting) and non-CLIL (non-immersive setting) secondary school pupils.

Several studies (e.g. Zydatiß 2007; Lorenzo & Moore 2010; Jexenflicker & Dalton-Puffer 2010; Gené-Gil et al. 2015; Martínez 2015; Bulon et al. forthcoming) have been carried out to compare the language proficiency of learners in immersive and non-immersive settings using global measures of complexity, accuracy and/or fluency, typically referred to as CAF (Housen et al. 2012; Norris & Ortega 2009). The present study focuses on the pupils' phrasicon, i.e. their phraseological lexicon, and reports on two main analyses, viz.: 1) an overview of the phraseological errors (focus on accuracy), and 2) an analysis of the variety/range of the phrasicon (focus on complexity). Since CLIL programs provide more target-like and input-rich environments than non-CLIL programs - and can therefore be considered closer to L1 acquisition because of their inherent usage-based approach - we hypothesize that CLIL pupils have a phrasicon that is both more accurate and more varied than that of pupils in non-CLIL settings.

The participants are 5th year French-speaking secondary school pupils in immersive settings (n=90) and non-immersive settings (n=90) learning English as a foreign language. The analysis is based on a corpus of written productions in the form of 180 e-mails (90 per category) on two similar topics.

Wordsmith Tools (Scott 2012) was initially used to extract the phrasemes (word clusters at that stage). The list was then manually checked, pruned and organized per category of phraseme (referential phrasemes, textual phrasemes and communicative phrasemes; see Granger & Paquot 2008). The errors found in the phrasemes produced by the learners were then classified on the basis of Thewissen's (2008) and Hong et al.'s taxonomies (2001). The various categories and error types used in the study will be presented and illustrated during the talk.

At the time of writing the present abstract various statistical analyses are still being carried out but our preliminary results reveal the following trends:

a) of all types of phraseological errors, the most frequent ones (both for immersive and non-immersive pupils) are: wrong choice of verbs and nouns in referential phrasemes (existing word but wrong selection); wrong choice of prepositions (be they dependent or

not), and the use of two separate words (open compounds) instead of one-word (solid) compounds;
b) both groups of pupils make more grammaticality errors (formally inexistent phraseme; e.g. in love for* [with]) than acceptability errors (formally existing phraseme but inappropriately used in context; e.g. The hotel had a great view on* the ocean [of/over]);
c) immersive pupils seem to have more difficulties with compounds than non-immersive pupils (more errors in each category);
d) with regard to the sources of phraseological errors, L1 transfer appears to be the most prominent among the three major categories (possible L1 transfer, possible transfer from learner's other L2 and possible intralingual errors), and this for both groups of pupils;
e) more intralingual errors are found in the non-immersive group;
f) overall, the phrasicon of immersive pupils tends to be more varied.

In the light of our current results, we can argue that phraseological use remains problematic for foreign language learners, even for those who are exposed to a much larger quantity of input. Even though immersive pupils tend to produce a more varied phrasicon than their peers, they still tend to rely heavily on their L1 and thus produce non-native like phrasemes. Our initial hypothesis that CLIL pupils have a phrasicon that is both more accurate and more varied than that of pupils in non-CLIL settings is thus only partly validated as only the variety was higher for CLIL pupils. Our results also support the idea that the acquisition of phraseological units requires explicit teaching, awareness-raising and focus on form activities (see Meunier (2012) for a review of possible instructional intervention types favoring the acquisition of the phrasicon in a second/foreign language), no matter the quantity of input being provided.

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. To appear in *Travaux du Cercle Belge de Linguistique (CBL)*.

Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Development of EFL writing over three years in secondary education: CLIL and non-CLIL settings. The Language Learning Journal, 43(3), 286-303.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), Phraseology: An interdisciplinary perspective. John Benjamins, Amsterdam/Philadelphia, 27-49.

Hiligsmann, P., Van Mensel, L., Galand, B., Mettewie, L., Meunier F., Szmalec, A., Van Goethem, K., Bulon, A., De Smet A., Hendrikx, I., Simonis, M. (in preparation). Assessing Content and Language Integrated Learning (CLIL) in the French-speaking Community of Belgium: linguistic, cognitive and educational perspectives. Cahiers du Girsef.

Hong, A. L., Rahim, H. A., Hua, T. K. & Salehuddin, K. (2001). Collocations in Malaysian English learners' writing: A corpus-based error analysis. 3L; Language, Linguistics and Literature, The Southeast Asian Journal of English Language Studies, 17 (special issue), 31-44.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), Dimensions of

L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. John Benjamins, Amsterdam/Philadelphia, 1-20.
Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), Language use and language learning in CLIL classrooms. John Benjamins, Amsterdam, 169-190.
Lorenzo, F., & Moore, P. (2010). On the natural emergence of language structures in CLIL: Towards a theory of European educational bilingualism. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), Language use and language learning in CLIL classrooms. John Benjamins, Amsterdam, 23-38.
Martínez, A. C. L. (2015). Analysis of the Written Competence of Secondary Education Students in Bilingual and Non-Bilingual Programmes. In Conference proceedings. ICT for language learning. libreriauniversitaria.it edizioni, 499-503.
Meunier, F. (2012). Formulaic Language and Language Teaching. Annual Review of Applied Linguistics, 32, 111–129.
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.
Scott, M. (2012). WordSmith Tools. Liverpool: Lexical Analysis Software.
Thewissen, J. (2008). The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. Proceedings from the 8th Teaching and Language Corpora Conference (TaLC 8). In Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 300-306.
Zydatiß, W. (2007). Deutsch-Englische Züge in Berlin (DEZIBEL). Eine Evaluation des bilingualen Sachfachunterrichts in Gymnasien: Kontext, Kompetenzen, Konsequenzen. Frankfurt am Main: Lang, 84.

Developing a CEFR-based vocabulary inventory for young learners - Comparing native-speaker and EFL learner corpus data

Marcus Callies[1], Veronica Benigno[2]
University of Bremen[1], Pearson Education[2]
callies@uni-bremen.de, veronica.benigno@pearson.com

This talk answers recent calls for the field of LCR to increase efforts to investigate a greater variety of learner demographics (esp. young learners) and to resort to text-based methods in the assessment of proficiency (Paquot & Plonsky 2017). We will present some first results of a project aimed at creating a vocabulary framework that indicates what word meanings learners aged 6 up to 11 years are expected to know at different proficiency levels in the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001) and Pearson's *Global Scale of* English (GSE; https://www.english.com/gse). Such an inventory for young learners will guide EFL teachers to identify level-appropriate vocabulary targets for their students. Young learners, i.e. learners aged between 6 to 12 years, an age range at which children attend primary/elementary school in many countries, differ from adults in the way they acquire a foreign language at least for two reasons: First, they have unique affective and cognitive characteristics and preferences, and second, their learning experience is linked to the daily here and now, largely concerned with the school domain, and thus much more affected by the respective linguistic and cultural background (Benigno & De Jong 2015). This is particularly important in the EFL context because of the very few chances children usually have to use the target language. Most importantly, research suggests that the L1 context is of great relevance to young learners; in fact, it has been shown that children map a word of the L2 onto concepts in their L1, unless the concept in the L2 is new and therefore adds to the L1 mental lexicon (Ellis 1997: 133).
The talk first outlines the methodology that has been applied and tested in the creation of the GSE Vocabulary Inventory for Adults, a graded lexical inventory aligned to the CEFR (Benigno & De Jong 2016) and recently launched by Pearson (https://www.english.com/gse/teacher-toolkit/user/vocabulary). To set up a comparable inventory for young learners, child-language samples of speech and writing form several reference corpora of L1 English were analysed to identify the frequency of occurrence of vocabulary items, yielding a frequency list of the top 2,000/3,000 word meanings. EFL coursebooks and teaching resources for young learners were consulted to integrate the corpus-based list with low-frequency items which were considered to be essential in the young learners' context of language use. The vocabulary items were then annotated for topics such as "classroom language", "hobbies and games", or "family and self". Since children mostly learn vocabulary in the classroom context, with very little exposure to the target language and therefore little chance of repetition and reinforcement, particular importance was given to vocabulary used in the school domain.
In the second and central part of the talk we report on the results of a study that validates the vocabulary inventory for young learners created on L1 child-language data against

comparable EFL data. We analysed written EFL data from the *International Corpus of Crosslinguistic Interlanguage* (ICCI; Tono & Díez-Bedmar 2014), i.e. short descriptive/narrative texts produced by primary-school learners from different L1 backgrounds. The primary objective is to gather evidence that the vocabulary inventory compiled by Pearson is relevant to the target group of EFL learners and that more proficient young learners tend to produce more "advanced" vocabulary than less proficient ones, possibly resulting in a correlation between learners' institutional grade (as a proxy for their proficiency level) and the order of the lexical items by CEFR and GSE. A minor objective is Match topic-specific L2 wordlists extracted from the learner corpus against the Pearson L1 wordlist to identify learner preferences in word use or culture-specific words.
It is possible to identify a core vocabulary of YLS, and in addition, a "localized" vocabulary which may reflect the learners' L1 (cultural) background/setting. The core vocabulary will cover essential notional and functional areas which refer to basic communicative acts carried out by children regardless of their L1, whereas the localized vocabulary will be linked to the experiential environment and therefore requires adaptation to concepts and situations that are particular to the children's L1.

References:

Benigno, V. & De Jong, J. (2015). The Global Scale of English learning objectives for young learners: A CEFR-based inventory of descriptors. In M. Nikolov (Ed.). *Assessing Young Learners of English: Global and Local Perspectives*. Berlin: Springer, 43-64.

Benigno, V. & De Jong, J. (2016). *Vocabulary White Paper*. https://prodengcom.s3.amazonaws.com/GSE_WhitePaper_Vocabulary.pdf.

Council of Europe (2001). The Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: CUP.

Ellis, N. C. (1997). Vocabulary acquisition: Word structure, collocation, grammar, and meaning. In M. McCarthy & N. Schmidt (Eds.). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: CUP, 122-139.

Paquot, M. & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1), 61-94.

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.

A longitudinal investigation of multi-word constructions in a learner corpus: a growth curve modelling approach

Duygu Candarli
University of Manchester
duygu.candarli@manchester.ac.uk

Phraseology has been a major area of interest within learner corpus studies and the field of English for academic purposes. It is reported that L2 learners of English have difficulty in using phraseological patterns in their essays (for an extensive review of the studies, see Paquot & Granger, 2012). Recent studies have investigated the use of lexical bundles in second language writing across different levels (e.g. Ädel & Römer 2012; Chen & Baker, 2014; Staples, Egbert, Biber, & McClair, 2013); however, few studies have researched phraseological units within a longitudinal design in second language writing (e.g. Bestgen & Granger, 2014; Li & Schmitt, 2009). Therefore, it is important to capture the developmental and interlanguage features of learner writing within a longitudinal research design. Also, little is known about the phraseological development of language learners/users in discipline-specific academic writing at an English-medium university.

The learner corpus of this study consists of 294 English essays of 98 Turkish students who were in their first year at an English-medium university in Turkey. The essays were collected at the beginning of the first semester, at the end of the first semester, and at the end of the second semester from the same students. The participants had an advanced level of English proficiency, and they submitted their assignments in English. In their first year at the university, they took 'Advanced Writing in English' courses at both fall and spring semesters in their first year; however, there was no explicit instruction on academic phrases. It could be said that the students were expected to internalise academic discourse and begin academic socialisation through academic writing during their first year at university. As Ortega and Iberri-Shea (2005) stated, longitudinal research is "better motivated when key events and turning points in the social or institutional context investigated are considered" (p. 38). Though one year may not be long enough to regard this study as longitudinal, it was designed to offer insights into Turkish EFL students' phraseological development. Ellis (2002) argues that L2 learners with a lower level of language proficiency rely on fixed phraseological patterns to a greater extent than those with a higher level of language proficiency. Therefore, it was hypothesised that there would be a slight decreasing trend of fixed phraseological patterns in learners' essays over one academic year.

The present study used a corpus-based approach in order to identify multi-word constructions (MWCs) that include two-, three- and four-word fixed sequences. Accordingly, in the corpus, I searched for Liu's (2012) list of the 228 most common MWCs in general academic written English organised by their semantic functions that include referential expressions, discourse organisers, and stance expressions which are in line with the discourse functions of lexical bundles proposed by Biber, Conrad and Cortes (2004). The occurrences of each semantic category of MWCs were recorded for each text, and the frequencies were normalised per 100 words in each text.

A multi-level/mixed-effects modelling is recommended in corpus linguistics studies (Gries, 2015). In order to capitalise on richness of the longitudinal data, a growth curve model (Mirman, 2014; Singer & Willet, 2003), a variant of mixed-effects models, was used to examine the trajectories of the frequencies of each semantic category of MWCs in participants' essays over time. Following Barr et al. (2013), participants and each semantic category of MWCs were treated as random effects, and maximal random effect structures; that is, the model that contained random intercepts and slopes for all independent variables were built, using the lme4 package in R. Overall, the results showed that the mean frequencies of MWCs, except referential expressions, were not static over time, and that the frequencies of discourse organisers showed a decreasing trend. The statistical significance of the first order polynominal random effects suggested that participants differed from each other in terms of the frequencies of MWCs, which shows heterogeneity of the learner data. The results suggest that the learners rely on target-like fixed expressions to a lesser extent as they gain experience in academic writing.

References:

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan corpus of upper-level student papers. *International Journal of Corpus Linguistics*, *17*(1), 3–34.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Chen, Y., & Baker, P. (2014). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, 1–33.

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.

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*(1), 95–125.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85-102.

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

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: Taylor & Francis.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45.

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225.

Teaching English for tourism: On the use of adjectives in texts written by (EFL) novice writers and by native and non-native professionals

Erik Castello
University of Padua
erik.castello@unipd.it

In order to write effective tourism texts in English, novice writers, and in particular EFL learners, need to become familiar with the keywords, collocations and grammar "institutionalised in the English currently and authentically used in the tourism industry" (Lam 2007). They need to learn the typical phraseology of this Language for Specific Purposes (LSP) and to find the right balance between clichéd writing and creative uses (Dann 1996). Teaching materials, however, often fall short of detailed descriptions of the vocabulary and grammar necessary for tourism text writing (Lam 2007), and more research is needed in this area.

The study presented in this paper is part of a larger project (Castello 2013), which uses (learner) corpus linguistic methodologies in the attempt to pinpoint the main lexico-grammatical and discourse features that are specific to this LSP, as recommended, among others, by Flowerdew (2015).

The paper explores the use of adjectives, which rank among the most prominent linguistic features of the language of tourism (Gotti 2006; Pierini 2007). Functionally, adjectives can be divided into evaluative/denotative, descriptive/connotative and intermediate ones on the continuum between the two poles (Pierini 2009). In some tourism text types they are often used attributively, in which case they pre-modify noun phrases and contribute to their density (Nilson 2000).

Broadly following the methodology put forward by Ädel (2006), the study compares texts promoting an Italian city and produced by various types of writers at different levels of English proficiency and expertise: internationally renowned publishing houses, local tourist boards, EFL learners, and English native-speaker students. It is (so far) based on four small corpora. The first one consists of the sections about Padua extracted from international online travel guides (6,073 words), while the second one is made up of the English sections about Padua taken from the official websites of local boards (6,675 words). The third corpus is a learner corpus of texts written by Italian university students in response to a prompt in Italian which provided some information about the city of Padua in the form of sketchy notes (4,647 words). Finally, the last is composed of texts written by native speaker students who were asked to write the same text as the Italian learners under the same conditions (5,533 words).

All of these writers clearly aim at promoting a given city and its attractions, yet they write for different audiences and with slightly different purposes, which is likely to impact on their lexico-grammatical choices. The study attempts to answer the following research questions:

- What are the main quantitative and qualitative differences, if any, between the four types of texts with respect to the use of adjectives?
- Why do these differences apply to the corpora?
- What target should non-professional EFL learners ultimately aim at?

The analyses involved retrieving all the adjectives used in the four corpora using *The Sketch Engine*[1]. The concordance lines were then inspected manually and sorted out and quantified using Microsoft Excel. The software *Range*[2] was also used to conduct some investigations.

The first step in the analysis explored the use of adjectives and their frequency of use across the corpora. The second step delved into their "lexical sophistication" by breaking down the adjectives according to the frequency word lists they are associated with by the software Range. The third step investigated some syntactic patters of noun pre-modification involving the use of adjectives (e.g. *adj. + n. + n.*). The last step focused on semantic and pragmatic aspects and distinguished between denotative and connotative adjectives, and if connotative between favourable and unfavourable evaluative ones (Partington 1998).

The preliminary results suggest that EFL learners tend to use higher percentages of tokens of adjectives, yet it is international publishing houses who produce the highest percentages of types, of low-frequency adjectives and of the *adj.+n.+n.* pattern. They also indicate the presence of some infelicitous uses in the writing of both types of novice writers and in that of the local tourist boards. Finally, the international publishing houses and the EFL learners often use adjectives with an unfavourable connotation, while local boards rarely do so.

The paper will discuss these and other findings and their implications for teaching English for tourism. Not only will it look at learners' mistakes, but also at the role of linguistic creativity in this LSP and at how it can be successfully achieved.

References:

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

Castello E. (2013). Exploring Existential and Locative Constructions in a Learner and in an Expert corpus of Promotional Tourist Texts. In D. Heller, C. Desoutter C. & M. Sala (Eds.). *Corpora in specialized communication - Korpora in der Fachkommunikation - Les corpus dans la communication spécialisée*. Bergamo: CELSB Libreria Universitaria, 385-410.

Dann, G. (1996). The Language of Tourism: A Sociolinguistic Perspective. London: CAB International.

Flowerdew, L. (2015). Learner corpora and language for academic and specific purposes. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 465-484.

Gotti, M. (2006). The Language of Tourism as Specialized Discourse. In O. Palusci & F. Francesconi (Eds.). *Translating Tourism: Linguistic/Cultural Representations*. Trento: Editrice Università degli Studi di Trento, 15-34.

[1] <https://www.sketchengine.co.uk>, last visited on 31st May 2017.

[2] <http://www.victoria.ac.nz/lals/about/staff/paul-nation>, last visited on 31st May 2017.

Lam, P. (2007). A Corpus-driven Lexico-grammatical Analysis of English Tourism Industry Texts and the Study of its Pedagogic Implications in English for Specific Purposes. In E. Hidalgo, L. Quereda, & J. Santana (Eds.). *Foreign Language Classroom: Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam: Rodopi, 71-89.

Nilson, T. (2000). Noun Phrases in British Travel Texts. In C. Mair & M. Hundt (Eds.). *Corpus Linguistics and Linguistic Theory.* Amsterdam: Rodopi, 267-274.

Partington, A. (1998). Patterns and Meanings: Using Corpora for English Language Research and Teaching. Amsterdam/Philadelphia: Benjamins.

Pierini, P. (2007). Quality in Web Translation: An Investigation into UK and Italian Tourism Web Sites. *The Journal of Specialised Translation* 8, 85-103.

Pierini, P. (2009). Adjectives in Tourism English on the Web: A Corpus-based Study. *CÍRCULO de Lingüística Aplicada a la Comunicación* 40, 93-116.

Individualising learner corpora in EAP: Doctoral students' use of corpus tools for editing

Maggie Charles
Oxford University Language Centre
maggie.charles@lang.ox.ac.uk

This paper aims to widen the scope of learner corpus research by showing how students can use their own writing as an individualised learner corpus and through the use of corpus tools make improvements in their texts. The use of learner corpus data has already had a considerable impact on language pedagogy (Granger, 2009). Not only has it contributed indirectly to the development of teaching materials such as dictionaries and grammar guidance (Gilquin et al., 2007), but it has also offered a basis for data-driven learning materials for direct classroom use (Nesselhauf, 2004). In this regard, it has been shown that providing students with access to a learner corpus of their own writing is particularly motivating (Lee and Swales, 2006; Seidlhofer, 2002) and that increased gains can be made when students work not only with native-speaker corpora but also with a local learner corpus (Cotos, 2014).

However, there has been little take-up of Mukherjee and Rohrbach's (2006) suggestion that students work on individual learner corpora of their own writing. This study aims to address this gap by reporting on a course in corpus-assisted editing for doctoral students with English as L2. After an initial session introducing corpus work, students built two corpora: 1) a learner corpus of their own writing, which ranged in size from 2,752-142,494 words; and 2) an expert corpus of research articles in their own field (size range: 77,000-3.3 million words). The freeware AntConc (Anthony, 2014) was used for examining the corpora. Class sessions provided demonstrations of how specific tools can be used for investigating the learner corpus, followed by individual practice in which students used the tools to facilitate editing of their theses.

The research questions addressed were:

1. How do students evaluate the use of their individual learner corpus and AntConc tools for editing?
2. What are the affordances of specific tools when applied to investigating an individual learner corpus?

The course has run nine times and evaluation data are available for 66 students (41% natural sciences; 30% social sciences; 29% humanities). All participants gave a positive answer to the question 'Is it helpful to use your corpus and AntConc for editing?' (79% 'yes definitely'*;* 21% 'yes probably'*)*. Students were asked to rate the individual tools for editing purposes as 'very useful', 'useful', 'fairly useful', 'of little use' or 'not useful'. Combining the 'very useful' and 'useful' categories shows that the most highly rated tool was the Concordancer at 95% of responses. This was followed by Clusters (82%), Collocates and Keyword List (both 74%), N-grams (70%), Concordance Plot (63%) and Word List (59%).

The utility of concordances in data-driven learning has already been shown, particularly in relation to comparisons of learner and expert corpus data (Granger and Tribble, 1998; Millar and Lehtinen, 2008). However, I argue that other tools such as N-grams, Keyword

List and Concordance Plot are also useful when applied to an individual learner corpus. For example, the N-Grams tool can be used to make a list of all the 3-grams in the student's own writing and compare it with those in an expert corpus, thereby revealing differences in phraseology, while both Keywords and Concordance Plot allow issues concerning the content of the student's thesis to be addressed. A keyword list of one section or thesis chapter compared to the rest of the text identifies the words that occur more (or less) frequently than expected. This tool can therefore reveal the most salient words in a section or chapter and thus the extent to which the writer deals adequately with the topic under discussion. Concordance Plot provides a graphic representation of the distribution of a search term throughout the corpus files. When the term chosen is central to the student's argument, this tool can show how the content develops over the course of the whole text. The present paper discusses further the affordances of these and other tools when used in conjunction with an individual learner corpus, illustrating the findings with examples of student investigations.

References:

Anthony, L. (2014). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. ReCALL, 26(Special Issue 02), 202–224.

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

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.). Corpora and Language Teaching. Amsterdam: Benjamins, 13–32.

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.). Learner English on Computer. London: Longman, 199–209.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. English for Specific Purposes, 25(1), 56–75.

Millar, N., & Lehtinen, B. (2008). DIY local learner corpora: Bridging gaps between theory and practice. The JALT CALL Journal, 4(2), 61–72.

Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (Eds.). Planing, gluing and painting corpora: Inside the applied corpus linguist's workshop. Frankfurt: Peter Lang, 205–232.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. M. Sinclair (Ed.). How to Use Corpora in Language Teaching. Amsterdam: Benjamins, 125–152.

Seidlhofer, B. (2002). Pedagogy and local learner corpora. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.). Computer learner corpora, Second language acquisition, and foreign language teaching. Amsterdam: Benjamins, 213–234.

Comprehensive Complexity Analysis of Large-scale Learner Corpora with the Common Text Analysis Platform

Xiaobin Chen, Detmar Meurers
University of Tübingen
xiaobin.chen@uni-tuebingen.de, dm@sfs.uni-tuebingen.de

The Common Text Analysis Platform, or CTAP (Chen & Meurers, 2016) is a Web-based computational system for automatic extraction of linguistic features from language productions. It combines state-of-the-art Natural Language Processing (NLP) technologies and complexity research to provide language researchers and education practitioners a tool to effectively and efficiently analyze large amount of learner data. The following first explains the rationales for complexity analysis of learner language to identify the needs for a platform that offers automatic and comprehensive analytical capabilities. Then the functionalities of the CTAP system will be presented.

The concepts of Complexity, Accuracy, and Fluency (CAF) have been widely used by researchers and practitioners as comprehensive and adequate constructs for measuring L2 performance and proficiency (Skehan, 1989; Ellis, 2003, 2008). Not only have CAF been used to evaluate the written and spoken performance of learners to determine their proficiency levels, but they have also been used as proxies to the developmental trajectories of their proficiencies (Housen et al., 2009). Among the CAF triad, complexity is the most researched construct in language acquisition studies (e.g. Norris and Ortega, 2009; Lu, 2010, 2012; Kyle and Crossley, 2015). It is defined as the elaborateness and variedness of the learner's language production (Ellis, 2003) on various linguistic levels such as lexical, morphological, syntactic, and phonological levels (Bulté and Housen, 2012). Modeling learner performance or language proficiency development from the complexity perspective requires analysis of large amount of learner production, which is difficult and laborious, if not impossible, without the help of modern NLP technologies. A number of automatic complexity analysis tools such as the Syntactic and Lexical Complexity Analyzers (Lu, 2010), CohMetrix (McNamara et al., 2014), and the Tool for the Automatic Analysis of Lexical Sophistication (Kyle and Crossley, 2015) have emerged in the past few years. Although these systems provide a valuable toolkit for analyzing learner language, they are geared more towards expert users of computers. Furthermore, a comprehensive analysis of large volumes of learner data is only achievable by utilizing the individual tools separately, because each of the tool deals with one or a few aspects of the complexity construct. Consequently, a platform that allows easy and comprehensive acquisition of complexity measures from learner corpora to support research on performance assessment and proficiency development is on demand.

The CTAP system is designed to meet these needs. It features 1) a consistent, easy-to-use, and friendly user interface, 2) modularized, reusable, and collaborative development of analysis components, and 3) flexible corpus and feature management. Four main user modules, namely the Corpus Manager, the Feature Selector, the Analysis Generator, and the Result Visualizer and a server module make up the CTAP system. The Corpus Manager helps users organize language materials to be analyzed into corpora, labeled groups, and

corpus folders. The Feature Selector allows for selection of different complexity measures for different analysis needs. The selected features could then be applied to different corpora as with the Analysis Generator. The analysis results or complexity feature values can be plotted with the Result Visualizer or be downloaded as Comma Separated Values (CSV) files for further analysis with external statistical tools.

More than 170 complexity features including measures of lexical density, lexical variation, lexical sophistication and syntactic complexity have been included into the CTAP system and the feature list will keep growing with contributions from developers all over the world thanks to the open-source nature of the system. This is by far the most comprehensive and easy-to-use system freely available online for complexity analysis. The open-source nature of the CTAP project also makes the system fully transparent in every aspect. New feature extractors can be easily plugged into the system by wrapping them as Unstructured Information Management (UIMA, https://uima.apache.org) analysis engines. This project setup encourages collaborative development and enhancement of the system among researchers and developers.

Analyzing learner corpus poses a great challenge to language researchers, especially those who are not programmers or expert computer users. Presented here is the CTAP platform that is designed to release researchers from the hustle and bustle of other automatic complexity analysis tools. The system provides a comprehensive set of complexity features and makes it easy and flexible to manage and analyze large learner corpora. The CTAP system as both a running production-level Web application and an open-source project for collaboration is accessible at http://ctapweb.com and https: //github.com/ctapweb respectively.

Acknowledgments

This research was funded by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Xiaobin Chen is a doctoral student at the LEAD Graduate School & Research Network.

References

Bulté, B. & Housen, A. (2012). Defining and operationalising l2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.). *Dimensions of L2 Performance and Proficiency*. John Benjamins, 21–46.

Chen, X.B. & Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity*, pp. 113–119, Osaka, Japan. The International Committee on Computational Linguistics.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd Ed). Oxford : OUP.

Housen, A., Kuiken, F., Zuengler, J., and Hyland, K. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786.

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.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge : CUP.
Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
Skehan, P. (1989). Individual Differences in Second Language Learning. London: Edward Arnold.

Where Does Alignment Occur? Analyzing Learner Corpus with the Common Text Analysis Platform

Xiaobin Chen, Detmar Meurers
University of Tübingen
xiaobin.chen@uni-tuebingen.de, dm@sfs.uni-tuebingen.de

This study demonstrates the use of an automatic text complexity analysis system—the Common Text Analysis Platform (CTAP, Chen & Meurers, 2016)—for analyzing a learner corpus of continuation writings on the purpose of identifying the aspects in which alignment occurs between the learner writings and the input texts.

Alignment is the general socio-cognitive process in which interlocutors coordinate with each other in a dynamic and adaptive way during a conversation to develop a common mental representation for successful communication (Pickering & Garrod, 2004). According to Pickering and Garrod's Interactive Alignment Model, alignment occurs on both the situational and linguistic levels, the latter of which manifests itself in structural priming or linguistic similarities between the parties involved in the communication. The alignment theory is not only accountable for L1 communicative interaction, but also L2 acquisition (Atkinson et al., 2007), because besides between human beings, alignment also happens between human beings and their social and physical environments.

Wang and Wang (2015) provided empirical evidence on the alignment between EFL learner production and the reading input through a series of continuation writing experiments. They found that students made significantly fewer grammar mistakes (more target like language) in continuation story writing tasks after reading the stories in English than when they were given the same stories in their mother tongue. Alignment also happened on the lexical level in terms of keyword overlap (p. 514) and on the morphological level manifested as more target like verb tense and plural nouns.

Supposedly, alignment should occur on various linguistic and structural levels between the learner production and the input they receive. However, the study by Wang and Wang (2015) did not test the other aspects besides error frequencies and keyword overlap. In this study, we propose a more comprehensive analysis of the same continuation writing corpus from Wang and Wang (2015) with the CTAP platform, aiming at finding whether the alignment effect is traceable from the other linguistic aspects.

The CTAP system (accessible at http://ctapweb.com) is a web-based platform that supports fully configurable linguistic feature extraction for a wide range of complexity analyses (Chen & Detmar, 2016). It integrates state-of-the-art Natural Language Processing (NLP) technologies and features a user-friendly interface, modularized and reusable analysis component integration, and flexible corpus and feature management. The latest version of the system contains more than 170 feature extractors capable of calculating lexical density, sophistication, variation, and syntactic complexity from user supplied corpora.

The learner corpus analyzed in the study is the continuation writings of 48 EFL students from Wang and Wang's (2015) study. Each student continued writing two stories after reading them with the endings removed. The treatment conditions were the language of

the input stories, either English (the target language) or Chinese (the mother tongue) in which the stories were written.

The corpus was imported into the CTAP system for extracting all the 173 textual features provided by the system. Shapiro-Wilk Normality Tests conducted for each feature and treatment condition helped identify 94 features that were normally distributed across the two treatment conditions. One-tail paired sample T-tests were then run on each of these features between writings with Chinese input and those with English input. Nineteen text features were found to yield significantly different measures across the English- and Chinese-treatment conditions. These were mainly lexical sophistication features measured by word frequency norms. No lexical variation, lexical density, or syntactic complexity measures were found to be significantly different between the two treatment conditions. The results of this study resonate with and further extend the findings from the original study by pinpointing the aspects of learner language that benefited from aligning to the authentic input. We also demonstrated a use case of the CTAP system which offers automatic extraction of comprehensive textual features from learner corpora.

Acknowledgments

This research was funded by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Xiaobin Chen is a doctoral student at the LEAD Graduate School & Research Network.

References

Atkinson, D., Churchill, E., Nishino, T., and Okada, H. (2007). Alignment and interaction in a sociocognitive approach to second language acquisition. *The Modern Language Journal*, 91(2), 169–188.

Chen, X.B. & Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity*, pp. 113–119, Osaka, Japan. The International Committee on Computational Linguistics.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190.

Wang, C. and Wang, M. (2015). Effect of alignment on L2 written production. *Applied Linguistics*, 36(5).

To parse or not to parse: the question of learner corpora

Elisa Corino, Cristina Bosco, Alessandro Mazzei
Università di Torino
elisa.corino@unito.it, bosco@di.unito.it , mazzei@di.unito.it

Modern learner corpora are now routinely PoS-tagged, whereas syntactic parsing is much less frequent. Nonetheless, an easier access to syntactic information could shed light onto particular phenomena which occur when dealing with learners' interlanguage.
Some attempts have been made so far especially for German (Nivre et al., 2007, Lüdeling 2008, Ott & Zai 2010) and English (Napoles et al., 2017), but the ground seems unexplored regarding the Italian panorama.
This contribution aims at discussing a preliminary study on suitable parsers for Italian as a foreign language, in an effort to identify key elements to be further used to parse such learner corpora in a proper way.
A complete error-free automatic parsing of learner language is unatteinable due to spelling and grammar mistakes (e.g. disagreement). However, parsing learner corpora has benefits for research on non native speakers' syntactic productions (e.g. its complexity or the avoidance of structure; Rosén & De Smedt 2010).
From a computational point of view, the technical difficulties of dealing with the learner varieties will be highlighted; whereas a qualitative linguistic analysis will show the possible use of a parsed learner corpus.
Starting from the data of a well-established learner corpus of Italian as L2 — VALICO (Corino/Marello 2017, Marello et al 2011, www.valico.org) — we will present a possible new structure for the PoS-tagged corpus, which includes some parsed sections. The pros and cons of parsing will be discussed with particular reference to the learner varieties and the difficulties related to the syntactic irregularity of interlanguage will be dealt with (eg. the use of gerund with attributive function).
The parsed sub-corpus of VALICO will be based on samples of texts of learners coming from different mother tongues, so that possible differences in the distribution of the observed phenomenon can be observed.
Dealing with the computational side of the research, a first attempt to apply a parser to VALICO (Corino/Russo 2016) has already brought to some results and has generated a first draft of the many hitches one has to face when a stochastic parser is applied to a learner corpus. Problems arise not only in relation to the verb or noun inflection, but emerge also where there is an accumulation of clitics, together with spellings that deviate from the norm (as in *averecela*). Wrong position might be an issue as well, as in *Il fratellino di Leo non capiva perchè e così lei ha spiegato glielo*, where the pronoun follows the verb instead of being in its canonical preverbal position.
The approach we propose is based on the application of a rule-based dependency parsing system, i.e. TULE (Lesmo 2007, 2009) which produces a morpho-syntactic annotation in the format of the Turin University Treebank. The best scores achieved by this parser on standard texts are Labeled Attachment Score 85.34 and Unlabeled Attachment Score 91.47 (Lesmo 2011). In order to morpho-syntactically process the sentences where an error

occurs, this will be provisionally replaced with the corresponding target form to be later replaced with the error itself. The fundamental idea is to *help* the parser to correctly analyze an ungrammatical sentence. So, in a first preliminary step the original sentence is modified by removing some ungrammatical morpho-syntactic fragments. Later, in a second step, the original ungrammatical morpho-syntactic fragments are inserted again in the final complete annotation of the sentence.

We will also consider the possibility to use a statistical parser trained on the Universal Dependency treebank released for Italian (http://universaldependencies.org/#it) in order to overcome the difficulties arising from the use of a rule-based parser, but also to work in the perspective of a format which is currently considered as a standard de facto and which has been applied to a large variety of languages. Moreover the application of two different annotation formats to our data may also give some hints about the usefulness of each of these formats for representing and detecting Italian learners' errors and give contribute to the investigation about this issue (Meurers/Dickinson, to appear 2017).

Despite the complexity of the operation, which necessarily requires a good deal of manual intervention, the parsing of a learner corpus can have positive impacts in more than one area, as it allows researchers to easily identify syntactic errors, deviations from the norm and distribution of categories and syntactic structures otherwise difficult to bring out querying a PoS-tagged corpus only.

From a linguistic point of view, a qualitative case study will deal with the relative position of nouns and adjectives within the NP. The study will discuss the difficulties the parser has had in processing the data and its tagging mistakes, with special reference to the order and the position of adjectives within the NP of leaners of typologically different L1, i.e. Spanish and French vs English and German. Possible differences in acquisition and features of the learners' interlaguage will thus be highlighted.

References:

Bosco, C., Dell'Orletta, F., Montemagni , S., Sanguinetti , M., Simi, M. (2014). The Evalita 2014 Dependency Parsing task. In C. Bosco, P. Cosi, F. Dell'Orletta, M. Falcone, S. Montemagni, M. Simi (a cura di), *Proceedings of the fourth International Workshop Evalita 2014,* Pisa.

Corino, E., Marello, C. (2017). *Italiano di stranieri. I corpora VALICO e VINCA*, Guerra, Perugia.

Corino, E., Russo, C. (2016). Parsing di corpora di apprendenti di italiano: un primo studio su VALICO. In A. Corazza, S. Montemagni, G. Semeraro (a cura di), *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016 5-6 December 2016, Napoli,* Accademia University Press.

Dickinson, M.; Ragheb, M. (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories* (TLT-8).

Lesmo, L. (2007). The rule-based parser of the NLP group of the University of Torino. Intelligenza Artificiale, 12.

Lesmo, L. (2009). The Turin University parser at Evalita 2009. In *Proceedings of Evalita'09*, Reggio Emilia

Lesmo, L. (2011). The Turin University Parser at Evalita 2011. In Evalita 2011 Working Notes, Roma

Lüdeling, A. et al. (2008). Syntactic Misuse, Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis

Lüdeling, A. et al. (2012). Das Falko-Handbuch Korpusaufbau und Annotationen Version 2.01.

Marello, C. et al (2011). I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte. In N. Maraschio, D. De Martino (a cura di), *La Piazza delle lingue L'italiano degli altri*. Firenze, 27-31 maggio 2010. Atti, Firenze, Accademia della Crusca, 2011 ("La Piazza delle lingue", 2). pp.49-61

Menzel, W.; Schröder, I. (1999). Error Diagnosis for Language Learning Systems [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.4723&rep=rep1&type=pdf]

Meurers, D., Dickinson, M. (to appear 2017). Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. Language Learning Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation. [http://www.sfs.uni-tuebingen.de/~dm/papers/Meurers.Dickinson-17-v14482.pdf]

Napoles, C. et al (2017). JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings of EACL2017*, Valencia

Nivre et al., (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. Natural Language Engineering 13(1), 1–41.

Ott, N.; Ziai, R. (2010). Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of the Ninth Workshop on Treebanks and Linguistic Theories* (TLT-9), Tartu.

Relating lexical differences and input variables in L1- and L2-acquisition of German by using the *Corpus Explorer* tool

Christine Czinglar[1], Katharina Korecky-Kröll[2,3], Lisa Buchegger[3], Jan Oliver Rüdiger[1]
University of Kassel[1], Austrian Academy of Sciences[2], University of Vienna[3]
christine.czinglar@uni-kassel.de, katharina.korecky-kroell@univie.ac.at, l.buchegger@univie.ac.at, e-mail@jan-oliver-ruediger.de

In first (L1) and even more so in second language (L2) acquisition we find great individual variation in the acquisition of the lexicon. One important factor in L1- and L2-acquisition is the socio-economic status (SES) of the parents, i.e. their highest education, job prestige and sometimes income (Hart & Risley 1995; Oller & Eilers 2002, Hoff 2003, Vasilyeva & Waterfall 2011). For bilingual children, amount and quality of input in the respective languages also play a big role, e.g. length and intensity of exposure, number of native speakers as conversational partners and input in childcare facilities (Hoff et al. 2012; Place & Hoff 2011; Mashburn et al. 2008; Unsworth 2016, 2013).
In this paper, we will investigate differences in the lexical abilities of 3-4-year-old children acquiring German as a second language in comparison to their monolingual peers: (i) What individual differences do we find in productive vocabulary (German spontaneous speech data in kindergarten) and in receptive vocabulary (German test data) in L1- and L2-acquisition? (ii) Which cluster of individual factors may explain these lexical differences? (iii) What is the relation between receptive and productive vocabulary, viz. test and spontaneous speech data?
In order to answer these questions, we will present data from an ongoing psycholinguistic project (INPUT): The database includes spontaneous speech and test data from 27 children successively acquiring German as L2 (L1 Turkish) and 29 monolingual L1-children. Interviews with parents and kindergarten teachers provide information on SES and on different input and other individual variables. The children were videotaped during spontaneous interactions with their caretakers 4 times over a period of 18-20 months in childcare facilities in Vienna. 30 minutes of spontaneous speech were transcribed for each session and coded for part-of-speech and morphology by using CHILDES/CLAN (MacWhinney 2000).
As our focus was to obtain ecologically valid data in child-care facilities, we did not control for communicative settings: The pedagogues and children could freely choose, what they wanted to do during the recording. But amount of speech and lexical diversity in interactions between children and caretakers differ with respect to communicative settings such as toy play, meal time or book reading (Hoff 2010; de Houwer 2009). So we devised a coding scheme for different types of communicative settings (e.g. morning circle vs. small groups, tutored vs. free activities, rule governed games, book reading) and coded the transcribed settings afterwards.
Coding on the word level, i.e. part-of-speech-tagging and morphological coding using a predefined lexicon, works well using CLAN. But coding more complex parts of speech, such

as phrases, clauses, interactions or even communicative settings, in CLAN is time-consuming and error-prone. To be able to combine both coding layers (morphological and communicative settings), we developed a procedure to integrate the morphological coding of CLAN with different kinds of higher-level coding in a new corpus linguistic tool called CorpusExplorer (Rüdiger 2016ab). The CorpusExplorer can also be used to perform in-depth analyses such as frequency over time, speaker influences and lexical diversity and richness (similarly to the tools in CLAN).

Previous analyses of our test data and input variables have shown significant influences of SES on L1 vocabulary, and of different input variables combined with SES on L2 vocabulary (Czinglar et al. 2015; Czinglar et al. in press). Previous analyses of our spontaneous speech data in kindergarten show that there are important differences regarding the quality of the input in different communicative settings, e.g. a significantly lower MLU in strongly tutored settings (e.g. handicraft work) or in strongly regulated settings involving board or card games (Templ et al. submitted).

In this talk, we will bring these different analyses together: receptive vocabulary from test data and productive vocabulary from spontaneous speech in different communicative settings and different variables regarding quantity and quality of the German input. This complex analysis becomes possible with the CorpusExplorer, which we will show to be a powerful tool to relate diverse sources of data and all types of text and metadata.

References:

Czinglar, Christine, Jan Oliver Rüdiger, Katharina Korecky-Kröll, Kumru Uzunkaya-Sharma & Wolfgang U. Dressler (in press). Inputfaktoren im DaZ-Erwerb von sukzessiv bilingualen Kindern mit L1 Türkisch. In: *Mehrsprachigkeit: Spracherwerb, Unterrichtsprozesse, Schulentwicklung. Beiträge zum 11. Workshop Kinder und Jugendliche mit Migrationshintergrund.* Hrsg. von Isabel Fuchs, Stefan Jeuk & Werner Knapp. Stuttgart, Fillibach bei Klett.

Czinglar, Christine, Katharina Korecky-Kröll, Kumru Uzunkaya-Sharma & Wolfgang U. Dressler (2015): Wie beeinflusst der sozioökonomische Status den Erwerb der Erst- und Zweitsprache? Wortschatzerwerb und Geschwindigkeit im NP/DP-Erwerb bei Kindergartenkindern im türkisch-deutschen Kontrast. In Klaus-Michael Köpcke & Arne Ziegler (Hrsg.): *Deutsche Grammatik in Kontakt. Deutsch als Zweitsprache in Schule und Unterricht*. Berlin: De Gruyter, 207-240.

De Houwer, Annick (2009): *Bilingual first language acquisition*. Bristol, Buffalo, Toronto: Multilingual Matters.

Hart, Betty & Todd R. Risley (1995): *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes.

Hoff, Erika (2003): The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development* 74 (5): 1368-1378.

Hoff, Erika (2010): Context effects on young children's language use: The influence of conversational setting and partner. *First Language* 30 (3-4): 461–472.

Hoff, Erika, Cynthia Core, Silvia Place, Rosario Rumiche, Melissa Señor & Marisol Parra (2012). Dual language exposure and early bilingual development. *Journal of Child Language* 39(01): 1-27.

MacWhinney, Brian (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, New Jersey: Erlbaum.

Mashburn, Andrew J., Robert C. Pianta, Bridget K. Hamre, Jason T. Downer, Oscar A. Barbarin, Donna Bryant, Margaret Burchinal, Diane M. Early & Carollee Howes (2008): Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development* 79 (3): 732-749.

Oller, D. Kimbrough & Rebecca E. Eilers (Hrsg.) (2002): *Language and Literacy in Bilingual Children*. Clevedon, Multilingual Matters.

Place, Silvia & Erika Hoff (2011). Properties of Dual Language Exposure That Influence 2-Year-Olds' Bilingual Proficiency. *Child Development* 82(6): 1834-1849.

Rüdiger, Jan Oliver (2016a): Corpus Explorer V2.0. http://www.corpusexplorer.de [Computer software] – Free and OpenSource.

Rüdiger, Jan Oliver (2016b): Korpushermeneutik: Ansatz und Werkzeug zur Analyse großer Textkorpora. Paper presented at DHd 2016 (Tagung der digital humanities im deutschsprachigen Raum), University of Leipzig, and distinguished with the Lisa Lena Opas-Hänninen Young Scholar Prize 2016.

Templ, Viktoria , Maria Weichselbaum, Katharina Korecky-Kröll & Wolfgang U. Dressler (submitted, 2017). Deutschspracherwerb ein- und zweisprachiger Wiener Kindergartenkinder. Der Einfluss des sozioökonomischen Status der Familie, des sprachlichen Hintergrunds und der Sprechsituationen. In: *Jahrbuch der Kommission für Migrations- und Integrationsforschung*. Hrsg. von. Göttingen, V&R unipress, Vienna University Press.

Unsworth, Sharon (2016). Early child L2 acquisition: Age or input effects? Neither, or both? *Journal of Child Language* 43(3): 608-634.

Unsworth, Sharon (2013a): Current Issues in Multilingual First Language Acquisition. Annual Review of Applied Linguistics 33,: 21-50.

Vasilyeva, Marina & Heidi Waterfall (2011): Variability in language development: Relation to socioeconomic status and environmental input. In Susan B. Neuman & David K. Dickinson (Hrsg.): *Handbook of early literacy research*. New York: Guilford Press. 3, 36-48.

Looking for common ground across globalized English varieties: A multivariate exploration of mental predicates

Sandra C. Deshors[1], Sandra Götz[2]
Michigan State University[1], Justus Liebig University Gießen[2]
sandracdeshors@gmail.com, sandra.goetz@anglistik.uni-giessen.de

At a time when English has become a world-wide language shaped by globalization, this study focuses on variety formation and adds to the discussion on the developmental pathways that characterize the evolution of non-native Englishes (EFL/ESL) in the 21st century. As a result of the rapid globalization of English, unifying (rather than distinguishing) theoretical approaches to world Englishes, such as Mair (2013) and Schneider (2014) have started to emerge. In this context, we raise the question how variety formation of globalized Englishes can be best approached within a unifying theoretical framework and to what extent linguistic features common to EFL/ESL varieties reflect a developmental pattern of globalization. Inspired by Mufwene's (2001) 'feature pool', we account for the fact that, through the media and international trade, speakers of world Englishes have access to – and input from – an international mix of native (ENL) and ESL varieties world-wide, facilitating new contact situations and a new mix of features that can become characteristic of newly emerging varieties (Schneider 2011).
While semantic investigations of ENLs, ESLs/EFLs have so far remained relatively rare and focused on the meanings of isolated lexemes, we take a unifying perspective to explore the uses of the near-synonymous mental predicates *I believe*, *I think*, *I suppose* and *I guess* across 8 native and ESL/EFL varieties. We explore those predicates by unpacking their semantic structure, following Krawczak (2015), who contrasts British and American Englishes and explores how mental predicates are construed in use, what functional components characterize their individual usage profiles and whether variation in their usage patterns is observed across speaker populations exposed to different socio-cultural contexts. While such contexts have been shown to lead to intralinguistic differences across ESLs, we test for the existence of invariant semantic patterns that transcend the language-contact situations of Singapore and India, as more established ESL varieties on the one hand, and Hong Kong English as a variety that has been discussed to stand between EFL and ESL status on the other. Given this background, we specifically explore whether

i. a semantic approach to cross-varietal variation can help us improve our understanding of what unifies English varieties world-wide;
ii. ii. ESL varieties at different stages of nativization demonstrate variation patterns that differ from those observed in varieties other than their historical input variety; and
iii. iii. whether mental predicates represent stable linguistic features across different native and ESL varieties.

Methodologically, we adopt a multivariate statistical technique (classification and regression tree analysis) to model the uses of 1,125 contextualized occurrences of our four

lexemes extracted from the *International Corpus of English* (ICE; specifically, the Great Britain, Ireland, New Zealand, America, Canada, Singapore, India and Hong Kong ICE sub-components) annotated for English variety, written genres (correspondence, press-editorials, creative, instructional, popular, student writing) and seven semantic variables (epistemic type, epistemic mode, epistemic class, argumentativity, verifiability, evaluation and negotiability). Overall, our findings show the usefulness of exploring invariant linguistic patterns across Englishes through the lens of their semantic structure. Although, on the surface, two groups of English varieties emerge with different preferential patterns of predicates (British, Indian, Irish and Singapore vs. Canadian, Hong Kong and American), at a more abstract level, those predicates share similar semantic combinatory patterns common to all varieties in focus. Further, our analysis confirmed that, as abstract semantic constructions, mental predicates can be approached as stable invariant features present in all the ESL varieties we investigated. Our results also suggest the existence of two different developmental pathways in the development of HKE and IndE, based on *think* and *believe*. Methodologically, it emerges that multivariate techniques can unveil the complex (semantic) structure that mental predicates hide across ESLs. With regard to Mufwene's (2001) 'feature pool', this metaphor allows us to identify the semantic structure of think constructions as a stable linguistic construct that transcends the language-contact situations of the ESL varieties in focus. Further, it helps us to show that invariant features are not just categorical in nature but can be present in different degrees which impacts on the shape of ESL varieties. Altogether, our results indicate that variety formation is a dynamic multidirectional process involving developmental paths both towards as well as away from native varieties and modeling this process using theoretical frameworks that account for the simultaneous development of generic (i.e. common to all Englishes) as well as specialized (i.e. specific to individual Englishes) linguistic patterns may be beneficial.

References

Krawczak, Karolina. 2015. "Near-synonymous epistemic stance predicates in English: A quantitative corpus-driven study of subjectivity". In: D. Glynn & M. Sjölin (eds.) *Subjectivity and Epistemicity. Stance strategies in discourse and narration*. Lund: Lund University Press, 311–339.

Mair, Christian. 2013. "The World System of Englishes: Accounting for the Transnational Importance of Mobile and Mediated Vernaculars", *English World-Wide* 34(3): 253-278.

Mufwene, Salikoko S. 2001. The Ecology of Language Evolution. Cambridge: Cambridge University Press.

Schneider, Edgar W. 2011. *English Around the World - An Introduction*. Cambridge: Cambridge University Press.

Schneider, E. W. 2014. "'Transnational Attraction': New reflections on the evolutionary dynamics of World Englishes", *World Englishes* 33(1): 9–32.

The progressive vs. non-progressive alternation: Non-native Englishes through the lens of collostructional analysis

Sandra C. Deshors[1], Paula Rautionaho[2]
Michigan State University[1], University of Tampere[2]
sandracdeshors@gmail.com, paurau01@gmail.com

This study investigates the progressive vs. non-progressive alternation in seven native (ENL) and non-native (EFL and ESL) varieties. Focusing on the semantic differences between the two alternating constructions, we model English learners' decision patterns when faced with the choice of opting for a progressive or non-progressive marking. Specifically, we address the following questions:

1. which lexical verbs, semantic domains and Aktionsart categories significantly attract the progressive and the non-progressive, respectively?
2. what are the exact degrees of association between those verbs, semantic domains and Aktionsart categories and (non-)progressive constructions?
3. to what extent do those associations vary systematically across ENL, EFL and ESL and across written genres?

Although ENL and EFL/ESL writers are known to differ in their uses of the progressive (Hundt & Vogel 2011; Rautionaho 2014), this difference is often based on frequency counts of progressive constructions. Recently, however, Deshors (2017) demonstrated the usefulness of reaching beyond 'over-' vs. 'underuses' approaches and focusing on assessing degrees of association between progressive constructions and their co-occurring linguistic features to unveil unattested L2 usage patterns. Building on Deshors' approach, we extend her analysis by (i) integrating the progressive vs. non-progressive alternation into our analysis and (ii) investigating a wider range of non-native varieties. Further, our collostructional analysis consists of a Distinctive Collexeme Analysis (DCA) instead of a co-varying collexeme analysis. This adjustment allows us to assess systematically the degrees of association between lexical verbs, semantic domains and Aktionsart categories on the one hand, and progressive and non-progressive constructions on the other hand. Concretely, we investigate over 7,000 progressive and non-progressive constructions in seven comparable corpora, the ICE Great Britain, USA, Ireland, India, Singapore and Nigeria, in addition to the recently released Corpus of Dutch English (Edwards 2014). Because the latter corpus follows the ICE design, it provides a valuable opportunity to investigate an as yet virtually unexplored population of EFL users whose linguistic background is historically and sociolinguistically unrelated to any of the English varieties traditionally covered in such analyses. Our approach consists of successive DCAs conducted for the variables LEMMA and CONSTRUCTION, then SEMANTIC DOMAIN and CONSTRUCTION, and finally AKTIONSART and CONTRUCTION, in order to identify the specific lexical items, semantic domains and Aktionsart categories that are distinctively associated with progressive and non-progressive constructions within each variety in focus. Overall, the results show, rather expectedly, that Existence verbs and States are associated

with the non-progressive construction, and that Activity verbs and Processes are associated with the progressive construction in most varieties. It is in the less central categories (e.g. Causative and Occurrence verbs), however, that we find the greatest variation. Interestingly, the DCA reveals that, despite its stative nature, the verb LIVE is significantly and systematically associated with the progressive construction in ICE-GB, ICE-IND and ICE-US, and States are not associated with the non-progressive construction in ICE-SIN. Further, we find evidence supporting the possible overextension of the progressive to non-delimited stative verbs in ESL varieties, as lemmas HAVE and KNOW do not associate with the non-progressive in these varieties, while they do so in ENL and EFL. Overall, our approach reveals fine-grained patterns of progressive and non-progressive usage that remain obscured in traditional frequency-based analyses, thus yielding more nuanced insight into the (dis)similarities among and within ENL, ESL and EFL.

References

Deshors, S. C. (2017). Zooming in on verbs in the progressive: A collostructional and correspondence analysis. *Journal of English Linguistics.*

Edwards, A. (2014). *English in the Netherlands: Functions, forms and attitudes*. Unpublished doctoral dissertation. University of Cambridge.

Hundt, M. & K. Vogel. (2011). Overuse of the progressive in SL and learner Englishes – fact or fiction? In J. Mukherjee & M. Hundt (Eds.). *Exploring Second Language Varieties of English and Learner Englishes, Bridging the paradigm gap.* Amsterdam: John Benjamins, 145-164.

Rautionaho, P. (2014). *Variation in the Progressive: A Corpus-Based Study into World Englishes*. Tampere: Tampere University Press.

Informing linguistic competence descriptors at CEFR A2 and B1 levels: insights from a fully-error tagged learner corpus by Spanish learners of English

María Belén Díez-Bedmar
University of Jaén
belendb@ujaen.es

Due to the overriding philosophy of the CEFR (Council of Europe, 2001), i.e. providing users with a document which may trigger reflection on the learning, teaching and assessment of any language and provide a common standard, the well-known descriptors or 'can-do statements' are written in such a way that they can be applied to any language. Consequently, their use may pose problems to users (Morrow, 2004; Hulstijn, 2007; North, 2014), as the ones for linguistic competence lack linguistic information on the language expected at each level, i.e., the quality of the language (Hulstijn, 2007). The consequent need to revisit the descriptors has motivated studies in which leaner corpus results complement the descriptors with information on the language used by learners at each CEFR (*The English Profile Project*; Díez-Bedmar, 2010, 2015; Hawkins & Filipović, 2012; Thewissen, 2013; Götz, 2015). However, the number of studies so far is still limited and most of them do not focus on one L1 learner group or restrict the findings to a restricted number of linguistic aspects.

To bridge this gap in the literature, this paper uses the information in a fully-error tagged learner corpus by Spanish learners of English to inform competence descriptors at CEFR A2 and CEFR B1 level by: a) describing 'negative grammatical features' at both levels; and b) revealing the criterial features found between A2 and B1 levels in Spanish learner writing. A Computer-aided Error Analysis (CEA) (Dagneaux, Denness & Granger, 1998) using the error-taxonomy in Dagneaux, Denness, Granger & Meunier (1996) was performed on a learner corpus composed of 80 compositions (40 at each level, 10119 words, 2004 errors analysed). After revealing the most frequent errors at each level, non-parametric tests were run to find out if there was any criterial feature which would distinguish both levels. The CEA results reveal that, at A2 level, there are nine error types which show a mean higher than one error per composition, the two highest error types per composition being spelling and vocabulary selection. At B1 level, the number of error types whose mean is higher than one per composition decreases to four, the errors with the highest means being the same as at A2 level. The results of the non-parametric tests show a statistical decrease in the mean of errors per composition from A2 to B1 in fifteen error-types.

The findings obtained have been used to fine-tune the phrasings of the competence descriptors (with a special focus on accuracy descriptors) for Spanish learners of English at B2 and B1 levels by including information on the aspects of language which improve from one level to the next. In a broader sense, this paper is also an example of the way in which LCR results may be used to inform LTA by providing valid and reliable descriptors which are clear to all CEFR users, SLA, as the errors at A2 and B1 levels are described and FLT, since

materials writers, teachers, etc. may become aware of the real limitations of Spanish students at these levels.

References:

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided Error Analysis. *System*, 26: 163-174.

Dagneaux, E., Denness, S., Granger, S., & Meunier, F. (1996). *Error Tagging Manual Version 1.1*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Díez-Bedmar, M. B. (2010). Analysis of the written expression in English in the University Entrance Examination at the University of Jaén. Unpublished PhD Dissertation. Universidad de Jaén.

Díez-Bedmar, M. B. (2015). Article use and criterial features in Spanish EFL writing: a pilot study from CEFR A2 to B2 levels. In M. Callies and S. Götz (Eds.). *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins, 163-190.

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

Götz, S. (2015). Tense and aspect errors in spoken learner English. Implications for language testing and assessment. In M. Callies and S. Götz (Eds.). *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins, 191-215.

Hawkins, J. A., & Filipović, L. (2012). *Criterial Features in L2 English*. Cambridge: CUP.

Hulstijn, J. H. (2007). The Shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91: 663-667.

Morrow, K. (Ed.) (2004). Insights from the Common European Framework. Oxford: OUP.

North, B. (2014). *The CEFR in Practice*. Cambridge: CUP.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1): 77-101.

Complexity in NPs in learner writing: a cross-sectional comparative study of Spanish and Israeli learners of English

María Belén Díez-Bedmar[1], Pascual Pérez-Paredes[2]
University of Jaén[1], University of Cambridge[2]
belendb@ujaen.es, pfp23@cam.ac.uk

The analysis of the Noun Phrase (NP) in learner writing has focused on aspects such as signalling nouns (Flowerdew, 2006, 2010), countable and uncountable nouns (Kobayashi, 2008), articles and noun count nouns (Osborne, 2004) or article use (Díez-Bedmar, 2015; Díez-Bedmar & Pérez Paredes, 2012; Leńko-Szymańska, 2012). However, little information is available so far regarding the complexity of the NPs used by learners at different levels. The analysis of complexity in NPs in learner writing at different levels and by learners from different L1s is necessary as it may contribute to SLA and other related disciplines, such as Language Testing and Assessment (LTA) and Foreign Language Teaching (FLT).
This paper sets out to answer the following research questions: a) which are the NP types (used by learners with different L1s?; b) does the use of such NP types vary at different levels?; and c) do NP types play a role in the characterization of learner writing at different level, i.e., criterial features (Hawkins & Filipović, 2012)?
Two learner corpora, part of the ICCI corpus (Tono and Díez-Bedmar, 2014), were used to provide a cross-sectional analysis of the syntactic complexity in NPs in learner writing by Spanish and Israeli secondary school students of English (8473 and 6225 words, respectively). A combination of both learner corpus analysis and syntactic complexity measures was used. To select the NPs analysed, the most frequently used nouns in each corpus were identified using Wmatrix (Rayson, 2009) and manually analysed. As a result, 635 NPs were manually parsed, which resulted in a classification of NP types in learner writing. Syntactic complexity analyses of noun complexity were run with TAASC 1.0 (Kyle, 2016).
Among the most interesting results, the number of NP types employed varies in each learner group: in the case of bare NPs or NPs accompanied by one or more determiners and premodified NPs, Israeli learners use more NP types than Spanish learners. The opposite is found in postmodified NP types and premodified and postmodified NP types. The number of NP types used varies from the lower to the higher grades. The main differences are found in postmodified NPs as well as in premodified and postmodified NPs, as Spanish learners use more postmodified NPs in Grade 11, whereas the opposite is found in Israeli learner writing. The non-parametric tests run to test if the types of NPs employed by the learners at each level would characterise learner writing at those levels reveal that NP complexity, as seen in the types of NPs used, remains the same at both levels. Syntactic complexity analyses of noun complexity confirmed such finding.

References:

Díez-Bedmar, M. B. (2015). Article use and criterial features in Spanish EFL writing: a pilot study from CEFR A2 to B2 levels. In M. Callies & S. Götz (Eds.). *Learner Corpora in Language Testing and Assessment*. Amsterdam: John Benjamins, 163-190.

Díez-Bedmar, M.B. & Pérez Paredes, P. (2012). A cross-sectional analysis of the use of the English articles in Spanish learner writing. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.). *Developmental and Crossslinguistic Perspectives in Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins, 139-157.

Flowerdew, J. (2010). Use of signalling nouns across L1 and L2 writer corpora. *International Journal of Corpus Linguistics*, 11(3), 345–362.

Flowerdew, J. (2006). Signalling nouns in a learner corpus. *International Journal of Corpus Linguistics*, 11(3), 345–362.

Hawkins, J. A. & Filipović, L. (2012). *Criterial Features in L2 English*. Cambridge: CUP.

Kobayashi, T. (2008). Usage of countable and uncountable nouns by Japanese learners of English: two studies using the ICLE error-tagged Japanese sub-corpus. *National Institute of Informatics Scholarly and Academic Information Navigator*, 816(10), 73–82.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Doctoral Dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35.

Leńko-Szymańska, A. (2012). The role of conventionalized language in the acquisition and use of articles by Polish EFL learners. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.). *Developmental and Crossslinguistic Perspectives in Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins, 83-103.

Osborne, J. (2004). Articles and non-count nouns in learner English: perception and production. In B. Lewandowska-Tomaszczyk (Ed.). *Practical Applications in Language and Computers (PALC 2003)*. Frankfurt: Peter Lang, 359-369.

The effects of speaking task on L2 fluency

Amandine Dumont
Université catholique de Louvain
amandine.dumont@uclouvain.be

Learner (like native speaker) fluency is characterised by an interplay between speed, breakdown and repair phenomena (e.g. De Jong 2016; Skehan 2003; Tavakoli & Skehan 2005). These include, among others, filled and unfilled pauses, reformulations, repetitions and temporal features such as speech rate. Research has provided valuable insights into the way these features interact with one another or are affected by factors such as mother-tongue background, time spent in an L2 environment or proficiency level (e.g. Freed 1995; Ginther, Dimova & Yang 2010; Götz 2013; Guz 2015; Towell, Hawkins & Bazergui 1996). With the advent of larger and more varied spoken corpora, a growing number of studies has turned to investigate how fluency might also be impacted by speaking task features such as task type (narrative, read-aloud task etc.), planning or level of interaction. Overall, research findings have revealed that variations in fluency measures can indeed partly be attributed to task type (e.g. Cucchiarini, Strik & Boves 2002; Foster & Skehan 1996) and that planning can have beneficial effects on learners' fluency performance (e.g. Crookes 1989; Foster & Skehan 1999; 2009; Mehnert 1998). With respect to the level of interaction, research indicates that speakers seem to be more fluent in dialogues than in monologues in terms of speed, length of silent pauses and repair measures (Tavakoli 2016; see also Ejzenberg 2000).

Against this backdrop, this study is an attempt to provide a more in-depth understanding into the extent to which speaking task influences the production of fluency features of French-speaking learners of English, and of B2 vs. C1 learners more particularly. For this purpose, filled and unfilled pauses, restarts, as well as speech rate and mean length of runs – which are all mentioned in the CEFR fluency descriptors (Council of Europe 2001) – are examined across three speaking tasks.

The study is based on the French component of the *Louvain International Database of Spoken English Interlanguage* (Gilquin, De Cock & Granger 2010). Like the other components, LINDSEI-FR contains interviews of 50 intermediate to advanced French-speaking learners of English as a foreign language. Each interview is made up of three speaking tasks. First, a set topic is presented to the learner, who has some time to prepare what he/she is going to say. This task is followed by a spontaneous free discussion on various topics and, finally, by a monologic picture description task. These speaking tasks differ along a number of variables, including the degree of elicitation, of preparedness, as well as interactivity, which, based on previous research, are all likely to impact on the learner's fluency. LINDSEI-FR interviews have been time-aligned at the level of words – a procedure that allows for the precise measurement of temporal variables such as speech rate or length of pauses – and subsequently semi-automatically annotated for a wide range of (dis)fluency features, including, among others, filled and unfilled pauses, restarts, repetitions, and discourse markers. Moreover, the fluency of each LINDSEI-FR learner has been assessed by three professionally-trained raters according to the CEFR grid and

descriptors (inter-rater agreement reached an acceptable level). The grades were pooled and mean ratings computed per speaker.

Results so far indicate that the frequency of unfilled pauses, speech rate as well as the mean length of runs are particularly affected by task properties, while the rate of filled pauses and restarts does not differ significantly across tasks. The monologic picture description task is, for example, characterized by a higher frequency of unfilled pauses and a lower speech rate as compared to the set topic and the free discussion task. The pre-planned set topic task differs from the spontaneous free discussion with respect to the mean length of runs. These result thus partly corroborate previous findings. Preliminary investigations into the fluency of B2 (*n*=22) and C1 (*n*=26) learners indicate that, although higher CEFR fluency level leads to better performances in the three tasks (i.e. fewer pauses and restarts, faster speech rate and longer speech runs), the previously mentioned differences between the tasks remain. Comparable analyses based on native speaker data from LOCNEC (the *Louvain Corpus of Native English Conversation*; De Cock 2004), LINDSEI's native speaker counterpart, indicate that speaking task characteristics also influence the fluency of native speakers, though not to the same extent as for the learners.

References:

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. 3rd printing. Cambridge: Cambridge University Press.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(4), 367-383.

Cucchiarini, C., Strik H., & L. Boves (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 111(6), 2862-2873.

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

De Jong, N. (2016). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.). *Handbook of Second Language Assessment*. Berlin, Boston: Mouton de Gruyter, 203-218.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.). *Perspectives on Fluency*. Ann Arbor: University of Michigan Press, 287-313.

Foster, P. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3), 299-323.

Foster, P. & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, 3(3), 215-247.

Foster, P. & Skehan, P. (2009). The influence of planning and task type on second language performance. In K. Van den Branden, M. Bygate & J.M. Norris (Eds.). *Task-based Language Teaching: A Reader*. Amsterdam: Benjamins, 275-300.

Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.). *Second Language Acquisition in a Study Abroad Context* (Studies in Bilingualism 9). Amsterdam & Philadelphia: John Benjamins, 123-148.

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

Ginther, A., Dimova, S., & R. Yang (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399.

Götz, S. (2013). How fluent are advanced German learners of English (perceived to be)? Corpus findings vs. native-speaker perception. In M. Huber & J. Mukherjee (Eds.). *Studies in Variation, Contacts and Change in English*. Giessen: University of Giessen.

Guz, E. (2015). Establishing the fluency gap between native and non-native-speech. *Research in Language*, 13(3), 230-247.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20(1), 83-108.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1-14.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 133-150.

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.). *Planning and Task Performance in a Second Language*. Language Learning & Language Teaching 11. Philadelphia: John Benjamins Publishing Company, 239–73.

Towell, R., Hawkins, R., & N. Bazergui. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84-119.

Error analysis in a speech corpus of Spanish learners of English as a foreign language

Patricia Elhazaz Walsh
Universidad CEU San Pablo
pelhazaz@ceu.es

This study is aimed at compiling a learner corpus of read speech to measure the oral reading fluency (ORF) and pronunciation errors in leveled texts by Spanish children who are learning English as a foreign language (EFL).

Oral reading fluency has been defined as the ability to read a level text accurately, at a natural speaking rate, and with proper expression (National Reading Panel, 2000; Meyer & Felton, 1999). Three components of oral reading fluency have generally been studied: accuracy (number of words read correctly), speech rate (words per minute) and expressiveness. Of these, accuracy and speech rate have been the most studied as primary indicators of oral reading fluency. Words read correctly per minute (wcpm) is a widely used metric for assessing ORF from the perspective of reading rate and accuracy (Good & Kaminski, 2001).

This concept of reading fluency is based on the theory of automaticity in reading (LaBerge & Samuels, 1974) which states that automatic decoding of words is necessary to enable students to read more accurately and rapidly. Most studies agree that fluency in reading is based on the accuracy and speed of word identification (Breznitz, 2006).

While oral reading fluency (ORF) does not measure comprehension directly, there is substantial evidence that estimates of ORF can predict future reading performance and correlate strongly with reading comprehension (Fuchs et al., 2001; Jenkins et al., 2003). Knowing which words the student cannot decode properly can be very relevant for instructional purposes in an L2 and EFL context, as skills such as word recognition in the reading of a second language can be influenced by first language orthographic features. According to the orthographic depth hypothesis, speakers of a transparent language such as Spanish can experience difficulties in decoding words when reading in English, which has an opaque orthography (Basetti, 2009).

A learner corpus of read speech developed specifically for this study was collected in Madrid (Spain). The corpus comprises 528 one-minute reading sessions (8 hours and 48 minutes) read by 176 students in 5th, 6th (primary education) and 1st (secondary education) grades. Each of the students read three different short passages aloud obtained from the DIBELS test (Good, Kaminski, Smith, Laimon, & Dill, 2001). We made sure the level of vocabulary and grammar in the stories matched the English proficiency level of the students.

All recorded text passages were segmented, aligned with the sound signal and orthographically transcribed by a trained bilingual speaker.

Additionally, two analysts, a native English speaking expert and a trained bilingual speaker marked all the reading errors in the corpus and oral reading fluency was measured by calculating the words read correctly per minute (wcpm).

Given that annotating non-native pronunciations of children is not an easy task, a training session was scheduled before the actual annotation so that both experts would develop a satisfactory level of agreement. It was agreed to label as errors those realizations which had been identified in contrastive studies as common pronunciation difficulties for Spanish learners of English (Mott, 2011). A total of 4.692 reading errors were identified. Intertranscriber agreement coefficients were calculated and the speech data was used to train the acoustic models of FLORA,a speech recognizer developed for the automatic assessment of oral reading fluency (Bolaños, D., Cole, R.A., Walsh, P.E., y Ward, W.H. 2012). So far, we have phonetically transcribed and classified 1.347 pronunciation errors according to a typology created for the present study. An external annotator is also verifying a subset of the errors (10%) in order to assess the accuracy of our transcription. Overall, students had more difficulty with the pronunciation of those phonemes which do not have a similar sound in their L1 sound inventories. They also showed decoding difficulties reading those words that can be considered to have an opaque orthographic representation as a consequence of a mismatch between the English and Spanish grapheme-conversion rules.

Our study aims to provide data concerning the types of decoding difficulties that Spanish students have in reading tasks. Our results could be used to design classroom instructional interventions as well as computer assisted pronunciation teaching tools to help Spanish students to improve their pronunciation and decoding skills in English as a foreign language.

References:

Bassetti, B. (2009). Orthographic Input and Second Language Phonology. In: Thorsten Piske & Martha Young-Scholten (eds.): *Input Matters in SLA. (Multilingual Matters*), 191-206.

Bolaños, D., Cole, R.A., Walsh, P.E., y Ward, W.H. (2012). Automatic assessment of oral reading fluency for Spanish speaking ELs. WOCCI.

Breznitz, Z. (2006). Fluency in reading: Synchronization of process. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral reading fluency as an indicator or reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–256.

Good R., Kaminski R., Smith S., Laimon D, & Dill S. (2001). *Dynamic indicators of basic early literacy skills*. 5th ed. Eugene, OR: University of Oregon.

Jenkins, J., Fuchs, L., van den Broek, P., Espin, C., & Deno, S. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research and Practice*, 18, 237-245.

LaBerge D, Samuels S. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology.

Meyer, M. S. & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia*, 49, 283-306.

Mott, Brian (2011a). *English Phonetics and Phonology for Spanish Speakers*. 2nd.ed. Barcelona : Publications i Edicions UB.

National Reading Panel. (2000). Report of the National Reading Panel: Teaching children to read. Report of the subgroups. Washington, DC: US Department of Health and Human Services, National Institutes of Health.

The Acquisition of the /w/-/v/ contrast by German-speaking Learners of English – A Case of Category Goodness Assimilation

Robert Fuchs
Hong Kong Baptist University
rfuchs@hkbu.edu.hk

Most studies based on learner corpora focus on the written word or written representations of spoken language, which allows investigations into the syntax and lexis, among others, of learner language. Phonology and pronunciation, by contrast, are more rarely studied in this field, likely due to the fact that comparatively few phonological learner corpora are available. Nevertheless, pronunciation is crucial in spoken communication, and learner corpus research can make a vital contribution to our understanding of learner phonology.

This study focuses on the production of the English contrast between the labio-velar approximant /w/ and the voiced labiodental fricative /v/, which has been described as particularly challenging for German-speaking learners of English (Kresta 2015: 127, Swan & Smith 2011: 39). The closest German equivalent, /v/, though often described as a labiodental fricative, differs from both the English phonemes.

The analysis aims to determine first of all whether L1 German learners of English maintain the contrast. A related question is whether learners who receive more input or phonetic training improve in accuracy in the production of the /w/-/v/ contrast. Secondly, if the learners realise the contrast in some way, it is still far from certain whether they do so through the same acoustic cues that are used in L1 English. From an articulatory and acoustic point of view, the contrast is relatively complex in that it involves several dimensions. /w/ is an approximant and involves both rounding of the lips and a velar constriction, but no frication. /v/, on the other hand, is a fricative. For learners with German as their first language, this contrast might be particularly challenging because the articulation of German /v/, although usually described as a fricative, in fact involves little or no frication (Hamann & Sennema 2005a, 2005b, Scherer & Wollmann 1986: 93).

This constellation, where a non-native contrast has a single phoneme in the learners' L1 as closest counterpart, is also of interest for speech learning theories. These theories, in turn, might allow us to explore the reasons why German-speaking learners find English /w/-/v/ so challenging. First of all, Flege's (1995) Speech Learning Model (SLM) states that speech sounds in the L1 and L2 that are similar to each other are harder to learn than those that are identical or different from each other. Further, the Perceptual Assimilation Model (Best 1994, 1995; Best & Tyler 2007) describes a constellation where an L2 contrast is assimilated to a single L1 phoneme as single-category assimilation. However, if the L1 phoneme is somewhat more similar to one of the phonemes of the L2 contrast, the dissimilar member of the contrast is predicted to be easier to learn. If the predictions of either the SLM or PAM are borne out by the analysis, this would (contra Walker 2010: 109) suggest that

spelling is perhaps only a contributing and not the principal factor that makes English /w/-/v/ problematic for German-speaking learners.
The analysis relies on data from the LeaP Corpus (Gut 2014), with 20 learners, 7 L1 speakers of German and 4 L1 speakers of English, who produced 283 /v/ and 990 /w/ phonemes. Four acoustic correlates were measured in Praat (Zero Crossing Rate, spectral centroid, F2, F3). The statistical analysis, based on mixed effects regression models in R, reveals a general proficiency effect contributing to target-like pronunciation. In addition, it shows that German /v/ is more similar to the English fricative /v/ than the English approximant /w/. The PAM would thus predict that the more dissimilar sound, English /w/, is easier to learn. While the analysis suggests that even very advanced learners struggle with the target-like production of the /w/-/v/ contrast, all learner groups are more successful at the production of /w/, which is the more dissimilar member of the contrast. This confirms a hypothesis from Best's Perceptual Assimilation Model, which states that the more dissimilar member of a non-native contrast (compared to the closest sound in the learner's L1) is easier to acquire. Pronunciation training or a stay abroad turned out to have a negligible effect on the target-likeness of the learners' productions of /v/ and /w/. The paper concludes with recommendations on how the results of this acoustic study can be applied to pronunciation teaching.

References

Best, C. T. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. In H. C. Nussbaum (ed.), The development of speech perception: The transition from speech sounds to spoken words. Cambridge: MIT Press, 167-244.

Best, C. T. 1995. A direct realist view of cross-language speech perception. In W. Strange (ed.), Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research. Timonium: York Press, 171-204.

Best, C. T. & M. D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In J. Munro & O.-S. Bohn (eds.), Language experience in second language speech learning. In honor of James Emil Flege. Amsterdam/Philadelphia: Benjamins, 13-34.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 229-273). Timonium, MD: York Press.

Gut, Ulrike. 2014. *The LeaP Corpus*. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *The Oxford handbook of corpus phonology*, 509-516. Oxford: Oxford University Press.

Hamann, S., & Sennema, A. (2005a). Acoustic differences between German and Dutch labiodentals. In C. Geng, J. Brunner, & D. Pape (Eds.), *ZAS Papers in Linguistics, 42* (pp. 33–41). Berlin: Zentrum für Allgemeine Sprachwissenschaft.

Hamann, S., & Sennema, A. (2005b). Voiced labiodental fricatives or glides—all the same to Germans? In *Proceedings of the ISCA Workshop on Plasticity in Speech Perception* (pp. 164–167). London: University College London.

Kresta, Ronald. 2015. "Aussprachefehler von deutschen und deutschsprachigen Studierenden technischer Studiengänge in englisch-sprachigen Fachvorträgen–Eine

empirische Untersuchung." In Ines-Andrea Busch-Lauer, ed. *Facetten der Fachsprachenvermittlung Englisch–Hands on ESP Teaching*: 113-136.
Scherer, G., & Wollmann, A. (1986, 3rd ed.). *Englische Phonetik und Phonologie*. Berlin: Erich Schmidt Verlag.
Swan, Michael & Bernard Smith. 2001, 2nd ed. *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.
Walker, Robin. 2010. Teaching the Pronunciation of English as a Lingua Franca. Oxford: Oxford University Press.

Studying collocations in learner language: Which statistic to use?

Dana Gablasova, Vaclav Brezina
Lancaster University
d.gablasova@lancaster.ac.uk, v.brezina@lancaster.ac.uk

Introduction & motivation

Formulaic language has occupied a prominent role in the study of language learning and use for several decades (Wray, 2013). Recently an even more notable increase in interest in the topic has led to an 'explosion of activity' in the field (Wray, 2012, p.23). Language learning research (LLR) in both first and second language acquisition has focused on examining the links between formulaic units and fundamental cognitive processes in language learning and use, such as representation of and access to these units in mental lexicon (Wray 2002, 2012, 2013; Ellis et al, 2015). Collocation, a specific unit of formulaic language, holds a prominent position in LLR, having been used in a number of studies on formulaicity in L2 (Schmitt, 2012). The statistical measures for identifying collocations, association measures (AMs), in these studies are of paramount importance as they directly and significantly affect the findings of these studies and consequently the insights into language learning that they provide.

However, so far, the statistical AMs in LLR have been largely used as apparently effective, but not fully understood mathematical procedures (Gablasova, Brezina & McEnery, 2017). The rationale behind the selection of MI-score in studies on formulaic development is not always fully transparent and systematic (González Fernández & Schmitt, 2015) and is often motivated by tradition rather than by specific aims of a given LLR study.As González Fernández & Schmitt (2015, p. 96) noted, "it is not clear which of these [MI-score and t-score] (or other) measures is the best to use in research, and to date, the selection of one or another seems to be somewhat arbitrary". The aim of this paper is therefore to propose a principled approach to the selection of AMs in language learning research. In particular, we will discuss three specific AMs, t-score, MI-score, and Log Dice, and consider their ability to highlight different aspects of formulaicity. T-score and MI-score were chosen because of their prominent role in recent corpus-based learner language studies; Log Dice is introduced as an alternative to MI-score.

Method

In order to address the above issues and research aims, the paper seeks to achieve the following two objectives: i) to discuss the differences between the three AMs. Special emphasis will be given to what type of linguistic pattern each of them highlights and how this may affect the conclusions draws about learner language, ii) propose general principles for selection of association measures in LLR. The study examines these questions using data from several native speaker and learner corpora. In particular, the British National Corpus and its subcomponents (e.g. the 5-million-word spoken informal component, the BNC-Demographic) was used to analyse the collocations identified with different AMs in language of native users of the British English; Trinity Lancaster Corpus, a 3.5 million corpus

of spoken learner English was used to examine the collocational patterns in learner production. The three AMs were examined using three types of collocations representing a range of constructions that commonly appear in language learning collocational research: verb + complementation (*make* + *sure/decision/point*), adjective + noun (*human* + *beings/rights/nature*) and adverb + adjective (*vitally/very/really* + *important*).

Results & Discussion

The results revealed the following pattern with respect to the first research objective: A difference between the three AMs (MI-score, Log Dice and t-score) in identifying the strength of the relationship between words. While the difference between measures such as t-score and MI-score was expected, the difference between MI-score and Log Dice deserves further attention as both measures reward similar linguistic properties of collocations (e.g. exclusivity of association). The difference has implications for the selection and interpretation of AMs in language learning research. Following these findings, with respect to the second research objective, we propose general principles for the selection of AMs for language learning research. These include the need to understand 1) the mathematical reasoning behind the measure, 2) the scale on which it operates and 3) its practical effect (what combinations of words get highlighted and what gets hidden/downgraded).

References:

Ellis, N.C., Simpson-Vlach, R., Römer, U., Brook 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.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Language Learning.*

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1-26.

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, *166*(1), 94-126.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Schmitt, N. (2012). Formulaic Language and Collocation. In Chapelle, C. (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: CUP.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231-254.

Wray, A. (2013). Formulaic language. *Language Teaching*, *46*(3), 316-334.

POS tagging a spoken learner corpus: Testing accuracy testing

Gaëtanelle Gilquin
Université catholique de Louvain
gaetanelle.gilquin@uclouvain.be

Part-of-speech (POS) tagging has extended the range of studies that can be carried out on the basis of a corpus. For written corpora of English, which represent the first type of corpora that were collected and, to this day, probably one of the most frequently used resource in corpus linguistics, it is quite common to have access to a POS tagged version of the data. If not, it has become relatively easy to POS tag a corpus, using a tool like the free CLAWS WWW tagger (see http://ucrel.lancs.ac.uk/claws/ trial.html). Not only are POS tagged corpora readily accessible, but they tend to be reliable, since state-of-the-art POS taggers are said to achieve an accuracy rate of about 97% (Manning 2011; Jurafsky & Martin forthcoming).

Turning to learner corpora, however, it appears that POS tagging is not so widespread. One major reason for this is that POS taggers have usually been designed to deal with standard language (see Gilquin & De Cock 2011: 149), which in effect means native written language (often English). Yet, attempts to POS tag learner corpora of written English have proved to be quite successful, with reported accuracy rates of about 95% (de Haan 2000) or even higher (van Rooy & Schäfer 2002; Granger et al. 2009: 16). The POS tagging of spoken learner corpora, on the other hand, has so far been rather neglected. With a view to testing the feasibility of automatically POS tagging one such corpus, the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010), made up of subcorpora representing the spoken English production of different L1 populations, an experiment was conducted which consisted in having POS tagged samples of the corpus checked for tag accuracy. The main objective of this experiment was to determine the success rate of a POS tagger on the LINDSEI data. At the same time, however, we wanted to test the procedure for checking tag accuracy, a methodological issue that is generally ignored in the literature, where accuracy rates are reported with no or little discussion of how these rates were obtained.

Eleven LINDSEI samples were selected for analysis, representing a total of about 22,000 words. In addition, a sample from the native equivalent of LINDSEI, the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004), was included as a point of reference. After some pre-treatment of the data, aimed at taking some of the specificities of the LINDSEI/LOCNEC transcription conventions into account, the CLAWS4 software (with the C7 tagset) was used to POS tag the twelve samples. Eleven LINDSEI teams were in charge of checking the tag accuracy of two files each: one from their national subcorpus and one from the French L1 subcorpus, which was thus checked by all the participants in the experiment. The CLAWS tagset and manual were sent to the participants, as well as some instructions on how to go about the checking task.

The average results indicate an encouraging accuracy rate of 92% for the LINDSEI data, as against 94% for the LOCNEC sample. However, it appears that this average rate hides a certain amount of variation. For one thing, some tags are more accurately assigned than

others: almost half of the tags have a 100% accuracy rate (this is the case, for example, of "VM", the tag for modal auxiliaries), whereas the others have an accuracy rate ranging between 99% (in the case of "VD0", the tag for the base form of *do*) and 0% (in the case of "MCMC", the tag for hyphenated number, which appears in the Brazilian subcorpus in relation to conventions used to indicate long pauses). For another thing, the different samples vary in their accuracy rate, which could be the result of differently tagged samples, some being perhaps more accurately tagged than others, or which could have to do with the variety of raters, who may have applied different criteria to check the accuracy of the tags. That the latter explanation at least partly accounts for the observed discrepancies appears from the comparison of the French L1 sample as checked by the different raters. With an overall Fleiss' kappa value of 0.514, the checking task shows a moderate inter-rater agreement, which suggests that it is not necessarily obvious to decide what a correct or incorrect tag is, especially when the data represent spoken learner language. Using examples from the checked POS tagged data, it will be shown what can lead to erroneous tags or trigger discrepancies among the raters, and how some of these problems can be solved, by improving the POS tagging or POS tag checking processes.

References:

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

de Haan, P. (2000). Tagging non-native English with the TOSCA–ICLE tagger. In C. Mair & M. Hundt (Eds.) *Corpus linguistics and Linguistic Theory*. Amsterdam: Rodopi, 69-79.

Gilquin, G. & De Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, 16(2), 141-172.

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.

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

Jurafsky, D. & Martin, J. H. (Forthcoming). *Speech and Language Processing*. 3rd edition. Englewood Cliffs: Prentice Hall.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In A. F. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer, 171-189.

van Rooy, B. & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, 20, 325-335.

Exploring word-formation in Learner Corpus Research: A case study on English negative affixes

Gaëtanelle Gilquin[1], **Marie-Aude Lefer**[1, 2]
[1]Université catholique de Louvain, [2]Université Saint-Louis – Bruxelles
gaetanelle.gilquin@uclouvain.be, marie-aude.lefer@uclouvain.be

It is well-known that knowledge of word-formation patterns, such as derivation, facilitates vocabulary expansion and can help work out the meaning of unfamiliar words. Back in 1994, Nation pointed out that "by focusing on frequent and regular prefixes and suffixes, vocabulary learning can be made much more manageable (…) because such a focus reduces the number of items to be learned and provides an approach to learning which relates to previous knowledge" (Nation 1994: 2584). To date, Second Language Acquisition research has mainly focused on inflectional morphology, to the detriment of derivational morphology. In the field of Learner Corpus Research (LCR), it is safe to say that word-formation in general, and derivational morphology in particular, is practically unchartered territory. Callies's (2015, 2016) corpus-based studies on lexical innovations in learner English are notable exceptions to this neglect. Callies uncovered two types of lexical innovations in interlanguage: (i) *intralingual*, L2-based innovations, which correspond to word coinage or overregularization (e.g. *unmerciful* instead of *merciless*) and (ii) *interlingual*, L1-based innovations, which correspond to cases of cross-linguistic influence (e.g. *refugiated*, from Spanish *refugiarse*) (see also Balteiro 2011 for similar observations). Our presentation aims to contribute to filling this gap in LCR by investigating the use of negative morphology in learner English, i.e. words with the derivational affixes *de-*, *dis-*, *in-*, *non-*, *un-* and *-less* (see e.g. Horn 1989/2001, Hamawand 2009, Bauer *et al*. 2013 on negative affixation in native English). This semantic category of affixes has been selected for two main reasons: first, it is the most frequent one (Bauer *et al*. 2013); and second, it lends itself to a more paradigmatic approach, as negation can be expressed morphologically (*unhappy*), lexically (*sad*) and syntactically (*not happy*).

We investigate four learner populations of intermediate to advanced learners of English, two of them with a Romance L1 – French- and Italian-speaking learners – and two with a Germanic L1 – Dutch- and German-speaking – and compare them with native speakers of English. Our dataset, extracted from the *International Corpus of Learner English* (ICLE; Granger *et al*. 2009) and the *Louvain Corpus of Native English Essays* (LOCNESS), consists of 5,500+ words with a negative affix, carefully validated following strict selection criteria (taking into account, among others, synchronic analyzability and semantic transparency). The data have been manually coded for transfer, semantic categories (distinguishing between contrary/contradictory negatives on the one hand and privatives/reversatives on the other; cf. Bauer *et al*. 2013) and creativity (teasing out creative forms from lexicalized derivatives, relying on four online dictionaries as reference tools), among others.

Overall, our results indicate that compared to native speakers, the German- and French-speaking learners underuse negative affixes, while the Italian-speaking learners overuse them. Looking at individual affixes, we see that there is a possible influence of the language family in that Romance prefixes tend to be preferred by the Italian- and French-speaking

learners, while Germanic affixes tend to be favoured by the German- and Dutch-speaking learners. Interestingly, for adjectives, we find that syntactic negation by means of *not* seems to be used as a compensation strategy for the underuse of negative affixes: there is a general tendency among learners to underuse morphologically derived adjectives and to overuse syntactically negated adjectives (e.g. *not married* instead of *unmarried*, *not healthy* instead of *unhealthy*). As regards creative forms, which account for 2 to 7% of the dataset, results show that in native writing they are predominantly built by means of *non-* and *un-*, which are said to be among the most productive affixes in English (Bauer *et al*. 2013), while learners exhibit different preferences according to their L1. Some of these tendencies seem to be transfer-related, such as the widespread use of *de-* in ICLE-French and of *in-* in ICLE-Italian, and the presence of calques from the learners' L1 (e.g. *incertain* and *deshumanized* in ICLE-French).

Our study provides empirical and quantitative evidence that derivational morphology can still present some difficulties for intermediate to advanced learners of English, with transfer playing a crucial role in the use of negative affixes, and, more generally, that there is a clear "need for more direct attention to the teaching of derivative forms" (Schmitt & Zimmerman 2002: 145).

References:

Balteiro, I. (2011). Awareness of L1 and L2 word-formation. Mechanisms for the development of a more autonomous L2 learner. *Porta Linguarum*, 15, 25-34.

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: OUP.

Callies, M. (2015). Effects of cross-linguistic influence in word formation. A comparative learner corpus study of advanced interlanguage production. In H. Peukert (Ed.) *Transfer Effects in Multilingual Language Development*. Amsterdam: Benjamins, 127-143.

Callies, M. (2016). Towards a process-oriented approach to comparing EFL and ESL varieties. A corpus-study of lexical innovations. *International Journal of Learner Corpus Research*, 2(2), 229-250.

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

Hamawand, Z. (2009). *The Semantics of English Negative Prefixes*. London: Equinox.

Horn, L. (1989/2001). *A Natural History of Negation*. Chicago: University of Chicago Press. Reprinted in 2001 by CSLI Publications, Stanford, CA.

Nation, I. S. P. (1994). Morphology in language learning and teaching. In R. E. Asher (Ed.) *The Encyclopedia of Language and Linguistics*. Vol. 5. Oxford: Pergamon Press, 2582-2585.

Schmitt, N. & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145-171.

Bridging the gap between learner corpus research and translation studies: The Multilingual Student Translation corpus

Sylviane Granger, Marie-Aude Lefer
Université catholique de Louvain
sylviane.granger@uclouvain.be, marie-aude.lefer@uclouvain.be

Learner corpus research (LCR) and corpus-based translation studies (CBTS) are two research strands that arose at approximately the same time, in the late 80s/early 90s (Granger 1993, 1994; Baker 1993, 1995). Although the two fields differ in their research agendas and pedagogical applications, their respective objects of study – learner language and translated language – share one major characteristic, i.e. they involve a process of interlingual mediation. As a result, LCR and CBTS are partly faced with similar issues, such as assessing the impact of transfer – from the first language for LCR and from the source text/source language for CBTS – and distinguishing between transfer-related effects and general features of foreign language acquisition or translation (e.g. increased explicitness, lexical and syntactic simplification). Granger (1996) advocated a rapprochement between the two fields in the form of the Integrated Contrastive Model, which involves to-ing and fro-ing between learner and multilingual corpora, with transfer as the key connecting point. Although this idea has received support from a number of scholars in both LCR and CBTS (Gilquin et al. 2008; Chesterman 2007; Johansson 2007) and has been implemented in a few studies (Gilquin 2000; Vanderbauwhede 2012), research along those lines has been fairly limited to date.

In our presentation we will contrast this approach with a different way of interfacing the two fields, viz. compiling learner translation corpora (LTC), which can be seen as two-in-one resources, as they contain translations produced by foreign language learners or trainee translators. The first LTC emerged in the early 2000s (Uzar 2002; Bowker & Bennison 2003) and were followed by several similar initiatives, such as the MeLLANGE corpus (Kübler 2008). However, as pointed out by Espunya (2014), "the field is clearly in its infancy, judging by the scarcity of publications reporting results or even research programmes". Starting from an overview of the existing LTC, we will show that most suffer from limitations in terms of metadata, language pairs covered, translation direction (mainly into the translator's mother tongue) and error annotation systems.

This overview will be followed by a presentation of the main characteristics of a new international corpus initiative, the Multilingual Student Translation (MUST) project, which aims to address some of the weaknesses of earlier collections. The following key features of MUST will be described: (i) the corpus is truly multilingual (at this stage, the project partners cover approximately 25 languages and 50 language pairs), (ii) it includes both direct (L2>L1) and inverse (L1>L2) translation, (iii) source texts can be general or specialized and represent a range of text types, genres, registers and topics, (iv) the corpus also contains expert translations, which can act as reference or model translations for the student translations, and (v) rich metadata are collected together with the source and target texts.

Our main concern in designing the corpus was to cater for the needs of the two research communities, i.e. learner corpus researchers and translation scholars. This concern is clearly evident in two key components developed for MUST, i.e. the metadata and the annotation system, both of which are standardized and will be used for all the translations included in the database to ensure full comparability of the data and reliable interpretation of the results. The MUST metadata are subdivided into three categories: translator metadata (language and study background, L2 proficiency level, translation experience, among others), source text metadata (e.g. general/specialized text, genre, topic, use of a translation brief) and translation task metadata (including detailed information on the resources used, the use of computer-aided translation tools and the revision phase). The annotation system draws on typologies used in both LCR and CBTS but presents two distinctive characteristics: first, it offers the possibility of highlighting both erroneous and correct use; second, it offers the option of marking translation strategies (such as transposition, simplification or explicitation), thereby allowing for theoretically oriented research. In view of these two features, it was decided to refer to the annotation system as "translation-oriented annotation" rather than "error annotation".
The last part of the presentation will focus on introducing the web-based interface of the corpus, Hypal4must, a tailor-made version of Obrusnik's (2014) Hypal tool, which includes POS tagging, automatic sentence alignment, annotation and corpus search, and contains both a research- and teaching-oriented environment.

References:

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

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. *Target* 7(2), 223-243.

Bowker, L. & Bennison, P. (2003). Student Translation Archive: Design, development and application. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education.* London & New York: Routledge, 103-117.

Chesterman, A. (2007). Similarity analysis and the translation profile. *Belgian Journal of Linguistics* 21, 53-66.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* 48(1), 33-43.

Gilquin, G. (2000). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3(1), 95-123.

Gilquin, G., Papp, S., & Díez-Bedmar, M.B. (eds.) (2008). *Linking up contrastive and learner corpus research.* Amsterdam & Atlanta: Rodopi.

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

Granger, S. (1994). The Learner Corpus: A Revolution in Applied Linguistics. *English Today* 39(10/3), 25-29.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson

(eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press, 37-51.
Johansson, S. (2007). Seeing through Multilingual Corpora. On the use of corpora in contrastive studies. Amsterdam & Philadelphia: Benjamins.
Kübler, N. (2008). A comparable Learner Translator Corpus: Creation and use. LREC 2008 Workshop on Comparable Corpora, 73-78.
Obrusnik, A. (2014). Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. Eleventh International Conference Teaching and Language Corpora, Lancaster, 20-23 July 2014, 67-69.
Uzar, R.S. (2002). A corpus methodology for analysing translation. In S.E.O. Tagnin (ed.) *Cadernos de Tradução: Corpora e Tradução.* Florianópolis: NUT, 1(9), 235-263.
Vanderbauwhede, G. (2012). The Integrated Contrastive Model evaluated: The French and Dutch demonstrative determiner in L1 and L2. *International Journal of Applied Linguistics* 22(3), 392-413.

Intensifying constructions in French-speaking L2 learners of Dutch and English: longitudinal results

Isa Hendrikx[1], **Kristel Van Goethem**[1, 2]
[1]Université catholique de Louvain, [2]F.R.S.-FNRS
isa.hendrikx@uclouvain.be, kristel.vangoethem@uclouvain.be

Intensification can be expressed cross-linguistically by several morphological and syntactic constructions (among others, Kirschbaum 2002; Hoeksema 2011, 2012; Zeschel 2012; Rainer 2015). The diversity of constructions available to express a single function implies a form-function asymmetry; alongside marked language-specific preferences for particular types of intensification complicate the acquisition of intensifying constructions for second language learners. In this contribution we will explore the longitudinal impact of Content and Language Integrated Learning (CLIL) on the acquisition of intensifying constructions in an L2[1].

Our research is situated within the theoretical framework of usage-based Construction Grammar (cf. Tomasello 2003; Goldberg 2010 among others). Second language acquisition is presumed to be complex because of the competition between L1 and L2 constructions (Ellis & Cadierno 2009). This study focuses on one specific case of such constructional competition, namely the expression of adjectival intensification in the interlanguage of French-speaking learners of Dutch or English.

More specifically, we will address three research questions:

(i) To what extent can we observe variation in the use of intensifying constructions between the native and learner language?
(ii) Does more input provided through a Content and Language Integrated Learning (CLIL) approach lead to a more native-like acquisition of intensifying constructions?
(iii) What developments can we observe in the learners' use of intensifying constructions from a longitudinal point of view (over the course of two academic years)? Do the learners in CLIL make more significant progress than those in traditional L2 educational settings?

The data for this study come from a corpus of the written productions in the form of fictional e-mails on the subject of a party or holidays. In 2015 we collected the first texts written by the participants, who were 5th year French-speaking secondary school pupils (aged 16-17), in CLIL and non-CLIL settings learning Dutch (CLIL n=132; non-CLIL n=100) or English (CLIL n=90; non-CLIL n=90) as a foreign language, and control groups of 63 native speakers of Dutch and 68 native speakers of English of about the same age. (The data of the English control group was collected in 2016). In April and May 2017, the French-speaking Belgian learners (who are in 6th grade now) will once more write e-mails in their

[1] This study is part of a broader interdisciplinary project on CLIL in French-speaking Belgium (Hiligsmann et. al. forth.).

target language (again on similar topics). In the present study the newly gathered data will be compared to the learner data collected in 2015 and the native data collected in 2015-2016, in order to examine developments in the pupils' use of intensification in their L2.
All instances of intensifying constructions observed in this corpus are subjected to a collostructional analysis, which expresses the degree of attraction/repulsion of a lexeme to an intensifying construction in the form of *p*bin-values[1] (Stefanowitsch and Gries 2003; Gries 2007; Ellis and Ferreira Junior 2009; Hoffmann 2011). We already conducted a covarying collexeme analysis (Gries 2007) on the data gathered during the first data collection, and showed its benefits: idiosyncratic uses of intensifying constructions are easily identified in the L1 corpora, and misuse (spelling mistakes, grammatical mistakes and semantic misuse) is efficiently detected in the learner corpora (Hendrikx et.al. 2017). Analysis of the data collected in 2015 and 2016 shows, for instance, that intensifying compounds are significant collostructions in the L1 corpora, e.g. *bloedheet* lit. 'blood-hot' (*p*bin=2,668 in native Dutch) and *crystal clear* (*p*bin= 2.792 L1 English) while learners use those particular constructions rarely or not at all. The collostructional analysis also unveiled the following erroneous [Intensifier + Adjective] collocations in the learner corpora: **veel leuk* 'many nice' (*p*bin 1,533 for non-CLIL learners), **so luxuous* (*p*bin 1.315 CLIL learners) and **amazingly delicious* (*p*bin 1.663 CLIL learners).
In the present study, the collostructional analysis will be utilized to investigate longitudinal developments in the learners' acquisition of intensifying constructions. In addition, the lexical diversity and productivity of the learners' use of intensifiers will be compared across groups and longitudinally, to gain insights into the impact of CLIL and traditional foreign language classes on the acquisition of intensification in a second language.

References:

Hendrikx, I., Van Goethem, K. & F. Meunier (2017). *The expression of intensification in the interlanguage of French-speaking CLIL and non-CLIL learners of English*. Oral presentation at Cogling7, January 5th and 6th, 2017. Radboud University, Nijmegen The Netherlands.

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

Ellis, N. & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics 7.* 187-220.

Goldberg, A. (2010) [2006]. *Constructions at Work. The nature of Generalization in Language.* Oxford: Oxford University Press.

Gries, Stefan Th. (2007). Coll.analysis 3.2a. A program for R for Windows 2.x.

Hiligsmann, P., Van Mensel, L., Galand, B., Mettewie, L., Meunier, F., Szmalec, A., Van Goethem, K., Bulon, A., De Smet, A., Hendrikx, I., Simonis, M. (forthcoming) Content and Language Integrated Learning: linguistic, cognitive and educational perspectives. *Cahiers du Girsef.*

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

[1] Bins are consecutive, non-overlapping intervals of a variable. In this case if pbin (bin of p) >3 than p<0.001; if pbin >2 than p<0.01; and if pbin >1.3 than p<0.05. The p value is adjusted according to the Bonferonni correction.

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

Hoffmann, T. (2011). *Preposition Placement in English: A Usage-Based Approach*. Cambridge: Cambridge University Press.

Kirschbaum, I. (2002). *Schrecklig Nett Und Voll Verrückt Muster Der Adjektiv-Intensivierung Im Deutschen*. Thesis. Düsseldorf, Universität, Diss 2002. Düsseldorf.

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

Riegel, M., Pellat, J.-C., Rioul, R. (1994). *Grammaire méthodique du français*. Paris: Presses Universitaires de France.

Stefanowitsch, A. & Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8*:2, 209-243.

Tomasello, M. (2003). *Constructing a language: A Usage-Based Theory of Language Acquisition*. Boston: Harvard University Press.

Zeschel, A. (2012). *Incipient Productivity. A Construction-Based Approach to Linguistic Creativity*. Berlin, Boston: De Gruyter Mouton. Retrieved 25 Jan. 2017, from http://www.degruyter.com/view/product/180582

Frequency and distribution of self-corrections in a spoken longitudinal learner corpus

Amanda Huensch, Nicole Tracy-Ventura, Taylor Chlopowski, Samantha Creel, Jessica Giovanni
University of South Florida
huensch@usf.edu, nkt@usf.edu, taylor31@mail.usf.edu, samantha93@mail.usf.edu, giovannij@mail.usf.edu

To date most of the research with learner corpora has focused on written language, and as such, studies examining unique features of spoken language have been scarce. For example, self-corrections (e.g., *and he had to <run across> [//] jump across the wall*) are a distinctive feature of spoken language that have been the subject of research with native speakers (Levelt, 1983, 1989) and bilinguals (de Bot, 1992) due to their relevance for understanding the process of speech production. In learner language they can be especially informative in regard to what learners are attending to while speaking, the fluency and accuracy of their speech, as well as the relationship to proficiency. By annotating spoken corpora for self-corrections it becomes possible to analyze the distribution and frequency of their use with corpus-based tools (Gilquin & DeCock, 2011). In the field of second language acquisition, a number of studies have examined second language users' self-corrections primarily with elicited data (e.g., Camps, 2003; Gilabert, 2007; Kormos, 2000) with very few studies utilizing more authentic data such as interviews (Belz et al., 2015; van Hest, 1996). Furthermore, while previous research has shown through cross-sectional research that more advanced students do not necessarily make fewer self-corrections (Kormos, 2000), longitudinal studies which investigate the same learners over time are rare. Therefore, based on the above mentioned gaps in the literature, the current study investigates the following research questions in a spoken longitudinal learner corpus:

1. How often do self-corrections occur in oral interviews before, during, and after a 9-month stay abroad?
2. To what extent is there a relationship between the frequency of self-corrections and proficiency over time?
3. What is the distribution of different types of self-corrections based on structural aspects of language (e.g., aspects of grammar, pronunciation and lexis) as well as the message content, and to what extent do these change over time?

The learner corpus used in this study (approximately 310,000 words) consists of interviews conducted in Spanish with 24 L1 English speakers who were pursuing a bachelor's degree in Spanish in the UK. As part of their degree program they were required to spend their third year (of a four-year degree) abroad. Data were collected six times over approximately two years: once before their year abroad, three times during, and twice after returning home to complete their degree. A proficiency test was also administered three times over the two-year period. In the interviews learners were asked a variety of questions related to their year abroad experience. For example, before going abroad they were asked to describe any

ideas they had for practicing the language and meeting people, and what their goals were for the year. Questions in-sojourn centered on things such as notable experiences, the people they interacted with, and their plans for the following months. The last two rounds of interviews were conducted back at the home university, and students were asked questions related to their ongoing contact with people met abroad and reflections about the experience. Interviews were transcribed in CHAT (Codes for the Human Analysis of Transcripts) and analyzed using the CLAN (Computerized Language Analysis) program, both available as part of the CHILDES project (MacWhinney, 2000). A range of features distinctive of spoken language were annotated in the transcripts including self-corrections, which were annotated with the retracing symbol [//] (described in the CHAT manual). The analysis is ongoing but thus far results of research question one demonstrate that learners made fewer self-corrections over time, with the lowest numbers at the end of their year abroad (2.36 corrections per 100 words) and 4 months after returning home (2.33 corrections per 100 words); however no significant correlations were found between the frequency of corrections and proficiency throughout the two years (research question two). Analysis related to research question three demonstrates that learners show evidence of a variety of self-correction types including rephrasing of content as well as corrections to lexical choice, pronunciation, and grammatical features such as morphology, articles, pronouns, and gender agreement. The presentation will describe these results in more detail and conclude with a discussion of the implications of this work for understanding models of speech production and the analysis of spoken learner corpora.

References:

Belz, M., Sauer, S., Lüdeling, A., & Mooshammer, C. (2015). Repair behaviour of advanced German learners in the Berlin Map Task Corpus. In: *Book of Extended Abstracts of the Workshop on Phonetic Learner Corpora, Satellite Workshop of the 18th International Congress of Phonetic Sciences*, Glasgow, UK.

Camps, J. (2003). The Analysis of Oral Self-Correction as a Window into the Development of Past Time Reference in Spanish. *Foreign Language Annals*, *36*, 233-242.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics, 13,* 1-24.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*, 215-240.

Gilquin, G., & De Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, *16*, 141-172.

Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning*, *50*, 343-384.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41-104.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Van Hest, E. (1996). *Self-repair in LI and L2 production*. Tilburg: Tilburg University Press.

Writing development of Swedish native writers: a comparison of product and process in a corpus of expository writing

Victoria Johansson
Lund University
victoria.johansson@ling.lu.se

Research on L1 writing development, from school age through adulthood, is not always so easy to find (cf. Scott 1988; Myhill 2008), although there are exceptions, like Berman & Verhoeven (2002), which describes the literacy development in writing and speech in seven languages. In this paper we are aiming at describing writing development from a linguistic perspective, through a corpus of expository texts, who have all been collected in a similar setting, independent of the participants' age. We are further interested in comparing the linguistic product, i.e., the finished text, and the linguistic process, i.e. the way in which the text was produced. The question we are addressing is whether the same developmental steps are identified if we study the text products as when we study the writing processes. Data consists of a small learner corpus of 115 written texts collected in a lab setting, which provides us with material that is comparable through the age groups: 10-year-olds (n=20), 13-year-olds (n=20), 15-year-olds (n=20), 17-year-olds (n=20), university students (n=19), and university students in a creative writing (CW) program (n=16). Inclusion criteria were Swedish as L1, no reading and writing difficulties, and basic typing skills. The age groups were selected to be able to describe the development through the school ages, up until university age. The CW-students represent another developmental step, where one can expect that writing expertise has increased. Data was collected using keystroke logging (ScriptLog), which enables the investigation of the writing process regarding for instance revisions. The participants were asked to discuss problems shown in a short, wordless film, prior to the elicitation, in a 30 minutes-paper.

The text product was explored through the following measures (using ANOVAs). 1. Text length (number of words) significantly increased with the 15-year-olds writing the longest texts. After this age, the texts became shorter for every age group, with significant differences between university students and CW-students. 2. Syntactic complexity (number of clauses per t-unit) was slowly increasing from age 13 to university students. The CW-students had significantly less complex sentences than the university group. 3. Lexical diversity showed no significant differences between the ages 10 and 13. But between the age of 13 and 15, and 15 and 17 the lexical diversity increased in the texts. There was however no differences between 17-year-olds and adult university students. The CW-students had the highest lexical diversity of all the participants.

The text process was explored through the following measures: 1. Text length in final text (number of characters) was significantly increasing through the school years, with the 17-year-olds writing the longest texts. After this, the length decreased, and the CW-students produced the shortest texts. 2. Text length in linear text (number of characters, including characters that were deleted) showed that the 17-year-olds by far produced the most characters. 3. Percentage deleted characters was significantly lower for the 15-year-olds than any other group (≈13%), while the groups of 17-year-olds and 13-year-olds both

deleted around 30%. The CW-students deleted a higher percentage than the university students.
Writing development look somewhat different depending on whether we study the product or the process. The longest texts measured in words are produced by 15-year-olds, but the longest texts, measured in number of characters, are produced by 17-year-olds. The syntactic complexity increases through the ages, but the most advanced group have less syntactically complex texts. This group has on the other hand the most lexically diversed texts. In the presentation we will relate the developmental patterns to cognitive theories of writing development, describing writers as knowledge tellers, knowledge transformers and knowledge crafters (Kellogg 2006).

References:

Berman, R. and Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. Written Language and Literacy, 5:1–44.

Kellogg, R. T. (2006). Professional writing expertise. In Ericsson, K. A., Charness, N., Feltovich, P. J., and Hoffman, R. R., editors, The Cambridge Handbook of Expertise and Expert Performance, pages 389–402. Cambridge University Press, New York.

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. Language and Education, 22(5):271–288.

Scott, C. M. (1988). Spoken and written syntax. In Nippold, M. A., editor, Later Language Development. Ages nine through nineteen, pages 49–95. Pro Ed, Austin, Texas.

High-frequency verbs in EFL learners' conversation: patterning of *do*, *have*, *make*, *give* and *take*

Rita Juknevičienė
Vilnius University
rita.jukneviciene@flf.vu.lt

Spontaneous speech leaves little time for planning of one's moves and demonstration of sophisticated lexical expressions. In grammars, spoken language is described as simplistic, consisting of high-frequency words and simple expressions (Biber et al. 1999: 1044-45). While these features are characteristic of native speaker English, learners of English as a foreign language (EFL) face a number of challenges while developing their competence in speech. One of the factors contributing to the naturalness of expression is related to the learners' ability to make use of high-frequency vocabulary, including such verbs as *do*, *make*, *have* etc. Corpus analyses of spoken English show that these verbs are often used as light or delexical verbs when they form 'verb + noun' collocations whose meaning is largely derived from the noun while the verb is delexicalized, e. g. *give a smile*, *have a go*, *make a discovery* (Huddleston & Pullum 2002; Sinclair 1991: 147–151; Biber et al. 1999: 428). Research into the use of high-frequency verbs in L2 English suggests that learners in comparison with native speakers underuse high-frequency verbs in light constructions which applies both to written English (Altenberg & Granger 1991; Wang 2016) and speech (Shirato & Stapleton 2007). But what is the global picture of high-frequency verbs in learner speech? If the light construction is underrepresented, what are the other patterns of high-frequency verbs in learners' conversation? Hypothetically, as simple words they should feature prominently in learner speech, but do they?

In order to investigate the patterning of high-frequency verbs in EFL learner speech, this study deals with five English verbs, namely, *have*, *do*, *make*, *take*, and *give*, and their collostructional analysis (Stefanowitsch & Gries 2003). The following research questions are at the focus:

(1) Which uses of the verbs—lexical, delexical, auxiliary or modal (where relevant)—are realised in EFL learner speech? What is the collostructional strength of the light verb construction?
(2) Does the patterning of the verbs vary across three groups of EFL learners whose first languages are different?

Conceived as a contrastive interlanguage analysis (Granger 2015), this research is concerned with spoken English produced by three groups of EFL learners—Lithuanian, Polish and Swedish—represented in the currently developed new edition of the LINDSEI corpus (cf. Gilquin et al. 2010 for the 1st edition). The total size of learner ('B') turns in three LINDSEI subcorpora is 246,261 words. The LOCNEC corpus of native speaker interviews compiled at the University of Louvain-la-Neuve (CECL) and having a similar design to the LINDSEI corpus was chosen as a reference corpus (118,517 words in 'B' turns). The analysis involves a combination of qualitative and quantitative research methods. First, the WordSmith Tools software (Scott 2008) is used to extract all instances of the five verbs from each subcorpus, categorise them into lexical, delexical/light, auxiliary (for *do* and

have), modal (*have to*) and obtain contingency data for statistical analysis. Next, the collostructional strength of the verbs was measured to establish which pattern of each verb is most likely to occur in learner speech and evaluate the attraction of the light 'verb + noun' pattern. For this purpose, the Fisher Exact Test (Levshina 2015) was used. Next, the test of statistical significance (chi-square) was run to check for statistically significant differences in the patterning of the five verbs among the three learner groups. All statistical tests are computed with the program R (R Core Team 2015).
Preliminary results suggest that the light uses of the verbs only partly account for the differences in the patterning of the verbs among the analysed subcorpora. For example, in the case of *have*, it is its auxiliary and possessive uses (especially in the Swedish subcorpus) which account for statistically significant differences in the data. The comparison of NS and NNS data, however, points to the light uses of the verbs as a major factor contributing to differences in the patterning of the verbs in speech.

References:

Altenberg, B. & Granger, S. (2001). The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. *Applied Linguistics,* 22(2), 173–195.

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

CECL (Centre for English Corpus Linguistics). LOCNEC corpus. [Online description] Available from: https://www.uclouvain.be/en-cecl-lindsei.html. Accessed on 4 January 2017.

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

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

Huddleston, R. & Pullum, G. K. (Eds) (2002). *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Levshina, N. (2015). *How to Do Linguistics with R. Data exploration and statistical analysis.* Amsterdam/Philadelphia. John Benjamins Publishing Company.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available from: http://www.R-project.org. Accessed 15 September 2016.

Shirato, J. & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research,* 11(4), 393–412.

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

Scott, M. (2008). *Wordsmith Tools*. Version 5. Oxford: OUP.

Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics,* 8(2), 209–243.

Wang, Y. (2016). The Idiom Principle and L1 influence. A Contrastive Learner-Corpus Study of Delexical Verb + Noun Collocations. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Do subject-internal factors predict third language acquisition? Preliminary evidence from a corpus of Cantonese learners of modern languages

Xin Kang, Kay Wong, Patrick C.M. Wong
The Chinese University of Hong Kong
xin.kang@cuhk.edu.hk, k.wong@cuhk.edu.hk, p.wong@cuhk.edu.hk

The goal of our present study is to examine subject-internal factors that may contribute to foreign language acquisition. We do so by capitalising on the large number of learners (reporting 500+ subjects here) of a relatively homogenous background (native speakers of Cantonese with high English language proficiency) who are learning different third languages (L3) at The Chinese University of Hong Kong.

A number of subject-internal factors have been found to predict second language learning, among them motivation (Dörnyei, 2001), cognitive abilities (Christopher, et al., 2012;Csapó & Nikolov, 2009), musical background (Wong & Perrachione 2007; Slevc & Miyake, 2006), and from more recent studies neurophysiology (Wong et al., 2007) and neuroanatomy (Wong et al., 2008). These studies are consistent with the view that "aptitude" can be measured before learning and can explain a portion of the variance in the success of learning a second language (Carroll, 1973). While some of these studies are of a larger scale (e.g., Ehrman, 1994), many are confined to a smaller group of subjects. As these predictor variables are likely to be correlated, the size of the effect for each variable might be limited when they are examined simultaneously. In the present study, we examined an array of subject-internal factors simultaneously, including attitude, motivation, performance IQ, musical background, socioeconomic status (SES) in order to determine their relative contribution to language learning success.

We present data from 568 participants (aged between 18-25 years old) who spoke Cantonese as their native language and had no report of hearing, neurological, or psychiatric disorders. They were learners of French (30%), German (30%), and Spanish (40%) at different proficiency levels. To assess their modern language achievement, three types of measures were included. First, spontaneous narrative production in the target language was elicited using Mercer Mayer's wordless picture book of Frog, Where are You? (1969)). Audio recordings were transcribed in the CHAT format (MacWhinney, 2000) and morphosyntactically tagged with the CLAN program. Second, exam scores were gathered from each individual and standardised as z-scores. Third, participants completed the Body Parts Picture Naming task (O'Grady et al., 2009) as a measure of language strength.

We first extracted a number of outcome measures from the narrative production, such as measures of expressive language development (e.g., mean length of utterance) and fluency (e.g., retracing). We found 17 variables out of 26 showing significant differences between High Level and Low Level learners across all three languages. We then adopted the Principal Component Analysis (PCA) with the Bartlett's approach to get the factors scores as a composite language measure. Thus, each individual learner had a global score of

language proficiency that took into account their narrative production, classroom performance, and language strength.

Hierarchical linear mixed-effects models were used to test which variables (if any) in the learners' profile have effects on their L3 achievement using the lme4 package (Bates, et al., 2016) in the R environment (R Core Team, 2016). The type of languages and classes were specified as random factors. As fixed effects, we entered language proficiency level, external motivation, internal motivation, anxiety, attitude, parents' SES, IQ, age, gender, and years of music training. Results revealed that a favourable attitude towards the target culture remained as the factor that was positively associated with language achievement when other factors were controlled in the model ($\beta = .18$, SE= .05, $p < .001$). Factors, such as internal motivation and age also had significant effects in subsets of language learners, but the effects were not generalised across the three languages. Variables that had no reliable predictive power the learners' achievement include IQ, gender, and parents' SES.

In conclusion, our results demonstrated the potential to use large-scale learner corpora to quantify language achievement by overcoming the limitation of sample size and the number of variables included in analysis. After examining many factors that were previously reported to be related to learners' language achievement, learners' attitude turned out to be the most robust factor associated with L3 learning even after considering collinearity, and when proficiency is measured more comprehensively. Thus, external factors may eventually play a more determining role in ultimate success in learning foreign languages instead of pre-training behavioural factors.

References:

Bates, D.M., Maechler, M., Bolker, B., Walker, S., Christensen, R.,…Green, P. (2016). *Linear miexed-effects models using 'Eigen' and S4.* R acakage version 1.1-12.

Carroll, J. (1973). Implications of aptitude test research and psycholinguistic theory for foreign language learning. *International Journal of PsychoLinguistics,* 2, 5–14.

Christopher, M. E., Miyake, A., Keenan, J. M., Pennington, B., DeFries, J. C., Wadsworth, S. J., … Olson, R. K. (2012). Predicting word reading and comprehension with executive function and speed measures across development: a latent variable analysis. *Journal of Experimental Psychology. General,* 141(3), 470–488.

Csapó, B. & Nikolov. M. (2009). The cognitive contribution to the development of proficiency in a foreign language. Learning and Individual Differences, 19, 209-218.

Dörnyei, Z. (2001). New themes and approaches in second language motivation research. *Annual Review of Applied Linguistics,* 21, 43-59.

Ehrman, M. E. (1994). The type differentiation indicator and adult foreign language learning success. *Journal of Psychological Type,* 30, 10-29.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, M. (1969). *Frog, Where Are You?* New York: Dial Press.

O'Grady, W., Schafer, A.J., Perla, J. Lee, O., & Wieting, J. (2009). A psycholinguistic tool for the assessment of language loss: the HALA project. *Language Documentation and Conservation,* 3(1), 100-112.

Slevc, L.R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science,* 17(8), 675-681.

Wong, P.C.M., Perrachione, T.K., Parrish, T.B. (2007). Neural Characteristics of Successful and Less Successful Speech and Word Learning in Adults. *Human Brain Mapping,* (28), 995–1006.

Wong P.C.M., Warrior, C.M., Penhune, V.B., Roy, A.K., Sadehh, A, et al. (2008) Volume of Left Heschl's Gyrus and Linguistic Pitch Learning. *Cerebral Cortex, (*18), 828–836.

On the way to a new multilingual learner corpus of foreign language learning in school: observations about task variations

Katharina Karges, Thomas Studer, Eva Wiedenkeller
University of Fribourg
katharina.karges@unifr.ch, thomas.studer@unifr.ch, eva.wiedenkeller@unifr.ch

Learner corpora are currently of significant interest in foreign language research (e.g. Callies & Paquot, 2015; Granger, Gilquin, & Meunier, 2013, 2015). A large part of these are concerned with adults and/or learners at intermediate or advanced levels of language proficiency and their corpus data is usually elicited through writing. Foreign language learning in primary and lower secondary school, however, has been somewhat neglected by corpus researchers so far, despite the fact that early language learning and acquisition of more than one foreign language have become common phenomena. One result of this is the fact that many school curricula and teaching materials are still predominantly based on grammar inventories and vocabulary lists which have little empirical foundation. Also, their alignment with the CEFR's action-oriented learning objectives (Council of Europe, 2001) seems to rely mainly on expert consensus instead of actual data on learner language development.

With this in mind, we are currently laying the foundations for a new corpus project. Its objective is to gather and analyse data on the linguistic competences of 12- and 15-year-olds in their language of schooling, their first foreign language (after 4 and 6 years of study) and their second foreign language (after 2 and 4 years of study). In the end, the corpus will contain texts in English, French and German. In creating two parallel foreign language subcorpora on the lower levels of language proficiency as well as a baseline corpus in the language of schooling, the project imitates the architecture of the FALKO corpus (e.g. Reznicek et al., 2012) and the design of the MERLIN project (Abel et al., 2014). These two projects also serve as models for the annotation and analyses of the learner texts (e.g. with respect to the use of (two) target hypotheses, the related error analyses and most notably the annotation systems).

The new project aims to describe the linguistic development of students' foreign language competences by using action- and content-oriented production tasks (cf. the concept of Dynamic Language Learning Progressions, e.g. Bailey & Heritage, 2014). In our contribution to LCR 2017, we would like to discuss preliminary investigations into the effect of task-related variables on student productions. A literature review during the early phases of the new project showed that the design and implementation of the tasks used for eliciting the learner language seem to receive little attention in learner corpus research. Most studies either rely on existing, unchanged tasks (e.g. from international standardised assessments) or on entirely new tasks which adhere to more or less specific design criteria. The effects of these criteria on the learners' productions, however, have rarely been examined further, although recent work suggests that variables such as text type or requirements in terms of

text content and length may influence the learner texts and have to be taken into account during corpus design (e.g. Révész, Michel, & Gilabert, 2016; Tracy-Ventura & Myles, 2015). Thus, in order to investigate what insight can be gained from relating language productions to the tasks from which they originated, we identified task criteria which are especially relevant in the school context. In a further step, we varied four task criteria during a preliminary data collection amongst 15-year-olds in the French-speaking part of Switzerland: mode (writing/speaking), medium (computer-based/paper-based, the spoken tasks being administered in one-one-interviews), text type or linguistic function (describing vs. arguing) and cognitive demands (every-day life and school-specific language, reflecting the BICS/CALP distinction). By comparing written and spoken texts on the same subject and across more or less demanding tasks, we want to explore to what extent written and spoken learner productions differ with respect to linguistic correctness, sophistication and formality (cf. e.g. Koch & Oesterreicher, 2007). In reaction to the continuing advance of electronic media in all areas of life, we will also investigate how the use of the computer influences formal and structural aspects of the students' written and oral texts. In our contribution, we will report on the results of these task variations and the conclusions we can draw from this for both our own main study and future corpus projects.

References:

Abel, A., Nicolas, L., Wisniewski, K., Boyd, A., & Hana, J. (2014). A trilingual learner corpus illustrating European reference levels. Ricognizioni. Rivista di lingue e letterature e culture moderne, 1(2), 111–126.

Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. TESOL Quarterly, 48(3), 480–506.

Callies, M., & Paquot, M. (2015). Learner corpus research: An interdisciplinary field on the move. International Journal of Learner Corpus Research, 1(1), 1–6.

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

Granger, S., Gilquin, G., & Meunier, F. (2013). Twenty years of learner corpus research. Looking back, moving ahead: Proceedings of the first Learner Corpus Research conference (LCR 2011). 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.

Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. Zeitschrift für germanistische Linguistik, 35(3), 346–375.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. Studies in Second Language Acquisition, 38(4), 703–737.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., Andreas, T. (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Humboldt-Universität zu Berlin.

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.

Investigating the effects of expertise and native language status in first and second language writing: p-frames across frequency profiles

Olesya Kisselev, Jungwan Yoon, Xiaofei Lu
Pennsylvania State University
ovk103@psu.edu, jxy201@psu.edu, xxl13@psu.edu

This paper addresses methodological concerns of defining and identifying formulaic sequences (FSs) with a special focus on Phrase Frames (p-frames, Fletcher 2007). Despite an abundance of research in FSs in various types of corpora there is a certain lack of agreement on the definition and methodological approaches to identification of FSs, even within the same theoretical orientation. For instance, the statistically-oriented approach, which identifies FSs through corpus-driven procedures with no criteria for idiomaticity pre-set by researchers, has utilized both frequency-based measures and strength of association measures (McEnery & Hardy 2014), which vary in regards to the underlying mathematical principles and, consequently, the results they produce. The issue is of great importance to applied research: for results to be comparable across studies and projects, the underlying methodology should be considered more carefully.

These methodological considerations are now being addressed with certain rigor: O'Donnell, Römer and Ellis (2013), for instance, investigated how four different operationalizations of FSs (n-gram frequency, n-gram association, p-frames and native norms) impacted the results of comparison of the use of formulaic language in expert vs. novice writer corpora, and L1 vs. L2 writer corpora. Having analyzed 3-, 4-, and 5-grams and p-frames, O'Donnell et al. found that different operationalizations led to different (and sometimes inconsistent) patterns of results.

Our paper presents a modified replication study of O'Donnell et al., in which we investigate further methodological issues. First, we investigated the issue of arbitrary frequency thresholds for FSs identification, which are not normally grounded in any empirical standards, but are rather informed by previous studies (Wood, 2015). O'Donnell et al., for instance, established the cut-off point of 3 following an established tradition. To explore the effect of an arbitrary frequency threshold, we examined the distribution of p-frames in different frequency bands, operationalized as four **quartiles** of frequency. The results show that the patterns of differences between the writer groups (L1 vs. L2 writers, expert vs. novice writers) vary substantially in different frequency quartiles. For example, the statistical difference in the usage of p-frame often appeared to be statistically significant in Quartile 4 (all p-frames above the threshold of 3) for either the factor of expertise or the L2 status, while in Quartiles 1 and 2 (less frequent p-frames) the effects either disappeared or reversed. These findings emphasize the importance of interpreting results of comparative studies on the usage of formulaic language more conservatively and with regard to particular frequency bands.

Secondly, we integrated the measure of type-token ratio (TTR) of p-frames to investigate the effect of incorporating diversity dimension on the result of FS usage in different writer

groups. The TTR measure, when compared to the results of frequency-based p-frame measure, produced a different - and more consistent - pattern of differences with consistent directionality of differences across all p-frame lengths and frequency quartiles. We suggest that TTR may be included into FS research as a reliable and informative measure.

Finally, we conducted a qualitative analysis of p-frames in order to investigate what types of p-frames that may prevail in different frequency quartiles. We focused this analysis on such instances where the directionality of differences between different writer groups was reversed from Quartile 4 to Quartile 1, and where the differences appeared to be compounded by the factor of p-frame length. In a brief discussion of the types of FSs, O'Donnell et al. suggest that the proportions of specific types of formulas extracted through different measures may vary: thus, they find more MI-defined genre specific formulas in the expert corpora, but more frequency-based (and evidently more topic-specific) formulas in the L2 undergraduate corpora. Since O'Donnell et al. did not observe the effects of expertise or L1 status on the use of p-frames, they only hypothesized that the differences could be there. In this modification study, we were able to show the qualitative (structural and functional) differences in the use of p-frames.

The results of the current study show the importance of establishing variable thresholds in extraction of FS, the usefulness of TTR as a reliable measure of p-frame variability, and the value of incorporating a qualitative analysis of p-frames in understanding the use of formulaic language among different groups.

References:

Fletcher, W.H. (2007). Implementing a BNC-comparable web corpus. In C. Fairon, H. Naets, A. Kilgariff, & G.M. de Schryver (Eds.), *Building and Exploring web Corpora*. *Proceedings of the WAC Conference* (pp. 43-56). Louvain: Presses Universitaires de Louvain.

McEnery, T., & Hardy, A. (2014). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

O'Donnell, M., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. International Journal of Corpus Linguistics, 18(1), 83-108.

Wood, D. (2015). Fundamentals of Formulaic Language: An Introduction. London: Bloomsbury Publishing.

Investigating fluency variables in learner language: Methodological concerns

Hege Larsson Aas, Susan Nacey
Inland Norway University of Applied Sciences
hege.aas@inn.no, susan.nacey@inn.no

This paper discusses methodological concerns related to the identification and analysis of fluency variables in studies of spoken native and interlanguage corpora. The specific focus here is pause behaviour observed in relation to interlanguage fluency, an area that has received growing attention in recent years. The paper addresses the following research question: How can a spoken learner corpus be compiled to make valid claims about utterance fluency variations?

Pauses, like other phenomena associated with hesitation, can be "welcome as overt, measurable indications of processing activity which requires a certain amount of time" (Chafe, 1980, p. 170). The frequency and duration of pauses can be studied independently as potential hesitation phenomena, but pauses are also an important component of other variables often measured in fluency research, such as mean length of run (MLR) and phonation-time ratio. These measures are viewed as *utterance fluency* variables in Segalowitz' (2010) trifold definition of second language fluency, described as any observable feature of the utterance that can potentially indicate a speaker's ability to process language (*cognitive fluency*), and/or affect listeners' perceptions of the same speaker's fluency level (*perceived fluency*):

> "it is not possible to globally characterize a person's L2 speech as "fluent" in some unidimensional, absolute fashion. All that one can say at this point is that under such and such circumstances a person's L2 speech has certain objectively measurable characteristics and that these can be interpreted by listeners to be fluent or dysfluent in particular ways" (Segalowitz, 2010, p. 39).

Contrastive studies of native and interlanguage speech production (e.g. Ginther et al., 2010; Götz, 2013) typically reveal between-group differences in measurements based on pause identification, and these differences are often seen as "fluency gaps" (Segalowitz, 2010), reflecting differences in the cognitive fluency levels of the speakers. Conclusions about interlanguage fluency from such studies necessarily rest upon the transcriptions and annotations of the spoken language under study, e.g. that the pauses analysed reflect what was actually said (or not said, in the case of unfilled pauses). Consequently, in studies based on transcription data, views of pause behaviour across languages may easily be constrained by the choices made at the transcription stage of spoken corpus compilation. A valid and reliable transcription of pauses in spoken language data – in particular data from spoken conversations – requires overt consideration of a number of issues that may not be immediately obvious. These include (a) the presence of initial silences, (b) pauses that occur in conjunction with overlapping speech and backchanneling, and (c) end-of turn pauses (when does a turn end?). As observed by Du Bois et al. (1992), "in some cases, the question of who a pause belongs to, how long it lasts, and even whether it has occurred in

a specific place, become subtly and inextricably linked to the interpretation of turn-taking and overlapping between speakers" (p. 42).

In an attempt to bring this perspective to the forefront, we present examples from our data to illustrate the various challenges involved. Our point of departure is the compilation of two related spoken corpora: the unpublished Norwegian version of the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin, De Cock, & Granger, 2010), and its smaller counterpart (NorwC) consisting of interviews with some of the same speakers speaking in their L1 Norwegian (*NL1,* cf. Gilquin (2008)). Based on a close analysis of the transcription of pause behaviour in six interviews, the paper argues that speech production as a whole should not be considered in isolation, and that utterances should not be viewed as independent from their immediate co-text. It suggests alternative transcription conventions, involving the segmentation into turns and utterances according to a set of criteria which includes discriminating between contributing and non-contributing utterances (cf. Linell & Gustavsson, 1987). The segmentation approach presented here is a step towards combining a dialogical analysis with an exploration into a specific utterance fluency variable, which in turn may contribute to a more valid description of fluency variations, and a more comprehensive view of fluency in both native and interlanguage speech.

References:

Chafe, W. L. (1980). Some reasons for hesitating. In H. W. Dechert & M. Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton Publishers, 169-180.

Du Bois, J. W., Cumming, S., Schuetze-Coburn, S., & Paolino, D. (Eds.). (1992). *Santa Barbara Papers in Linguistics, Discourse transcription* (Vol. 4). Santa Barbara: Department of Linguistics, University of California.

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.

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer. Detection, explanation, evaluation. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 3-33). Amsterdam: Brill | Rodopi.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing,* 27(3), 379-399.

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

Linell, P., & Gustavsson, L. (1987). *Initiativ och respons: Om dialogens dynamik, dominans och koherens.* Linköping: Department of Communication Studies.

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York: Routledge.

Is there a correlation between form and function? An investigation of the introductory *it* pattern in non-native-speaker and native-speaker academic writing

Tove Larsson
Uppsala University
tove.larsson@me.com

The introductory *it* pattern, as in *It is interesting to note the difference*, is an important tool used by academic writers for a range of different purposes. For example, it is often used for information-structural purposes or as a means of persuading the reader of the validity of one's claims. This paper aims to investigate the interaction of functional and syntactic characteristics of the pattern in academic writing by non-native-speaker (NNS) and native-speaker (NS) students in linguistics and literature, as outlined below.

The introductory *it* pattern, which includes an introductory *it* and an extraposed clausal subject (Quirk et al., 1985:1391; Larsson, 2016), enables writers to comment on the content of the extraposed clause, for example by using adjectives such as *interesting*. However, while the pattern is used frequently by expert academic writers, it has been found to be problematic for learners (Hewings & Hewings, 2002; Römer, 2009). For example, with regard to its functional distribution, learners have been reported to have a tendency to underuse the pattern to hedge claims (e.g. *it might be that…*) compared to NS-student and expert writers (Hewings & Hewings, 2002; Larsson, under review). Learners have also been found to struggle with making appropriate use of high-frequency syntactic realizations of the pattern, such as subject-verb-complement (SVC: *it is important to…*) and subject-verb (SV: *it seems that…*) (Larsson, forthcoming; see Quirk et al., 1985:1392 for an overview of the syntactic types). The question thus arises whether the functional and syntactic distribution might be linked, as is put forth by theories such as Pattern Grammar (Hunston & Francis, 2000). Very limited attention has been given to investigations of whether there is a correlation between the functional and syntactic distribution of this pattern in academic writing and whether this distribution differs across disciplines. In order to shed light on these research issues, the present study investigates the following research questions:

- Is there a statistically significant correlation between the function and syntactic form of the pattern?
- Are there any disciplinary differences (linguistics vs. literature)?
- What differences and similarities can be found when comparing the NNS and NS corpora? Do the NNS students use any such form-function pairings similarly to the NS students?

Subsets from three corpora are used in the present study: ALEC, BAWE and MICUSP. ALEC is a recently compiled corpus of mainly L1 Swedish learner writing that, in contrast to most learner corpora, allows the investigator to control for levels of achievement (i.e. what grade the paper was awarded); the reference NS corpus is composed of subsets of BAWE

and MICUSP. All the papers included were written by students in linguistics or literature, which are two disciplines that are typically placed in the same language department in a European setting, despite their apparent differences.

The study uses Quirk et al.'s (1985:1392) syntactic classification, as well as a functional model developed in Larsson (under review). While the functional model builds on previous models by Hewings & Hewings (2002) and Groom (2005), it diverges from these by limiting the reliance on word semantics, with the intent of increasing the replicability of the results. Using R (R Core Team 2016), *multinomial log-linear models* were fitted onto the data to test for statistical significance.

The results show that while there is a correlation between form and function for some categories of the introductory *it* pattern, this does not extend to all categories. For example, *neutral observations* are significantly more likely to be realized as SV_{pass} (e.g. *it can be seen in Figure 1 that*…) than through any other form, whereas the *hedging* function can be realized through three different syntactic types: SV (e.g. *it seems that*…), SV_{pass} (e.g. *it could be argued that*…) and SVC (e.g. *it is possible that*…). However, the results also show that unlike the NS students who use all three forms, the learners, in particularly in linguistics, tend to mainly use the SV type to hedge claims, which might at least partially explain the underuse of this functional category. The findings of the present study will not only lead to a deeper understanding of the uses of the introductory *it* pattern, but will also help facilitate more targeted, discipline-specific teaching.

References:

Advanced Learner English Corpus (ALEC). Corpus compiled at Uppsala University in 2013.

British Academic Written English (BAWE). Corpus compiled at the Universities of Warwick, Reading and Oxford Brookes in 2004–2007. http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4 (3), 257–277.

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

Hunston, S. & Francis, G. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Larsson, T. (2016). The introductory *it* pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22(2016), 64–79.

Larsson, T. (forthcoming). A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg & V. Wiegand (Eds.), *Corpus Linguistics, Context and Culture*. Berlin: De Gruyter Mouton.

Larsson, T. (under review). A functional classification of the introductory *it* pattern: Investigating academic writing by non-native-speaker and native-speaker students.

Michigan Corpus of Upper-level Student Papers (MICUSP). Ann Arbor, MI: The Regents of the University of Michigan. Corpus compiled at the University of Michigan in 2009. http://micusp.elicorpora.info/about-micusp

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, UK: Longman.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7 (1), 140–162.

The effects of non-lexical factors on lexical complexity measures applied to FL learners' and native speakers' texts

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

In the last ten years there has been an increased interest in the application of various measures of lexical, phraseological and syntactic complexity of FL learners' written and oral production for assigning students to proficiency levels. One type of studies have attempted to examine if automatically-computed indices converge with marks assigned to learners' essays by human raters. Another line of research has investigated whether the indices can discriminate between the groups of learners already established to represent different proficiency levels, as well as between FL learners and native speakers (NS). All these inquiries were based on an assumption that the complexity of learners' production depends directly on their linguistic competence and the indices which capture this complexity in an objective and systematic way can be used to gauge the command of a foreign language. Little attention has been given to other factors which can influence learners' choice of vocabulary, phraseology and syntax, including the text's purpose, audience, genre, style, topic or students' motivation (for counterexamples see Bouwer, Béguin, Sanders, & Bergh, 2015; In'nami & Koizumi, 2016).

This paper will present selected results from a large-scale study which point to the influence of these non-lexical factors on the values taken by various lexical indices. The aim of the project was to investigate which of several automatically-computed indices of lexical complexity proposed in the literature in the last twenty years can discriminate between two advanced groups of learners of English at different proficiency levels and English NS. The studies conducted to date have rarely attempted such a comparison simultaneously (see Tidball & Treffers-Daller, 2007), focusing either solely on learners at two different levels or comparing a homogeneous group of FL learners with NS.

The data used in this study were drawn from the PELCRA Learner English corpus (PLEC, Pęzik, 2012). It included two sets of 50 argumentative essays produced by Polish English-major students in Year 1 and 4 of their studies. The essays were written on the same topic (The curse or the blessing of the mobile phones) and in identical conditions (timed in-class writing). This collection was supplemented with 50 essays also written on the same topic and in similar conditions by American NS students at the university level. Twenty-four lexical indices were computed for each essay with the help of publically-available software for lexical analysis (Lexical Complexity Analyser, Coh-metrix and TAALES). The values for these indices were compared between the three groups of essays using ANOVA or Kruskal-Wallis statistical tests and post-hoc analyses.

The comparison of the indices computed for the three groups of texts produced very interesting results which in part ran counter to the expected outcome. Only five indices – selected gauges of lexical diversity and sophistication – demonstrated statistically significant differences between the three groups. A large size effect was observed in each

of these cases. The difference between the indices in the two groups of language learners confirmed the predicted direction of change. It indicated that Year 4 students produced more diverse texts, as well as used a larger proportion of a range of advanced lexemes and more academic vocabulary than Year 1 learners. This finding confirmed the results produced by earlier studies (e.g. Laufer & Nation, 1995; Daller & Xue, 2007; Tidball & Treffers-Daller, 2007). However, the same five indices rendered very surprising results for NS. Contrary to intuition and the outcomes of earlier studies (e.g. Linnarud, 1986; Vermeer, 2004) the native essays displayed significantly lower values than those of both groups of learners. This means that American students had written the least diverse and the least lexically advanced texts of the three analysed groups of writers. The same effect was observed for 18 other analysed measures referring to various lexical characteristics of the texts, which only demonstrated the difference between FL (as a whole) and NS essays.

A qualitative scrutiny of the analysed samples conducted by four human raters offered a plausible explanation of such a counter-intuitive disparity between the essays written by FL learners and NS. It demonstrated that the low values of lexical indices obtained for the NS texts may be a consequence of the different genres applied by the Polish and American students in developing the same topic. The Polish writers approached the task as an academic essay whereas the American writers treated it as a sample of journalistic prose, and the two genres sanction the use of a distinct style. Moreover, it can be suggested that the Polish students, as experienced foreign language learners, were likely to be more aware of the dual function of writing in language assessment. They might have been conscious that their written production may be used not only to evaluate their effectiveness as writers but also their proficiency in language use. Therefore, they were more likely to make an effort to present their full linguistic potential, while the native students may not have felt such a need.

The results of the study demonstrate that the values of lexical measures computed for FL learners' and native speakers' texts are not a straightforward reflection of writers' lexical competence and can be influenced by a range of other factors which go beyond their linguistic command. This finding serves as a warning that various automatically-computed indices need to be applied very cautiously to gauge a FL learner's proficiency.

References:

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1), 83–100.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press. 150–164.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing,* 33(3). 341–366.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16(3), 307–322.

Linnarud, M. (1986). Lexis in Composition: A Performance Analysis of Swedish Learners' Written English. Malmö: C.W.K. Gleerup.

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

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press. 133–149.

Vermeer, A. (2004). The Relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.). *Vocabulary in a Second Language*. Amsterdam: John Benjamins. 173–189.

Phrasal Verbs in Spoken L2 English: The Effect of L2 Proficiency and L1 Background

Irene Marin-Cervantes, Dana Gablasova
Lancaster University
i.marincervantes@lancaster.ac.uk, d.gablasova@lancaster.ac.uk

Phrasal verbs (PVs) are a prominent component of multi-word expressions that are ubiquitous in the English language and lie at the heart of successful communication. This study examines the use of phrasal verbs in spoken, interactive communication of L2 English users. In particular, it addresses the following research questions: 1) What is the effect of English proficiency on the frequency of PVs in L2 production? 2) What is the effect of a particular L1 background on the frequency of PVs in L2 production? In addition to complementing and extending previous findings about L2 learners' use of PVs, the study aims to demonstrate trends in the frequency of PV use of L2 speakers from a wide range of proficiency levels and L1 backgrounds when engaged in spoken, interactive communication, an area which has not yet been systematically studied at a large scale.
The study is based on the data from the Trinity Lancaster Corpus (TLC), a large corpus of spoken L2 English production recorded in a semi-formal, institutional setting (Gablasova et al., 2015). The corpus consists of more than 3.5 million tokens of L2 production from 1449 speakers from three proficiency levels: B1, B2 and C1/C2 of the Common European Framework of Reference, and from eight different L1 backgrounds (e.g. Spain, China, Italy, Russia, etc.). A one-way ANOVA and Bonferroni post-hoc test were used to establish the effect of the independent variables (L2 proficiency and L1 background) on the frequency of PVs in learner speech.
The findings indicate that the effect of L2 proficiency on the production of PVs was statistically significant and that the number of PVs increased with speaker's proficiency. While the results show that PVs represent a relatively small proportion of the language used by L2 speakers at the lower-intermediate and intermediate levels, the overall frequency of PVs in the language produced by advanced speakers in the TLC was found to resemble that reported for native speakers of English in conversation (Biber et al., 1999). However, the range of the most frequent PVs remains largely unchanged across proficiency levels. A subsequent analysis of the meaning of the twenty most frequent PVs in the corpus revealed that L2 learners in the TLC seem to be unaware of the highly polysemous nature of these verbs and tend to associate an average of one or two meanings (often their core meaning) with those verbal forms. The analysis also showed that intermediate learners appeared unaware of the contribution of particles to the meaning of PVs.
The effect of L1 background on PV frequency was also found to be statistically significant. The Bonferroni post-hoc comparisons revealed that the main difference was between Chinese and Spanish speakers, with speakers of Chinese producing, on average, more PVs than speakers from the other L1 backgrounds analysed. These findings are explained in light of what previous experimental and corpus-based studies have reported regarding the role of cross-linguistic differences (Chen, 2013). It is argued that factors other than L1-L2

syntactic correspondence have an impact on the production of PVs in L2 speech (e.g. L2 exposure or previous language training).

Ultimately, important pedagogical implications are discussed, namely the need i) to introduce and practice PVs early on in the language learning process given the importance of PVs and other multi-word expressions for the development of fluency in spoken production (Wray, 2012; Siyanova-Chanturia & Martinez, 2015), ii) to pay further attention both in research and teaching practice to appropriate use of PVs and depth of PV knowledge as well as to frequency and range of PV production, and iii) to carry out further corpus-based explorations of PV knowledge for understanding the acquisition and use of multi-word constructions in L2 speech.

References:

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

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

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, Advance Access.

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, 36(5), 549-569.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32, 231-254.

L2 French learners' longitudinal morphosyntactic development: A Conceptual replication

Kevin McManus[1], **Rosamond Mitchell**[2]
Pennsylvania State University[1], **University of Southampton**[2]
kmcmanus@psu.edu, r.f.mitchell@soton.ac.uk

French SLA research continues to benefit from oral learner corpora to understand (i) the learner's linguistic repertoire as a whole and changes within in it as well as (ii) the development and use of a specific linguistic feature (e.g., Bartning 1997, 2009, 2016; Myles 2005, 2008, 2015). Drawing on the InterFra and Lund oral learner corpora made up of longitudinal and cross-sectional oral interviews and retellings of films and comic strips from 58 Swedish-speaking learners of L2 French, Bartning and Schlyter (2004) proposed developmental sequences for eight morphosyntactic features: (1) verbal morphology agreement, (2) tense, (3) mood, (4) aspect, (5) negation, (6) object pronouns, (7) gender agreement, (8) subordination. Using their multi-stage, developmental analyses for each of the eight linguistic features, Bartning and Schlyter additionally proposed a series of grammatical profiles called 'the advanced learner variety' based on learners' use of these eight morphosyntactic features.

In this paper, we present a conceptual replication of Bartning & Schlyter designed to address these limitations. Importantly, a conceptual replication¸ as noted by Porte (2012), typically draws on a different research design to verify the original findings. As a result, our research questions are the exact same as the original study, namely to describe the different stages of development for each of the eight morphosyntactic features: (1) verbal morphology agreement, (2) tense, (3) mood, (4) aspect, (5) negation, (6) object pronouns, (7) gender agreement, (8) subordination.

Participants were 27 English-speaking learners of French L2 majoring in French at a UK university. Mean age was 21 (range: 20-24 years), mean length of previous French study was 11 years (range: 9-15 years), and mean age of first exposure to French was at 9.5 years old (range: 0-15 years old). Six participants were workplace interns, fifteen were teaching assistants, and six were university exchange students, all situated throughout France including small and large, urban and rural French cities.

Our corpus design included only longitudinal speech samples collected from semi-structured L2 interviews. Each interview was administered by a member of the research team at each of the six data collection points. All interviewers used the same list of pre-established questions focusing on sojourners' experiences and opinions about the sojourn. Each interview lasted approximately 20 minutes. All interviews were digitally recorded and later transcribed following CHAT conventions (MacWhinney, 2000). Each transcript was then checked for transcription accuracy by at least two members of the research team before analysis. All files were automatically tagged and manually disambiguated using CHAT's part-of-speech morphosyntactic tagger (MOR and POST commands).

Our findings indicate ongoing, long-term development on all features, with different rates of development between different structures (e.g. grammatical gender vs. mood). Our clear gains in some areas of development will be contrasted with others showing less

development. Lastly, drawing on usage-based explanations of learning (Ellis 2006, O'Grady 2015), we will contextualize our findings in terms of L1-L2 processing differences, the roles played by frequency, saliency, and contingency in L2 learning, as well as the roles played by differences in context (abroad vs. at home).

Extraction of unsuitable pragmalinguistic features of requests produced by Japanese learners of English with low proficiency

Aika Miura
Tokyo University of Agriculture
dawn1110am@gmail.com

This study presents how learner corpora can be applied to the investigation of interlanguage pragmatics, which has been mainly based on elicitation tasks such as the Discourse Completion Task (e.g. Blum-Kulka, House, and Kasper, 1989; Flores Salgado, 2011; Hill, 1997; Trosborg, 1995; Schauer, 2009). The author (Miura, 2015; in press) developed a multi-layered annotation scheme based on the coding scheme of Blum-Kulka et al. (1989) and studied *criterial* pragmalinguistic features of Japanese learners of English by examining the requests made in a shopping role play in the NICT JLE Corpus, which contains written transcripts of an oral interview test called the Standard Speaking Test. In this, the interlocutor role-plays a shop assistant, and the test-taker is given the task of purchasing a particular item as a customer. It was observed that learners with lower proficiency used *direct* request strategies (e.g. the desire verb 'want'); however, the use of *conventionally indirect* strategies (e.g. the ability/permission modal 'can') increased with the enhancement of proficiency. The results confirmed several previous studies (Flores Salgado, 2011; Hill, 1997; Trosborg, 1995).

In this study, the author presents linguistic features that specifically represent unsuitable learner data in different developmental stages. The data of 68 CEFR A1-level and 114 CEFR A2-level learners were investigated and nine situations where requests occurred were identified. The extracted linguistic features of the requests are shown in the following frequently occurring situations: when the customer (i) expresses their intention to purchase a particular item and (ii) makes an enquiry or requests further information regarding an item.

The study addresses the following research questions:

(1) What proportions of suitable and unsuitable uses of linguistic features are produced by learners at two different proficiencies in each situation?

(2) What types of unsuitable linguistic features of two different proficiency levels are observed in each situation?

The author modified the coding scheme of Blum-Kulka et al. (1989) to analyse the learners' requests specifically observed in shopping situations in the corpus. The requests were manually identified and classified into either direct or conventionally indirect strategies, depending on the choice of lexico-grammatical features, as shown in Table 1. Among them, the *statement*, *not-classifiable*, and *yes/no* categories were specific to the learners' unsuitable production, and *existence* (such as *'Do you have another one?'*) and *intention* (such as '*I will have it.'*) were specific to a shopping situation, which were added to the original coding scheme.

Table 1. Annotation Scheme for Extracting Pragmalinguistic Features of Requests

Direct Strategies	Conventional Indirect Strategies
• Obligation: have to/must • Non-sentential phrase • Desire: want/need/would rather/would like • Imperative • Statement • Not-classifiable • Yes/No	• Ability/Permission: *can/could/may* • Willingness: will you/would you/would you mind • Possibility: *Is it possible* etc. • Suggestory: *how about* etc. • Subjectivizer: *hope that* etc. • Existence • Intention

As a result, in Situation (i), A1 group presented 67 direct (69.79%) and 29 conventionally indirect (30.21%) strategies, whereas A2 showed 105 direct (56.45%) and 81 conventionally indirect (43.55%) patterns. Further, 31 requests (10.99%) contained unsuitable features, including patterns such as 'I want to some guitar.' (addition of a to-infinitive), 'Yes, I want.' (omission of a noun), and 'I wanted to get a baby's present to my friend.' (incorrect choice of tense) in the *desire* subcategory, as well as 'I buy this suits.' (omission of modals or tense inflections) in the *statement*; 'Buy it.' in the *not-classifiable* category; and 'Yes', which was a response to the interlocutor's prompt 'Would you like this?', in the *yes/no* subcategory. Regarding Situation (ii), A1 group showed 94 direct (63.09%) and 55 conventionally indirect (36.91%) strategies, whereas A2 showed 179 direct (55.94%) and 141 conventionally indirect (44.06%) patterns. Among them, 72 requests (15.35%) contained unsuitable linguistic features, including 46 *statement* patterns (12.75%), for example, 'My size is M.' and 'And its color is black.', which seemed to be influenced by the topic-comment structure of Japanese (L1).

References:

Blum-Kulka, S., House, J., & G. Kasper (1989). *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.

Flores Salgado, E. (2011). The Pragmatics of Requests and Apologies: Developmental Patterns of Mexican Students. Amsterdam: John Benjamins.

Hill, T. (1997). *The Development of Pragmatic Competence in an EFL Context.* (Unpublished Doctoral Dissertation.) Tokyo: Temple University.

Miura, A. (2015). Criterial features of pragmatic competence in a spoken corpus of Japanese learners of English: Distinguishing different levels of proficiency. *Language, Area and Culture Studies*, 21, 147-171.

Miura, A. (in press). Assessing politeness of requestive speech acts produced by Japanese learners of English in a spoken corpus. *Language Value*, 9.

Schauer, G. (2009). Interlanguage Pragmatic Development: The Study Abroad Context. London: Continuum.

Trosborg, A. (1995). *Interlanguage Pragmatics. Requests, Complaints and Apologies.* Berlin/New York City: Mouton de Gruyter.

From pedagogical input to learner output: How teaching materials affect the use of the English passive in learner writing

Verena Möller
Université catholique de Louvain
verena.moeller@uclouvain.be

In her assessment of the contribution of learner corpus research to the design of teaching materials, Granger (2015a: 494) states that "[w]hile learner corpus data has had a significant though still modest effect on dictionaries and grammars, its impact on coursebooks has been more nominal than real." So far, textbooks have mostly been analysed in terms of how well they reflect authentic language. Römer (2005: 296) identifies what is generally missing: "A third component that I also would have liked to include in my comparative analysis, in addition to what I have termed 'real' and 'ideal' language learner *input* [...], is language learner *output*."

The present study analyses how the language represented in teaching materials affects the output produced by learners. In the framework of *Contrastive Interlanguage Analysis* (CIA, Granger 2015b: 17), we compare learner texts and teaching materials derived from *English as a Foreign Language* (EFL) and *Content and Language Integrated Learning* (CLIL) contexts. Interlanguage text data were collected from three groups of learners: Participants in EFL and CLIL (CLIL+), learners who had taken part in EFL, but had chosen not to participate in CLIL (CLIL-), and learners to whom only EFL lessons were available (CLIL0). Learners were attending Year 11 and, according to the EFL syllabus (cf. MKJS 2004: 109), were therefore expected to have attained levels B1 to B2 of the CEFR. Learner texts were compiled in a corpus of argumentative essays, the *Secondary-Level Corpus of Learner English* (SCooLE). Three reference language varieties were collected in the *Teaching Materials Corpus* (TeaMC): CLIL materials for Year 7-10, EFL materials for Year 7-10, and EFL materials for Year 11/12, with the latter category being subdivided according to genre (i.e. imaginative, argumentative, and informative texts). In addition, a subcorpus of the *Louvain Corpus of Native English Essays* (LOCNESS) was used because "[n]ative controls performing the same tasks as the learners should be included" (Myles 2015: 329).

For the comparison of input and output, we have chosen the English passive as it is both a genre marker (cf., for instance, Svartvik 1966, Granger 1983) and a proficiency marker (cf., for instance, Kameen 1993, Granger 2013). In addition to passive frequency, we analysed the adjectivalness cline, i.e. the prevalence of central passives (e.g. *he was killed)* and non-central passives such as attitudinal/emotive passives (e.g. *he was interested/he was annoyed*, cf. Svartvik 1966: 134) and statal passives (e.g. *the building is demolished*, cf. Granger 1983: 114).

The highest passive frequencies in the TeaMC were found in scientifically oriented CLIL materials and informative EFL materials, thus mirroring findings from previous research (cf. Svartvik 1966: 155, Biber et al. 1999: 476). These genres comprised comparatively few non-central passives, amongst which statal combinations prevailed. The lowest passive

frequencies were observed in imaginative EFL materials in the Year 7-10 and the Year 11/12 subcorpora, which displayed a notable proportion of non-central passives, consisting to almost equal parts of statal and attitudinal/emotive passives.
Data from the SCooLE reveals that CLIL participants produced a significantly greater number of passives than non-CLIL learners, which made their texts more target-like when compared to LOCNESS and the argumentative subcorpus of the TeaMC. The number of erroneous passives was found to be lowest in this group as well. Proportions of non-central passives in CLIL+ texts and LOCNESS were almost identical. CLIL+ learners favoured statal passives over attitudinal/emotive passives, again reflecting input from CLIL materials and resembling LOCNESS texts. Output produced by the two non-CLIL groups was more similar to imaginative EFL materials regarding passive frequency and the adjectivalness cline. Results suggest that EFL materials do not contain sufficient input to help non-CLIL learners produce texts which are target-like with respect to the passive. We are by no means suggesting that teaching materials are the only source of differences between CLIL and non-CLIL students, as multiple selectivity issues regarding CLIL have been revealed (cf. Möller forthcoming 2017). Following Meunier & Reppen (2015: 514), however, who state that it should "be part of the corpus linguists' agenda to provide textbook writers with clear guidelines", we suggest that a wider variety of genres should be included in EFL materials from the early stages of L2 acquisition, allowing for a realistic representation of the passive, and thus counteracting transfer of frequency phenomena across genres.

References:

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

Granger, S. (1983). *The be + past participle Construction in Spoken English*. Amsterdam: Elsevier Science.

Granger, S. (2013). The passive in learner English. Corpus insights and implications for pedagogical grammar. In S. I. Ishikawa (Ed.). *Learner Corpus Studies in Asia and the World. Vol. 1. Papers from LCSAW2013.* Kobe: School of Languages and Communication, Kobe University, 5-15.

Granger, S. (2015a). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: CUP, 485-510.

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

Kameen, P. T. (1993). Syntactic skill and ESL writing quality. In A. Freedman (Ed.). *Learning to Write: First Language, Second Language. Selected Papers from the 1979 CCTE Conference*, Ottawa, Canada, 162-170.

Meunier, F. & Reppen, R. (2015). Corpus versus non-corpus-informed pedagogical materials: grammar as the focus. In D. Biber & R. Reppen (Eds.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: CUP, 498-514.

MKJS Ministerium für Kultus, Jugend und Sport Baden-Württemberg (2004). *Bildungsplan 2004. Allgemein bildendes Gmynasium*. Ditzingen: Philipp Reclam Jun.

Möller, V. (forthcoming 2017). *Language Acquisition in CLIL and Non-CLIL Settings: Learner Corpus and Experimental Evidence on Passive Constructions*. Amsterdam: John Benjamins.

Myles, F. (2015). Second Language Acquisition theory and Learner Corpus Research. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 309-331.

Römer, U. (2005). *Progressives, Patterns, Pedagogy. A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts and Didactics.* Amsterdam/Philadelphia: John Benjamins.

Svartvik, J. (1966). *On Voice in the English Verb*. The Hague/Paris: Mouton.

Using a learner corpus to support online intelligent tutoring: the Alegro project

Penny MacDonald[1], Michael O'Donnell[2]
Universitat Politècnica de València[1], Universidad Autónoma de Madrid[2]
penny@idm.upv.es, michael.odonnell@uam.es

This talk will outline the use of a learner corpus as a source of information within an intelligent online learning system. This system, currently being developed within a Spanish national project, is called Alegro (Adaptive Learning of English Grammar Online). The system keeps track of the grammatical concepts that have been acquired by each learner, those which they are still developing, and those which they have yet to acquire. In operation, the system will keep the learner working within their Zone of Proximal Development (Vygotsky, 1978), developing those areas which the learner is ready for but has not yet fully developed. Prior research has shown that when students are focused within this ZPD, they are maximally engaged (e.g., Hamilton & Cherniavsky 2006), and, in such a state of flow, learning is maximised (Csikszentmihalyi 1988).

To support this functionality, we have been making use of a large learner corpus of texts, written by Spanish learners of English at University level (these are at least initially also the target users of the online system). The corpus consists of two parts, the Wricle Corpus (Rollinson & Mendikoetxea 2010) and the UPV Learner Corpus (Andreu Andrés et al. 2010). Wricle involved students in an English Studies degree, while the UPV corpus involved students in English for Specific Purposes courses. The combined corpus consists of 730,000 words.

Both corpora have proficiency information associated (CEFR levels, derived using the Oxford Quick Placement test). Each text has been automatically tagged for syntactic structure (cf. O'Donnell 2012). Additionally, a 112,000 word subset of the corpus has been tagged for
linguistic errors, identifying 16,000 errors.

Both the manual error tagging and the automatic syntactic tagging are used within the Alegro system to help provide an intelligent and targeted experience for the learner. Firstly, the error annotation has been used to identify "critical language areas": there are large numbers of linguistic features involved in a language, but some features are more problematic than others for language learners from a particular mother tongue. We have used our error analysis to identify exactly those linguistic areas where the learners make the most errors. The learning system focuses on these areas, rather than spending time teaching structures which are not problematic for the learner. We have identified a list of the 15 most frequent language errors produced by our students (the error scheme is detailed, including over 127 error types).

We have found however that the identification of these critical language areas is not enough: within each area, there are a number of subsidiary concerns that need to be addressed. For instance, the most frequent error for Spanish learners of English is to use an article where it is not appropriate (cf. Diez Bedmar, 2010a). However, to properly teach this area, we analysed the 1087 instances of this error in our corpus, to derive exactly the

linguistic concepts not understood by the learner in producing the error. Here, more delicate coding revealed (in line with earlier work) that the vast majority of the errors occurred with generic plurals ("The Cats are a mammal") and generic uncountable nouns ("The black is my favourite colour."). The
in-depth more delicate coding within each of our critical language areas is producing a longer list of "critical language concepts" that need to be acquired by our learners.
A second way we are using our learner corpus is to derive an understanding of the relative difficulty of each of the critical language concepts in relation to each other. As this work is covered elsewhere, only a brief summary of how the learner corpus is used for this purpose is dealt with here. The list of critical concepts ordered in difficulty is essential in the system, as it is this which allows us to locate the zone of proximal development of learners (exactly the least difficult concepts which are not yet fully acquired).

References:

Andreu M.A., Astor A., Boquera M., MacDonald P., Montero B. y Pérez C. (2010). Analysing EFL Learner Output in the MiLC Project: An error *it's, but which tag?. In M.C. Campoy, B. Belles-Fortuno and M.L. Gea-Valor (Eds.). *Corpus-based Approaches to English Language Teaching*. London: Continuum, 167-180.

Csikszentmihalyi, M. (1988). The Flow Experience and its Significance for Human Psychology. In M. Csikszentmihalyi and I. S. Csikszentmihalyi (Eds.). *Optimal Experience: Psychological Studies of Flow in Consciousness*. London: CUP, 15-35.

Díez-Bedmar, María Belén (2010a). From Secondary School to University: The Use of the English Article System by Spanish Learners. In Belles-Fortuño, Begoña, Campoy, María Carmen & Gea-Valor, Lluisa (Eds.). *Exploring Corpus Linguistics in English Language Teaching*. Castelló: Publicacions de la Universitat Jaume I, 45–55.

Hamilton, E. & J. Cherniavsky (2006). Issues in synchronous versus asynchronous E-learning platforms. In H. O'Neill and R. Perez (Eds.). *Web-Based Learning: Theory, Research and Practice.* Mahwah, NJ, Lawrence Erlbaum, 87-106.

Murcia Bielsa, Susana & MacDonald, Penny (2013). The TREACLE project: Profiling learner proficiency using error and syntactic analysis. In S. Granger, G. Gilquin and F. Meunier (Eds.). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use* – Proceedings 1. Louvain-la-Neuve: Presses universitaires de Louvain.

O'Donnell, Mick (2012). Using Learner Corpora to Redesign University-level EFL Grammar Education. *Revista Española de Lingüística Aplicada* (RESLA), Vol. Extra 1, 145–160.

Rollinson, Paul & Mendikoetxea, Amaya (2010). Learner Corpora and Second Language Acquisition: Introducing WriCLE. In Bueno Alonso, Jorge L. et al. (Eds). *Analizar datos: Describir variación/ Analysing Data: Describing Variation*. Vigo: Universidade de Vigo, 1–12.

Vygotsky, L. S. (1978) Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Development of L2 metaphorical competence from ages 10-19

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 we produce (linguistic metaphors). Empirical research has since confirmed that linguistic metaphor is ubiquitous in language (see e.g. Nacey, 2013; Steen et al., 2010). As a consequence, metaphor necessarily plays a central role in language learning, including all aspects of communicative competence in a second language (L2) (see e.g. Littlemore & Low, 2006).

This investigation details a pseudo-longitudinal corpus-based exploration into the development of metaphorical competence of L2 learners as they progress through their school career. The particular focus here is on the written production of linguistic metaphors in L2 English written by parallel groups of pupils from the ages of 10-19 in Norway, where the subject of English is obligatory from the first grade (at age six). The Norwegian government defines English as one of three 'core' subjects (along with Norwegian and mathematics), and considers it as both a key language subject and as a subject for the personal growth and development of pupils (Norwegian Directorate for Education and Training, 2013).

The particular objectives and methods are adapted from a Littlemore *et al*. (2014) investigation into the metaphor use of Greek and German-speaking learners of English with varying degrees of English proficiency. More specifically, the MIPVU metaphor identification procedure will first be applied to 180 texts (20 per grade level), to identify all linguistic metaphors in these texts (see Steen et al., 2010). The main objective is to measure how the metaphorical density varies per grade level - that is, variation in number of linguistic metaphors per lexical unit. A second goal is to compare patterns for open-class versus closed-class metaphors across grade levels, to identify whether any particular level at which the use of the former overtakes the latter as has been observed in previous research (Littlemore et al., 2014).

The empirical data is retrieved from the "Tracking Written Learner Language" corpus (TraWL), a compilation of authentic texts written by Norwegian pupils. TraWL is a longitudinal corpus, currently under compilation as part of a wider, ground-breaking project into the development of L2 writing in the Norwegian school system. The corpus consists of texts written in L1 Norwegian, L2 English, L2 Spanish, L2 German and L2 French, which are being collected from schools in differently populated geographical regions in Norway, ranging from the capital city to rural municipalities. All texts have been submitted

as class work by pupils from the fifth grade in primary school to the final year of upper secondary school.

Compilation of TraWL began in the fall of 2016 and will continue for the foreseeable future, to accommodate longitudinal studies of writing development of either individuals or groups, in the L1 and in the L2(s). As of this writing, however; AP only allows for cross-sectional (pseudo-longitudinal) studies of learners at various stages of development, starting from the fifth grade. The present investigation therefore represents an initial exploratory look into the metaphor production in second language writing from different groups of pupils from nine different grade levels, from the primary through the upper secondary school levels. This investigation is innovative, since no previous work has tracked the development of metaphorical competence from such a young age (from age 10) and over such a wide age range (until age 19).

References:

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

Littlemore, J., Krennmayr, T., Turner, J., & Turner, S. (2014). An investigation in metaphor use at different levels of second language writing. *Applied Linguistics,* 35(2), 117-144.

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

Nacey, S. (2013). *Metaphors in learner language*. Amsterdam: John Benjamins.

Norwegian Directorate for Education and Training. (2013). Læreplan i engelsk. Retrieved from http://www.udir.no/kl06/ENG1-03

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: from MIP to MIPVU*. Amsterdam: John Benjamins.

Building a Learner Corpus for Irish as part of the development of Speech Technology for Computer-Assisted Language Learning

Neasa Ní Chiaráin, Ailbhe Ní Chasaide,
Trinity College, Dublin
neasa.nichiarain@gmail.com, anchsid@tcd.ie

This paper describes how speech technology-based interactive platforms can be exploited for the development of learner corpora. This facilitates not only the design of targeted pedagogical materials, but can also be exploited in research aimed at further speech technology development.

Background: the Irish context

The Irish language has been in decline for more than a century and a half. The Irish State has, however, supported Irish since its foundation in the 1920s. Despite official support, the numbers of native speakers are in fast decline. Irish is a compulsory study for all school-going children up to the age of 18. Attitudes towards the learning of Irish are varied and at times learners make very little progress. The Irish language context is quite unique in that the number of learners in the schooling system (N=c.750K) far outweighs the number of native speakers in the country (N=c.40K, McCloskey, 2001). Native speaker models are generally not available to learners.

The ABAIR initiative is concerned with speech technology development for Irish, which includes the development of multi-dialect synthetic voices and the development of platforms where these voices may be used for language teaching purposes (www.abair.ie). The related project, CabairE, entails the development of teaching-specific facilities that will exploit the synthetic voices and resources. Future developments of the ABAIR initiative include speech recognition for Irish. To date, large spoken corpora have been gathered from three native speakers, one for each of the three main dialects of Irish. This involved the speakers reading many hours of specially selected materials written in their own dialects.

In the development of CabairE, we envisage applications that will serve a dual purpose. We are pursuing the approach that the development of pedagogically-oriented materials in CabairE will offer unique opportunities for the related building of learner corpora. In this paper, a specific learner-orientated pedagogical tool, named *An Scéalaí* (the storyteller), is described which illustrates this approach. On the one hand its purpose is to serve as a language learning tool, helping learners reflect on and self-correct both their writing and pronunciation. At the same time, it allows the researcher to collect learner data which we envisage using for (1) linguistic research (2) the development of speech technology and (3) computer-assisted language learning (CALL) tools for Irish. The remainder of this abstract describes *An Scéalaí* as its intended future application. This is currently work in progress and is at a preliminary stage of development.

An Scéalaí: pedagogical tool that facilitates learner corpus building

An Scéalaí is an online platform designed to gather a written and spoken corpus from learners of Irish, who are at B1-C1 level of the CEFR framework (Council of Europe, 2011). These include trainee teachers and senior second level school pupils. It consists of an easy-to-access webpage which can be accessed from the main ABAIR website (www.abair.ie). The main data-gathering interface consists of an open textbox into which learners can type a continuous piece of authentic text which may reflect their opinions or beliefs based on general prompts given to them by the researcher. When submitted, the learner can choose to hear the text being read back using synthetic voices and may choose to use a spelling and grammar check facility, intended to promote *noticing* (Swain & Lapkin, 1995). The errors are much more easily identified aurally than visually and in the case of Irish, where the spelling to sound correspondences are opaque, the aural feedback is invaluable. The learner can then self-correct their input and resubmit to the system. The learners' interaction with the system is being saved at 30-second intervals and researchers have access to the learner data at each of the timepoints. Learners are also be asked to record themselves reading aloud from their text and may listen back to check and self-correct pronunciation and lexical errors. *An Scéalaí* also elicits some personal characteristic of learners such as gender; L1; level of Irish; attitude towards technology; school-type, etc. *An Scéalaí* has two benefits. Firstly, it functions as a pedagogical tool in its own right. It is giving greater accessibility to Irish synthetic speech and to a specific spellchecking system and is available to learners at a time and place of their own choosing. This has been pilot tested with a cohort of trainee teachers who have reported a high degree of satisfaction with the tool. Secondly, it forms a corpus of learner Irish which has been gathered from authentic language learning situations in which the exercise is of mutual benefit to the learner and to the researcher.

Application

While there are a number of Irish language corpora available, there is not yet a corpus gathered exclusively from learners involved in the education sphere where the context and subject matter are of direct language learning relevance. The present tool will produce data that will be of interest in the following areas:

1. **Linguistics**

Error analysis: there is a need to identify both general error patterns and variations on error types in both written and oral formats, which may be associated with differences in the personal data elicited from the learners, such as L1 not being English. A contrastive analysis, which would compare the present learner data with that which was analysed by Ó Domhnalláin and Ó Baoill (1978) and Ó Baoill (1981), would also be valuable.

2. **Speech Technology Development**

A crucial next step in the ABAIR initiative will be the development of a speech recognition system for Irish. Although the development will initially focus on native speakers, we foresee that in the future a major area of application will be in the area of Irish language education. In designing a speech recognition system it is important that the system is built on language samples gathered from a large cohort representative of the targeted end

users. *An Scéalaí* will be in a position to gather considerable quantities of the necessary data type.

3. Computer-Assisted Language Learning

Among other CALL projects, work is currently underway on an interactive spoken dialogue system for the teaching/learning of Irish (Ní Chiaráin & Ní Chasaide, 2016). It comprises of a chatbot, built using Pandorabots AIML, which interacts with the learner in the target language. Currently, the learner must type into the system, as speech recognition is not yet available for Irish, and the system responds using synthetic voices. This works using pattern matching techniques where the chatbot's responses have been preprogrammed into the system and anticipated common learner errors, devised on the basis of intuition and experience, have been manually coded. *An Scéalaí* will provide invaluable information on the actual errors learners are making and will allow us to provide intelligent corrective feedback as part of the system. This will result in a much more interactive and engaging learner experience.

The information gathered from *An Scéalaí* will enable development of more targeted learning resources for learners of Irish.

References:

Council of Europe (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe

McCloskey, James. 2001. Voices Silenced: Has Irish a Future? Dublin: Cois Life Teoranta.

Ní Chiaráin, N., & Ní Chasaide, A. (2016). Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for Irish. In *10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016* (pp. 3429–3435). Portorož (Slovenia).

Ó Domhnalláin, T., & Ó Baoill, D. (1978). *Earráidí scríofa Gaeilge. Cuid 1, earráidí briathra: cuntas ar na hearráidí Gaeilge a rinne sampla ionadaíoch de na hiarrthóiri ar scrúdú Gaeilge na hArdteistiméireachta*. Baile Átha Cliath: Institiúd Teangeolaíochta Éireann.

Ó Baoill, D. (1981). *Earráidí scríofa Gaeilge. Cuid 3, réamhfhocail agus comhréir : earráidí a tharla in aistí Gaeilge na hÁrdteistiméireachta, 1975*. Baile Átha Cliath: Institiúd Teangeolaíochta Éireann.

Pandorabots. (2016). Pandorabots: A multilingual chatbot hosting service. Retrieved January 20, 2017, from http://www.pandorabots.com/

Swain, M., & Lapkin, S. (1995). Problems in Output and the Cognitive Processes they Generate: A step towards Second Language Learning. *Applied Linguistics*, *16*, 371–389.

The most probable translations explain learner errors: Arab learners' use of prepositions

Noom Ordan, Omaima Abboud
The Arab Academic College of Education, Haifa
noam.ordan@gmail.com, omaima.abboud@gmail.com

This work is concerned with mother tongue interference1 in foreign language learning, in particular the use of English prepositions by native speakers of Arabic. We use parallel data of Arabic source language units and their English target language translations to measure the translation potential, analyze different translation strategies, and eventually bring evidence that the non-native speakers' strategy is to opt for the more probable translation equivalents, even though others are possible.

In translation studies, interference has been studied extensively (Gellerstam 1986, Toury, 1979 and 1995). Although most translators translate into their native language, they seem to share many similarities with non-native speakers as hypothesized by Toury (1979), and as has been empirically shown by Rabinovich et al. (2016). When studying translations, the researcher always has access to the source text which triggered the target equivalents, whereas in second language production, there is no, so to speak, source text, and we only assume that the mother tongue of the learners affects their production (on the use of various kinds of corpora in foreign language acquisition, consider .

This work suggests that, since translations and non-native production share similar features, we can, in fact, trace back certain errors made by non-native speakers by observing parallel corpora and estimating the probabilities of lexical items in L2, given a lexical item in L1. Additionally, we conduct a small-scale experiment where we provide Arabic native-speakers from a small set of sentences from the corpus and ask them to translate from Arabic to English. We compare their performance and analyze their errors based on the parallel data (see further below).

We used two learner corpora and one parallel corpus, respectively as follows:

ArabCC: a work-in-progress, medium-scale learner corpus compiled from short essays written by English majors in the Academic Arab College for Education in Haifa, currently consisting of about 500 essays and 131,202 tokens.

A subsection of the corpus reported in Tetreault et al. (2013), which includes 900 TOEFL exams of native Arabic speakers.

The Arabic-English portion of OPUS (Tiedemann & Nygaard, 2004), which consists of about 55 million sentence pairs of English and Arabic. In particular, we focused on data from The OpenSubtitles Corpus and TheUnited Nations Corpus.

Since both learner corpora are rather small, we focused our attention on a highly frequent event, namely prepositions, following the rule of thumb that the more frequent an event is, the smaller a corpus can be in order to represent the phenomenon at hand (McEnery &

[1] Since this work is cast in the paradigm of translation studies, we prefer the term interference over the more complex term cross-language influences.

Wilson. ,2001). In particular, we picked two prepositions in Arabic – fī and bī – that are both roughly equivalent to the English in (but see more below), and while the first is more typical to Modern Standard Arabic, the second is more common in spoken dialects, particularly in the Palestinian dialect of ArabCC.
As bī is a proclitic appearing before its host, we first use a morphological analyzer to tokenize the text (Pasha et al. 2014), then estimate the probabilities of English equivalents of fī and bī using an open-source software package (Och & Ney 2000). The output of this system are words or phrases in Arabic, their translations to English, and the probability for each translation to occur in the parallel corpus on which the algorithm was trained. The result is known as a phrase table, from which we extract, at this stage, all the fī and bī occurrences, their possible equivalents in English, and their probabilities.

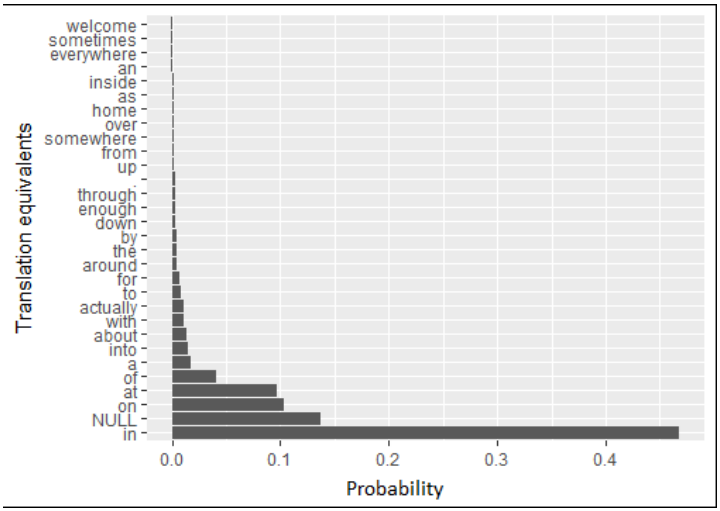


Figure 1: translation equivalents of the Arabic preposition fī

The first striking finding is that the number of possible translations is immense. Like many phenomena in natural language, the distribution is Zipfian (Baroni 2009), such that a few events hold most of the probability mass and many, although possible, are not very probable (cf. Tsui 2004, in the context of second language acquisition). The plot (Figure 1) shows that the most probable translation of fī is in, which will feature as the most common error in non-native production and that the second most probable translation is NULL, i.e. there is no corresponding word in English in many cases (just as, for example, when translating the definite article the from English to Russian, omission should take place).
We categorize translation equivalents in different sub-cases. For the cases where no translation equivalent is required, we note that strategy is rarely adopted by non-native speakers. We then move on to the other equivalents, categorized to sub-cases, and finally provide evidence to our main claim, that non-native speakers tend to choose more probable events. Due to lack of better knowledge, selecting the most probable translation is, statistically, a good bet.

References:

Baroni, M. (2009). Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), Corpus linguistics. An international handbook, 803–821. Berlin & New York: Mouton de Gruyter.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. Translation studies in Scandinavia, 88-95.

Granger, Sylviane. (2010). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. In: Journal of Shanghai Jiaotong University, Vol. 2, p. 14-21.

McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction. Edinburgh University Press.

Och, F. J., & Ney, H. (2000). Giza++: Training of statistical translation models.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In LREC (Vol. 14, pp. 1094-1101).

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016). On the Similarities Between Native, Non-native and Translated Texts. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1870–1881, Berlin, Germany, August 7-12.

Tiedemann, J., & Nygaard, L. (2004). The OPUS Corpus-Parallel and Free: http://logos. uio. no/opus. In LREC.

Toury, G. (1979). Interlanguage and its manifestations in translation. Meta: Journal des traducteursMeta:/Translators' Journal, 24(2), 223-231.

Toury, G. (1995). Descriptive translation studies and beyond. Amsterdam & Philadelphia: John Benjamins.

Tsui, A. B. (2004). What teachers have always wanted to know–and how corpora can help. How to use corpora in language teaching, 39-61.

Particle placement alternation in EFL learner speech vs. native and ESL spoken Englishes: core probabilistic grammar and/or L1-specific preferences?

Magali Paquot[1], Jason Grafmiller[2], Benedikt Szmrecsanyi[2]
FNRS / Université catholique de Louvain [1], KU Leuven[2]
magali.paquot@uclouvain.be, jason.grafmiller@gmail.com, Benedikt.Szmrecsanyi@kuleuven.be

Szmrecsanyi et al. (2016) explored three syntactic alternations (the particle placement, genitive and dative alternations) in four varieties of English (British, Canadian, Indian and Singapore English) as represented in the *International Corpus of English* and reported that the varieties studied share a core probabilistic grammar, i.e. the choice between syntactic variants is motivated by probabilistic constraints rather than categorical rules (cf. Bresnan, 2007). However, they also showed that grammatical variation is subject to indigenization "at various degrees of subtlety, depending on the abstractness and the lexical embedding of the syntactic pattern involved" (p. 2), with particle placement alternation exhibiting the most robust variety effects.

The main objective of the case study presented here is to shed some light on whether English as a Foreign Language (EFL) learners share a core probabilistic grammar with users of first and second language varieties of English. The study focuses on particle placement (as this alternation is more likely to exhibit variety effects, cf. Szmrecsanyi et al., 2016) and is driven by the following research questions:

1. What factors influence EFL learners' particle placement alternation?
2. How do EFL learners' particle placement preferences compare with those of users of first and second language varieties of English?

The study makes use of the French, German, Swedish and Dutch L1 components of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) (Gilquin et al., 2010) and largely replicates the methods used in Szmrecsanyi et al. (2016) to identify interchangeable transitive phrasal verbs with *around, away, back, down, in, off, out, over, on*, and *up*, and code particle placement variants in EFL learner speech. Unlike in Szmrecsanyi et al. (2016), however, identification and annotation of particle placement variants are done fully manually for two main reasons: (1) tagging learner speech as represented in the LINDSEI proves unreliable and (2) the LINDSEI components are much smaller (50 interviews each, between 75,000 and 95,000 words) than the corpora used in Szmrecsanyi et al. (2016).

Results are compared with corpus data from the « Exploring probabilistic grammar(s) in varieties of English around the world » research project which explored particle placement alternation in 9 varieties of English as represented in the *International Corpus of English* (ICE): British, Canadian, Hong-Kong, Indian, Irish, Jamaican, New Zealand, Philippine and Singapore English. For comparability purposes, results are also compared with data from the *Louvain Corpus of Native English Conversation*, i.e. a corpus of interviews with native speakers of English (LOCNEC; De Cock 2004).

Predictors included in the analysis are variety, nativeness, type of the direct object, length of the direct object in number of words and letters, animacy, definiteness, giveness and thematicity of the direct object, frequency of the direct object, the presence of a directional PP following the target VP and the semantics of the verb. Like in Szmrecsanyi et al. (2016), the effect of the different variables is investigated with conditional inference trees (mostly for visualization of interactions among predictors) and conditional random forest (to measure the overall importance of each predictor).

Preliminary results show that the type of the direct object (i.e. nominal head vs. pronominal head) and its length are the most important predictors of particle placement choice by EFL and L1 speakers (as represented in the LOCNEC corpus). EFL learners' particle placement preferences, however, differ from L1 speakers of English in two main ways: (1) there is a bias towards V-Part-DO in learner speech and (2) unlike LOCNEC speakers, EFL learners do not seem to be sensitive to (in)definiteness. Findings thus suggest so far that EFL and L1 speakers share a core (albeit simplified) probabilistic grammar (i.e. the main effect of the type of the direct object and its length is found in the 5 varieties investigated so far; the direction of the effects is stable across all varieties) but there are also clear EFL-specific preferences (see Wulff et al, 2014 for similar findings about that-variation). However, no significant differences between learner groups were noted despite the fact that the first languages represented in the learner dataset differ by the presence or absence of linguistic structures similar to English phrasal verbs.

We are now adding ICE data in our analyses so as to investigate how EFL learners' particle placement preferences compare with those of users of second language varieties of English. Based on results reported in Szmrecsanyi et al. (2016), we hypothesize that EFL learners' probabilistic grammar will resemble that of ESL speakers more than that of speakers of first language varieties.

References

Bresnan, J. (2007). Is syntactic knowledge probabilistic ? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (eds). *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton de Gruyter, 75-96.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. BELL : Belgian journal of English language and literatures, 225-246.

Gilquin, G., De Cock, S. & Granger, S. (2010). Louvain International Database of Spoken English Interlanguage (CD-Rom + handbook). Louvain-la-Neuve: Presses universitaires de Louvain.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller and Melanie Röthlisberger (2016). Around the world in three alternations: modeling syntactic variation in varieties of English. *English World-Wide*, 37(2) : 109-137.

Wulff, S., Lester, N. & Martinez-Garcia, M. (2014). That-variation in German and Spanish L2 English. *Language and Cognition* 6(2). 271–299.

Native Language Identification in a Portuguese learner corpus

Adriana Picoral, Jungyeul Park
University of Arizona
adrianaps@email.arizona.edu, jungyeul@email.arizona.edu

In this paper, we explore the classification accuracy of different feature sets for Natural Language Identification (NLI) in a 'cheap' learner corpus built from online journal entries written in Portuguese by learners who report speaking English and Spanish as their first languages. Falling into the broad task of authorship profiling, the topic of NLI has gathered some interest in the last decade or so, especially regarding English production by second language learners (Jarvis, Bestgen & Pepper, 2013; Kyle, Crossley & Kim, 2015; Koppel, Schler, & Zigdon, 2005; Tsur & Rappoport, 2007; Wong & Dras, 2011), with little attention given to other languages. The focus of this study is on Portuguese as an additional language (henceforth L2), and the detection of two native languages (L1s), Spanish and English. This combination of L1s and L2 was chosen due to the high relevance of these three languages not only to the authors' own context, but also to areas where they are often spoken and learned (e.g., the Iberian Peninsula). Also of interest is the typological proximity of the L1s and the L2. Thus, the main purpose of this study is to use text classification techniques for NLI in a newly built Portuguese learner corpus using the standard approach defined in previous literature, namely support vector machines (SVMs) (Brooke & Hirst, 2012).

There are few large learner corpora readily accessible in languages other than English, and Portuguese is no exception to this rule. In fact, there is only one learner corpus of Portuguese publicly available, a sub-corpus of The University of Toronto Romance Phonetics Database (RPD), which, due to its limited size and accessibility, does not really fit the requirements for NLI. Thus, following the work of Brooke and Hirst (2011), we built a 'cheap' Portuguese learner corpus from scraped online data by extracting 2,220 entries (a total of 256,794 tokens) from Lang-8, a website where language learners share their writing in an additional language so that native speakers can offer them feedback. We were able to collect at least 1,000 entries for each L1 group (i.e., Spanish and English). After normalizing punctuation and splitting sentences, the corpus was tokenized and POS tagged using two taggers: for universal tags we used the hunpos tagger (Halácsy, Kornai, & Oravecz, 2007) and for a more fine-grained tagging, we used Python's NLTK tagger trained with the floresta corpus (Freitas, Rocha, & Bick, 2008), a publicly available Treebank for Portuguese. For the SVM implementation, we used Python's scikit-learn, and our data were randomized and then divided into three parts: 80% train data, 10% development data, and 10% test data. Training was done using two kernels: RBF and linear. Different penalty parameters (C) were also tried out in development. We report results on the test data, and in our presentation we will also report the best kernel and penalty parameter combination found during development for each feature set. We also ran ten-fold cross-validation experiments, where the data were randomly partitioned into 10 equal-sized subsets.

The results point to a dominance of shallow lexical features, including token n-grams (classification accuracy of 83.67%), which reflects previous results for English learner corpora (Brooke & Hirst, 2012). In addition, character trigrams (76.77% accuracy), a feature that does not depend on specific language processing (e.g., POS tagging requires a tagged

corpus in the target language for training), performed really well in comparison to other more costly features, confirming the findings in (Ionescu, Popescu, & Cahill, 2016). Some feature sets that had not been used in previous studies, namely token-POS tuples extracted from hunpos statistical tagger, and idiosyncrasies based on tagging divergences and unknown tokens originated from a lemmatizer, performed quite well too (accuracy of 79.72% and 72.94%, respectively).

To the best of our knowledge, this is the first published study on NLI with data in Portuguese. Our research shows that NLI techniques previously used in learner corpora in English, Chinese (Malmasi & Dras, 2014b), Arabic (Ionescu et al., 2016; Malmasi & Dras, 2014a), and Norwegian (Ionescu et al., 2016), also work for romance languages such as Portuguese. Furthermore, the methods we used can be easily replicated by other researchers, in a multitude of L2s and L1s, including languages that are typologically close.

References:

Brooke, J., & Hirst, G. (2011). Native language detection with 'cheap'learner corpora. In Twenty years of learner corpus research. looking back, moving ahead: Proceedings of the first learner corpus research conference (lcr 2011) (Vol. 1, p. 37).

Brooke, J., & Hirst, G. (2012, 12). Robust, Lexicalized Native Language Identification. In Proceedings of coling 2012 (pp. 391–408). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from http://www.aclweb.org/anthology/C12-1025

Freitas, C., Rocha, P., & Bick, E. (2008). Floresta Sintá(c)tica: Bigger, Thicker and Easier. In A. Teixeira, V. L. S. d. Lima, L. C. d. Oliveira, & P. Quaresma (Eds.), Computational processing of the Portuguese language, 8th international conference (propor 2008) (pp. 216–219). Springer Verlag.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). Poster paper: HunPos – an open source trigram tagger. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions (pp. 209–212). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P07-2053

Ionescu, R. T., Popescu, M., & Cahill, A. (2016, 6). String Kernels for Native Language Identification: Insights from Behind the Curtains. Computational Linguistics, 42 (3), 491–525. Retrieved from http://dx.doi.org/10.1162/COLI{_}a{_}00256 doi: 10.1162/COLI{_}a{_}00256

Jarvis, S., Bestgen, Y., & Pepper, S. (2013). Maximizing classification accuracy in native language identification.

Kyle, K., Crossley, S. A., & Kim, Y. J. (2015). Native language identification and writing proficiency. International Journal of Learner Corpus Research, 1(2), 187-209.

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an Author's Native Language by Mining a Text for Errors. In Proceedings of the eleventh acm sigkdd international conference on knowledge discovery in data mining (pp. 624–628). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1081870.1081947 doi: 10.1145/1081870.1081947 Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP) (pp. 180–186). Doha, Qatar: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W14-3625

Malmasi, S., & Dras, M. (2014b). Chinese Native Language Identification. In Proceedings of the 14th conference of the European chapter of the association for computational linguistics, volume 2: Short papers (pp. 95–99). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/E14-4019

Tsur, O., & Rappoport, A. (2007, 6). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In Proceedings of the workshop on cognitive aspects of computational language acquisition (pp. 9–16). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W07/W07-0602

Wong, S. M. J., & Dras, M. (2011, 7). Exploiting Parse Structures for Native Language Identification.

In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 1600–1610). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D11-1148

Passives and Expletive Subjects in Learner English

Tom Rankin[1], **Elaine Lopez**[2]
Vienna University of Economics and Business[1], University of Newcastle[2]
tom.rankin@wu.ac.at, Elaine.Lopez@newcastle.ac.uk

There has been extensive interest in SLA research on the interplay between unaccusativity and word order phenomena (e.g. Rutherford 1989, Zobl 1989, Lozano & Mendikoetxea 2010), and the occurrence passive unaccusative errors (e.g. Oshita 2000, Hirakawa 2013) in L2 English. The basic insight from this work is that the underlying argument structure of predicates is a constraint on grammatical alternations and certain types of non-target production. Subject-verb inversion (VS) and overuse of passive morphology are restricted to unaccusative predicates in L2 English, see (1) and (2) respectively.

(1) It has occurred some important events (L1 Spanish, Lozano & Mendikoetxea 2013)

(2) Terrorism is happened very often (L1 Korean, Oshita 2000)

Lozano & Mendikoetxea (2010, 2013) have recently highlighted the occurrence of the generic expletive it in non-target unaccusative SV structures produced by L1 Spanish-speaking learners, see (1). In this paper, we propose that such it-insertion can be analysed as an instance of optional subject raising analogous to true subject-raising predicates, illustrated by the optionality in (3) and (4).

(3) It seems that important events have occurred.

(4) Important events seem to have occurred.

We explore this analysis by extending the empirical domain of previous studies. These studies extracted a range of unaccusative and unergative predicates from learner production and examined the linguistic constraints on the occurrence of inversion or overpassivisation structures, thus identifying differences between the syntactic behaviour of the different predicates. The present study explores expletive usage systematically by taking instances of expletive it and there as the starting point and examining their occurrence with unaccusative, passive and unergative predicates. The hypotheses are that learners will produce non-target VS with unaccusatives but not unergatives (providing support for previous work), but that this should also occur with passive predicates, which share the same underlying unaccusative argument structure with a single internal argument. In addition, if such structures are due to universal properties of argument structure, we expect similar non-target production from different L1 groups.

All occurrences of it and there were extracted from the Chinese, German and Spanish components of the International Corpus of Learner English (Granger et al 2003) and the Louvain Corpus of Native Essays (REF). Uses of referential it or expletives in clefts, extraposition, time expressions and weather predicates were removed. Occurrences of locational and presentational there were removed. The resulting database of sentences was made up of it/there + PASSIVE and it/there + (aux)+V. These were coded for the type of predicate (passive, unaccusative, unergative) and for the occurrence of subject NPs (SV or VS).

Target-like impersonal and expletive passives, involving presentational passives or verbs of reporting, thinking or saying plus clausal complements occur across the corpora. As predicted, non-target expletive passive VS errors also occur (see (5) and (6), but only in the Chinese and Spanish data, and markedly more frequently in the Spanish at a rate of .11 per 1000 words versus .04 in the Chinese.

(5) It can be avoided social problem (L1 Chinese)

(6) I think, it should be taught something more related with the life (L1 Spanish)

In addition, non-target structures involving post-verbal NP subjects with other raising predicates (see (7) occurred only in the Spanish data.

(7) It is necessary a lot of changes.

While non-target inversion with it is absent from the German data, L1 German learners are more likely to produce non-target inversion with expletive there, as in (8).

(8) Of course there have to be find solutions…

This pattern of results partially confirms the hypotheses to the extent that passives induce the expected patterns of non-target performance with it as identified previously for unaccusatives. We analyse this as a general learnability issue analogous to optional subject-raising constructions. However, the non-occurrence of such structures in the German data and the unique non-target production by Spanish-speakers requires an analysis in terms of the interaction of input with the L1 grammar. We propose that the occurrence of expletive constructions in German may facilitate acquisition of the English structure. We outline potential further research necessary to further investigate the acquisition of passive expletive structures and any constraints on their optionality in use, and to rule out potential proficiency effects between the different L1 groups.

References:

Lozano, C. and Mendikoetxea, A. (2010) Interface conditions on postverbal subjects: a corpus study of L2 English Bilingualism: Language and Cognition 13(4), 475-497.

Lozano, C. and Mendikoetxea, A. (2013) Corpus and experimental data: Subjects in second language research. In S. Granger, G. Gilquin & F. Meunier (eds) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use–Proceedings 1, Louvain-la-Nueve: Presses universitaires de Louvain, 313-323.

Oshita, H. (2000) What is happened may not be what appears to be happening: A corpus study of 'passive' unaccusatives in L2 English. Second Language Research, 16 (4), 293–324.

Oshita, H. (2004) Is there anything there when there is not there? Null expletives and second language data. Second Language Research, 20 (2), 95–130.

Rutherford, W. (1989) Interlanguage and pragmatic word order. In Gass S.M. and Schachter, J. (eds.) Linguistic perspectives on second language acquisition. Cambridge: Cambridge University Press, 163-82.

Zobl, H. (1989) Canonical typological structures and ergativity in English L2 acquisition. In Gass S. M. and Schachter, J. (eds.) Linguistic perspectives on second language acquisition. Cambridge: Cambridge University Press, 203-21.

Thematic structure in English and Norwegian academic texts in the field of didactics: novice writers vs. expert writers

Sylvi Rørvik, Marte Monsen
Inland Norway University of Applied Sciences
sylvi.rorvik@inn.no, marte.monsen@inn.no

At Inland Norway University of Applied Sciences, there is a master's program in the teaching and learning of language subjects during which the students, specializing in either English or Norwegian, have to write a report on a small empirical study of a topic related to didactics. Hence, the texts they produce are a ready-made comparable corpus of novice academic writing. The present study aims to investigate the extent to which novice academic writers of L2 English and L1 Norwegian are able to conform to the discourse conventions of expert academic texts regarding thematic structure within the field of didactics. It was decided to focus on thematic structure (cf. Halliday 2004: 64-105) since previous research has shown that this is an area where novice writers may struggle (cf. e.g. Berry 1995; Hawes & Thomas 1997; Hasselgård 2009; and Rørvik 2013), and it is also an area where disciplinary differences have been identified (cf. e.g. North 2005a & b). In addition, there is a relative dearth of studies of student academic writing in Norway, at least as regards studies including a contrastive perspective (but see e.g. Fossan 2011) and yet previous contrastive studies of other text types have identified differences in thematic structure between English and Norwegian (Hasselgård 1998, 2005).
In an attempt to fill this gap, a contrastive study was carried out of the above-mentioned student texts, alongside a comparison of texts by expert writers of English and Norwegian (i.e. published academics), in a procedure roughly following the Integrated Contrastive Model (Gilquin 2000/2001: 100-101). The novice material was taken from a corpus of student academic writing that is in the process of being compiled at Inland Norway University of Applied Sciences, and the comparable expert texts were collected from journals publishing papers within the field of didactics (*Acta Didactica Norge* for the English expert material and four different journals within didactics for the Norwegian expert material). Since the students write about education in a Norwegian context, it was decided to use expert texts that had a similar focus. Altogether, the material comprises 11 texts in each of the four categories, and these were divided into T-units (cf. Fries 1995: 318) and manually analyzed for features related to thematic structure. Statistical calculations were then carried out to compare the results for each corpus, by means of a one-way ANOVA with a Tukey post-hoc test. The table below provides an overview of the size of the material terms of the total number of words and total number of T-units in each subcorpus.

	Number of texts	Number of words	Number of T-units
Norwegian experts	11	56,161	2,732
Norwegian novices	11	42,966	2,311
English experts	11	75,529	3,130
English novices	11	41,132	1,813

The results show that the expert writers of Norwegian display a lower proportion of unmarked themes (i.e. subjects as themes) than the English writers (p=0.0080607), although it should be noted that there is a greater degree of variation within the Norwegian expert subcorpus than is found within the corresponding English subcorpus. This contrastive difference does not seem to cause problems for the novice writers, however, since neither of the novice subcorpora are different from the expert subcorpora in their respective languages when it comes to the proportion of unmarked themes.

There are no significant differences in the proportion or realization of marked themes (i.e. non-subjects as themes) between the Norwegian and English expert texts, nor when it comes to the distribution of meanings expressed by the marked themes. However, there are several areas where the two groups of novice writers differ from each other, for instance as regards the types of constructions they employ as marked themes: dependent clauses are more frequent in English than in Norwegian (p=0.0411287), while the opposite is true for prepositional phrases (p=0.0242974). Given that the expert texts do not exhibit the same differences, we conclude that the novice writers need advice in these areas in order to conform to the conventions of the text type and field.

References:

Berry, M. (1995). Thematic options and success in writing. In M. Ghadessy (Ed.). *Thematic Development in English Texts*. London & New York: Pinter, 55-84.

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.* Master's thesis, University of Oslo.

Fries, P. H. (1995). Themes, development and texts. In R. Hasan and P. Fries (Eds.). *On Subject and Theme.* Amsterdam: John Benjamins, 317-359.

Gilquin, Gaëtanelle. (2000/2001). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3(1), 95-123.

Halliday, M. A. K. (2004). *An Introduction to Functional Grammar.* 3rd edition, revised by C. M. I. M. Matthiessen. London: Arnold.

Hasselgård, H. (1998). Thematic structure in translation between English and Norwegian. In S. Johansson & S. Oksefjell (Eds.). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies.* Amsterdam: Rodopi, 145-167.

Hasselgård, H. (2005). Theme in Norwegian. In K.L. Berge & E. Maagerø (Eds.). *Semiotics from the North. Nordic approaches to systemic functional linguistics.* Oslo: Novus Press, 35-47.

Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.). *Corpora and Language Teaching*. Amsterdam: John Benjamins, 121-139.

Hawes, T. & Thomas, S. (1997). Problems of Thematisation in Student Writing. *RELC Journal* 28, 35-55.

North, S. (2005a). Disciplinary Variation in the Use of Theme in Undergraduate Essays. *Applied Linguistics* 26(3), 431-452.

North, S. (2005b). Different values, different skills? A comparison of essay writing by students from arts and science backgrounds. *Studies in Higher Education* 30:5, 517-533.

Rørvik, S. (2013). *Texture in learner language.* Doctoral dissertation, University of Oslo.

Intertextuality in pedagogic genres: Examining the influence of genre- and task-based factors on source-based business writing

Christine Sing
Vienna University of Business and Economics / University of Vienna
csing@wu.ac.at

The concept of task has been shown to be vital for L2 teaching and learning (e.g., Ellis 2003). Research into task-based language teaching has only recently shifted its focus from spoken to written communication (Ortega 2012: 405). Little is thus known about how task relates to writing, particularly in pedagogic genres such as L2 assignment writing. In these 'occluded genres', the "authors cannot fashion their discourse on prototypical texts; rather, they must rely upon other genres to structure their texts" (Conner Loudermilk 2007: 202). Focussing on the immediate pedagogical context in which the L2 writing task is situated, it will be argued that these students tend to rely on the expert models they find in the source literature.

Intertextuality is therefore integral to source-based academic writing and covers a great variety of language phenomena, including direct quotations, copy-and-paste jobs or paraphrasing (e.g., Petrić 2012, Pecorari & Shaw 2012, Davis & Morley 2015). There is considerable evidence that L2 writers struggle with source appropriation, i.e., an intertextual strategy of effectively using and restructuring source texts, frequently giving rise to 'transgressive intertextuality' (Abasi & Akbari 2008).

For the purpose of analysis, it is however important to differentiate between language re-use and source use. Writers re-use language for different reasons; they may have pooled commonly used language resources for future (re-)use; they may imitate recently encountered linguistic options; they may have been otherwise prompted or primed (Hoey 2005). Source use, on the other hand, is inextricably tied up in the nexus of reading and writing. Effective source use depends on successful sense-making strategies and good comprehension skills.

There are several aims of this study: 1. To ascertain the intertextual strategies used to carry out the writing task. 2. To investigate to what extent these practices are indicative of language prompts that have permeated into the students' writing. 3. To determine the textual functions which are susceptible to intertextuality. 4. To develop a model of language re-uses in terms of the task-based factors accounted for in the corpus data. In order to examine the influence of genre- and task-based factors on this particular pedagogic genre, the research corpus will be compared to a reference corpus consisting of all sources used in the writing assignment.

The ESP setting of a business school, in which the writing task originated, showcases a pedagogic context typical of occluded genres. The database of this study is made up of a self-compiled specialised corpus, the corpus of Academic Business English (ABE), which consists of c. 1 million running words. Its compilation was guided by a clear set of design criteria, drawing on Flowerdew's (2004: 21) parameters for specialized corpora and

Tribble's (2002: 133) contextual-analysis framework. The ABE corpus contains more than 400 papers produced by advanced students of international business administration. Drawing on this rich source of data, the present study combines bottom-up, inductive, corpus analyses with top-down analyses focussing on larger portions of discourse (Flowerdew 2005).

The findings show that intertextual practices, while pervasive throughout the corpus texts, have limited range and tend to cluster in specific sections of the papers, thus pointing to both localized and global intertextual practices. It would thus seem that task-based factors strongly influence the writing, causing an effect of 'persistence' (Szmrecsanyi 2005), i.e. the idea that language users will rely on recently encountered language patterns whenever possible. Another important, interrelated, finding is that the textual sources used by the students in text production tend to be 'language re-uses' (Flowerdew & Li 2007).

These results, while preliminary, suggest that the concepts of task and context are pivotal to ESP writing. This calls for extending the concept of task to include "the psycholinguistic and textual nature of writing tasks in terms of a focus on the linguistic resources for meaning-making" (Byrnes & Manchón 2014: 7). Some of the issues emerging from these findings relate specifically to ESP writing instruction, suggesting a strong influence of two interrelated factors, one task-based and the other teaching-induced. The heavy reliance on expert uses modelled on the source literature provides strong evidence in favour of learning by imitation (Limburg 2014). This study should, therefore, be of value to practitioners wishing to blend task-based and genre-based writing instruction.

References:

Abasi, A. R., & Akbari, N. (2008). Are we encouraging patchwriting? Reconsidering the role of the pedagogical context in ESL student writers' transgressive intertextuality. *English for Specific Purposes, 27*(3), 267–284.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (Eds.). (2012). *Corpus-informed research and learning in ESP: Issues and applications*. Amsterdam: Benjamins.

Byrnes, H., & Manchón, R. (Eds.). (2014). *Task-based language learning: Insights from and for L2 writing*. Amsterdam: Benjamins.

Conner Loudermilk, B. (2007). Occluded academic genres: An analysis of the MBA Thought Essay. *Journal of English for Academic Purposes, 6*(3), 190–205.

Davis, M., & Morley, J. (2015). Phrasal intertextuality: The responses of academics from different disciplines to students' re-use of phrases. *Journal of Second Language Writing, 28,* 20–35.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2004). Supporting genre-based literacy pedagogy with technology - the implications for the framing and classification of the pedagogy. In L. J. Ravelli & R. Ellis (Eds.), *Analysing academic writing. Contextualized frameworks*. London: Continuum, 210–232.

Evans, S. (2013). Designing tasks for the Business English classroom. *ELT Journal, 67*(3), 281–293.

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics, 28,* 440–465.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions. Perspectives from corpus linguistics*. Amsterdam: Benjamins, 11–37.
Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*(3), 321–332.
Flowerdew, L. (2011). *Corpora and Language Education*: London: Palgrave Macmillan.
Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam: Benjamins.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
Limburg, A. (2014). Imitationslernen in den Wirtschaftswissenschaften: Unterschiede zwischen Examensarbeiten und Forschungsartikeln. *Zeitschrift Schreiben. Schreiben in Schule, Hochschule und Beruf* (http://www.zeitschrift-schreiben.eu).
Ortega, L. (2012). Epilogue: Exploring L2 writing–SLA interfaces. *Journal of Second Language Writing, 21*(4), 404–415
Pecorari, D. (2015). Plagiarism in second language writing: Is it time to close the case? *Journal of Second Language Writing, 30,* 94–99.
Pecorari, D., & Shaw, P. (2012). Types of student intertextuality and faculty attitudes. *Journal of Second Language Writing, 21*(2), 149–164.
Petrić, B. (2012). Legitimate textual borrowing: Direct quotation in L2 student writing. *Journal of Second Language Writing, 21*(2), 102–117.
Seidlhofer, B. (2000). Operationalizing intertextuality: using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective. Papers from the Third International Conference on Teaching and Language Corpora*. Wien: Lang, 207–223.
Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory, 1*(1), 113–149.
Tardy, C. M. (2012). Writing and Language for Specific Purposes. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell, 6266–6274.
Tribble, C. (2002). Corpora and corpus analysis: new windows on academic writing. In J. Flowerdew (Ed.), *Academic discourse*. Harlow: Longman, 131–149.

First Language Proficiency Predicts Second Language Proficiency: An Investigation of Linguistic Complexity in L1 and L2 Academic Writing

Marcus Stroebel[1], Elma Kerz[1], Daniel Wiechmann[2]
RWTH Aachen University[1], University of Amsterdam[2]
marcus.stroebel@rwth-aachen.de, elma.kerz@ifaar.rwth-aachen.de, d.wiechmann@uva.nl

Recent years have witnessed an increasing interest in usage-based/experience-driven and emergentist models of first and second language acquisition (cf., e.g. Ellis & Larsen-Freeman, 2006; Beckner et al. 2009; McClelland, et al. 2010; Larsen-Freeman, 2011; MacWhinney, 2012; Ambridge & Lieven, 2015). In these models, language learning (both first and second) is a continuous process, which does not end at some discrete point of time in ontogenetic development but instead takes place across the lifespan. Correspondingly, language learning does not result in the establishment of a static knowledge system; rather, as long as there is exposure to linguistic input, an individual's knowledge of a language is in constant flux. These models render the notion of 'ultimate attainment' superfluous. It has been often assumed that L2 learners have a perfect command of their L1 (cf. Hulstijn, 2015 for a discussion). Hence, with regard to L1 performance, L2 learners are treated as members of a homogeneous group of individuals who can only differ in their L2-specific proficiency. However, more recently, there has been an explosion of studies uncovering substantial individual differences across multiple components of language across the lifespan in native speakers (cf., e.g., Dabrowska & Street, 2006; Dabrowska, 2012). The variability in both non-native and native language proficiency raises the question regarding the relationship between L1 and L2 proficiency. This relationship has typically been investigated in controlled experimental settings with a primary focus on receptive skills. Learner Corpus Research (LCR) has a unique contribution to make to understanding this relationship by targeting productive skills and relying on naturalistic, ecologically valid data.

The present study showcases how learner corpus data can be used to investigate whether individual variation in L1 proficiency can explain variation in L2 proficiency. We used the Aachen Corpus of Academic Writing (ACAW, Kerz & Stroebel, 2017). ACAW is a multilingual learner corpus of intermediate to advanced L1 German and L2 English academic writing. In its current form, ACAW consists of 80 pairs of L2 - L1 texts (mean $length_{L2}$ = 5,084 words, SD = 2,019; mean $length_{L1}$ = 4,650, SD = 1,695) produced by the same set of undergraduate and graduate university students (mean age = 23.92, SD = 2.61) at the same stage of development (see, Kerz & Stroebel, 2017, for details). Prior to the assessment of complexity, all texts were analyzed with several annotators from Stanford CoreNLP (Manning et al. 2014): tokenizer, sentence splitter, POS tagger, lemmatizer, named entity recognizer and syntactic (PCFG) parser. Thirteen lexical and grammatical complexity measures (CMs) were included into the analysis. These measures have been seen as basic

descriptors of L2 performance and as indicators of L2 proficiency (cf. Wolfe-Quintero, Inagaki, & Kim, 1998; Housen, Kuiken, & Vedder, 2012).
All measurements of complexity were automatically obtained using CoCoGen (short for Complexity Contour Generator; cf. Ströbel, 2014; Ströbel, Kerz, Wiechmann & Neumann, 2016). Rather than providing a single complexity score per text, CoCoGen uses a sliding-window technique to assess the complexity of a text based on a series of measurements. The large number of words of the ACAW texts allowed us to gather 10 measurements from each text by assessing text-complexity for 10 equally sized text-partitions (or windows). This corresponded to one measurement of complexity every 400-500 words of a given text. We thus obtained 20 data points per learner (10 measurements per text * 2 texts), which allowed us to investigate whether the effect of L1 proficiency differs reliably between subjects.
The role of L1 complexity on L2 proficiency was evaluated using mixed-effect linear regression models implemented with the lme4 package (Bates, Maechler, & Bolker, 2012) in the R environment (R Development Core Team, 2014). Separate models were fitted for each of the 13 CMs. In each model, the outcome variable L2 complexity was regressed onto the predictor variable L1 complexity. In addition, three experience-related control variables were entered into the model as fixed effects (participant age, number of months spent in an English speaking country, and years of formal English education). All models included (correlated) by-subject random intercepts and slopes for L1 proficiency as well as random intercepts per text-partition (window). For the assessment of the significance of L1 complexity, we examined its fixed effect in the presence of the corresponding random slopes (Pinheiro & Bates, 2000).
We found L1 complexity to be a significant predictor for 12 out of the 13 measures (all with the exception Mean Length of Words), after controlling for the effects of the experience-related control variables. The implications of our findings for the assessment of proficiency in a second language as well as for dynamic perspectives of L2 development are discussed.

References:

Ambridge, B., & Lieven, E. (2015). 22 A Constructivist Account of Child Language Acquisition. The handbook of language emergence, 87, 478.

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. 2012. URL: http://CRAN. R-project. org/package= lme4.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D. and Schoenemann, T. (2009), Language Is a Complex Adaptive System: Position Paper. Language Learning, 59: 1–26.

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. Linguistic Approaches to Bilingualism, 2(3), 219-253.

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. Language Sciences, 28(6), 604-615.

Ehret, K. & Szmrecsanyi, B. (2011). An information-theoretic approach to assess linguistic complexity. Complexity and isolation. Berlin: de Gruyter.

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. Applied linguistics, 27(4), 558-589.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (Vol. 32). John Benjamins Publishing.

Hulstijn, J.H. (2015). Language proficiency in native and non-native speakers: Theory and research. Amsterdam: John Benjamins.

Kerz, E. & Stroebel, M. (2017) Aachen Corpus of Academic Writing (ACAW): A Multilingual Corpus of First and Second Language Writing. Paper presented at LCR 2017, Bolzano, Oct. 5-7.

Larsen-Freeman, D. (2011). A complexity theory approach to second language development/acquisition. Alternative approaches to second language acquisition, 4872.

MacWhinney, B. (2012). The logic of the Unified Model. The Routledge handbook of second language acquisition, 211-227.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. & McClosky, D. (2014) The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. Trends in cognitive sciences, 14(8), 348-356.

Pinheiro, J. C. and Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. Statistics and Computing. Springer, New York.

Stroebel, M. (2014). Tracking complexity of l2 academic texts: A sliding-window approach. Unpublishe master's thesis. RWTH Aachen University, Aachen, Germany.

Stroebel, M., Kerz, E., Wiechmann, D., & Neumann, S. (2016). CoCoGen-Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-Window Technique. CL4LC 2016, 23.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity (No. 17). University of Hawaii Press.

What Kind of Linguistic Features Distinguish Second Language Learners' Texts from Those of Native Speakers, and Why?

Masatoshi Sugiura, Daisuke Abe, Yoshito Nishimura
Nagoya University
sugiura@nagoya-u.jp, abe.gsid@gmail.com, nishimura@nagoya-u.jp

This study attempts to find the differences between the first language (L1) and the second language (L2) by investigating linguistic features that distinguish L1 and L2 texts. The preliminary goal is to determine critical linguistic features that can be automatically computed and used to distinguish L1 and L2 texts irrespective of text length. This analysis would provide insight into the differences between L1 and L2.

Crossley and McNamara (2009) attempted native/non-native classification by using lexical features computed by Coh-Metrix. They reported a classification accuracy of 79% using 10 features. As mentioned in their study, this method was limited in that it did not take into account variables other than lexical features.

The current study attempted to expand on this approach by selecting from lexical and syntactic features computed by Lu's L2 Syntactic Complexity Analyzer (SCA) (2010) and Lexical Complexity Analyzer (LCA) (2012). The SCA includes 23 measures such as frequencies of words, sentences, and T-units and their ratios measured against other frequencies. The LCA outputs 34 measures including types and tokens of certain word categories and other indices computed from these numbers. Besides the measures obtained from these two analyzers, another measure, D, was added (Malvern et al., 2004; MacWhinney, 2000). After removing the overlapping measures in the SCA and the LCA, 56 measures were left.

Data from English essays written by participants were used. The original data consisted of 36 L1 texts (*M*: 989.8; *SD*: 327.8) and 185 L2 texts (*M*: 312.6; *SD*: 102.1). In order to balance the text lengths between the L1 and L2 data for comparison, the 36 longest texts were selected from the L2 texts. The 36 L1 texts and 36 L2 texts were then truncated to around 383 words each, which was the length of the shortest L2 essay.

To classify the L1 and L2 texts, random forest classification was conducted using the 56 measures. When all of the measures were considered, the error rate was quite low (6.94%). Many of the measures, however, are highly correlated, and some of them are products or quotients of other measures. To narrow down the measures to the most significant ones, the mean decrease in Gini coefficient was used as a guideline for feature importance. Random forest categorization was conducted with several different combinations of measures, removing overlapping measures and other measures that were likely to be functions of text length.

The remaining two measures, Mean Length of T-unit (MLT) and D, yielded a reasonably low error rate of around 10%. Both measures are relatively unaffected by differences in text length, although McCarthy and Jarvis (2010) report a significant, albeit small, correlation between D and text length.

To provide further evidence that these two measures alone can be used to correctly classify the texts, discriminant analyses were conducted. Subgroups were created from the L2 texts by randomly selecting five subgroups of 36 texts without overlap from the 185 texts. These five subgroups were paired with the 36 L1 texts. Because homogeneity of variance could not be assumed with these two measures, text classification was demonstrated using quadratic discriminant analyses based on Mahalanobis generalized distance, which yielded error rates with an average of 5.6%. Five hyperbolic curves were obtained discriminating the two groups. In addition, a "general" discriminant function was derived, using the 36 L1 texts as one group and all five sets of the L2 texts as the other group:

$$f(x, y) = .0003x^2 + .0029xy - .1228y^2 - .2268x + 1.4742y + 19.9610,$$

where *x* represents D and *y* represents MLT. Using the values of D and MLT of any particular text, it should be possible to determine the nativeness of the writer, a negative value indicating nativeness.

From these results, MLT and D may be suggested as measures used to distinguish between L1 and L2 texts. These measures are versatile because they can be used regardless of text length.

D has been used as a major index of lexical diversity along with MTLD (McCarthy and Jarvis, 2010). As for MLT, however, the significant role of such a simple index is a mystery. This issue may be further discussed from the point of view of the "Shallow-Structure Hypothesis" (Clahsen and Felser, 2006) and the "complexity devices" of "non-clausal features embedded in noun phrases" (Biber, Gray and Poonpon, 2011).

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,* 1-31.

Clahsen, H. & Fesler, C. (2006). Grammatical processing in language learners, *Applied Psycholinguistics, 27,* 3-42.

Crossley, S. A. & McNamara, D. S. (2009). Computational assessment of lexical differences in second language writing. *Journal of Second Language Writing*, *17*(2), 119-135.

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.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Hampshire: Palgrave Macmillan. Doi: 10.1057/9780230511804

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392.

Tracking the long-term evolution of foreign language proficiency through development and analysis of a bilingual, multimodal and longitudinal corpus

Nicole Tracy-Ventura[1], Amanda Huensch[1], Rosamond Mitchell[2]
University of South Florida[1], University of Southampton[2]
nkt@usf.edu, huensch@usf.edu, R.F.Mitchell@soton.ac.uk

Much previous learner corpus research has focused on investigating aspects of second language (L2) writing. Fewer studies have investigated spoken learner corpora, and even fewer have examined spoken and written data from the same learners over time. Yet longitudinal corpora are important because they can potentially yield the most valuable insights into L2 development (Ortega & Byrnes, 2009; Ortega & Iberri-Shea, 2005). By tracking learners longitudinally using corpus-based methods to collect multimodal data (both L2 speech and L2 writing), it becomes possible to compare their linguistic development across modes as well as across time. Longitudinal corpora are rare, in large part because of the cost and effort required for data collection, storage, and curation. However, learner-corpus tools and methods provide an efficient means by which longitudinal L2 data can be analyzed, stored, and shared with other research teams, and indeed a few researchers have begun to analyze written longitudinal learner corpora (Meunier & Littre, 2013; Vyatkina, 2013; Yuldashev, Fernandez, & Thorne, 2013). Yet to our knowledge no longitudinal learner corpus exists that includes both oral and written language from the same participants.

This presentation will describe the longitudinal and multimodal learner corpora collected for the Languages and Social Networks abroad Project (LANGSNAP) which began in May 2011, in order to investigate the long-term evolution of L2 proficiency (development, maintenance, attrition, and the relationship between development in L2 speech and writing). To date, there have been 7 data collection waves, the most recent in June 2016. Data come from English L1 learners of French (n=29) and Spanish (n=27) who, when the project began, were university students in the UK majoring in French and/ or Spanish. As part of their degree requirements they had to spend the 2011-2012 academic year abroad in France, Mexico, or Spain. The main purpose of the initial study was to investigate the influence of social, individual, and contextual factors on language learning during residence/study abroad. Because the participants were already of intermediate proficiency, we expected they would be developing a sense of style/genre, and therefore chose to include three communicative activities to examine how their linguistic abilities changed over time and across genres: 1) an oral interview, 2) an oral narrative, and 3) argumentative writing. These activities were completed on 6 occasions over 21 months during the later part of the participants' undergraduate programme: presojourn x 1, insojourn x 3, and postsojourn x 2. All data were transcribed according to CHAT conventions (MacWhinney, 2000) and later analysed using CLAN for fluency, accuracy, and lexical and syntactic complexity. Those results are described in Mitchell, Tracy-Ventura, and McManus (2017).

A follow-up study was launched in 2015 to add new data from a subset of the same participants (n=33) to the existing longitudinal corpus. Because the participants had since graduated from university and many were living and working in a mainly English-speaking environment, we anticipated that some participants would be experiencing foreign language attrition. However, others had spent further time abroad and/or were employed in L2-using professions (e.g., school teaching), therefore likely following a rather different developmental trajectory. The longitudinal nature of project, the inclusion of two different target L2s, the analysis of both speech and writing, and the systematic collection of background data on changing patterns of social networking and L2 use, provide a powerful means to investigate the long-term evolution of foreign language proficiency.
In addition to the learner corpus, data from this project include a proficiency test and a test of lexical knowledge. In this methodologically-oriented presentation we will discuss how we are utilizing those data in addition to questionnaire data about L2 use in the follow-up project. The design, methods, and results of our current project provide implications for future research in LCR and second language acquisition (SLA). In particular they highlight how SLA-informed questions can be investigated using learner-corpus tools, and by using learner-corpus tools and making our corpus freely available, we are sharing a valuable resource with other researchers interested in similar questions about the long-term evolution of a foreign language.

References:

MacWhinney, B. (2000). *The Childes Project: Tools for Analyzing Talk. Transcription format and programs*. Psychology Press.

Meunier, F., & Littre, D. (2013). Tracking Learners' Progress: Adopting a Dual "Corpus cum Experimental Data" Approach. *The Modern Language Journal*, *97*(S1), 61–76. https://doi.org/10.1111/j.1540-4781.2012.01424.x

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone Students Abroad: Identity, social relationships and language learning.* New York: Routledge.

Ortega, L., & Byrnes, H. (2009). *The Longitudinal Study of Advanced L2 Capacities*. New York: Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45. https://doi.org/10.1017/S0267190505000024

Vyatkina, N. (2013). Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus. *The Modern Language Journal*, *97*(S1), 11–30. https://doi.org/10.1111/j.1540-4781.2012.01421.x

Yuldashev, A., Fernandez, J., & Thorne, S. L. (2013). Second Language Learners' Contiguous and Discontiguous Multi-Word Unit Use Over Time. *The Modern Language Journal*, *97*(S1), 31–45. https://doi.org/10.1111/j.1540-4781.2012.01420.x

A Corpus-based Approach to the Use and Acquisition of Prepositions by Learners of German as a Foreign Language: On the Effect of Specification

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

In German SLA studies, the effect of distinct syntactic functions of prepositional phrases (PPs) on the use and acquisition of prepositions has been rather neglected so far. Taking up the corpus-based study of learner language outlined in Weber (2014, 2015), this paper presents a pilot learner corpus study focusing on distinct syntactic functions of PPs with different specifications of prepositions: (a) complements containing specified prepositions licensed by verbs and adjectives and (b) adjuncts (and adjunct-like complements) containing unspecified prepositions (Huddleston & Pullum, 2006: p. 215, p. 272f.)[1]:
(a)
I. Ich warte auf Ihre Antwort (I'm waiting for your response)
II. Ich bin gespannt auf deine Antwort (I'm curious about your response)

(b)
I. … einfach auf der Straße spazieren (simply strolling on the street)
II. … auf dem Land zu wohnen (to live in the country)

The research questions of the pilot study are:

1. How do accuracy rates for prepositional usage differ across the distinct functions of PPs?
2. (How) Do error types differ in the distinct functions of PPs?
3. What error types are frequent?

The pilot study uses the German MERLIN corpus (Abel et al., 2014)[2]. The German subcorpus contains 1033 texts from high-quality language tests written by GFL learners and rated according to the competence levels of the CEFR. MERLIN includes multi-layer error annotations and target hypotheses (Lüdeling, 2008). For the pilot study, instances of the German local prepositions *an/am* (*at*) and *auf* (*on*) were extracted from the corpus and the PPs annotated by two annotators according to their syntactic functions (see above)[3]. The prepositions were chosen based on their high frequency rates in the lexical database

[1] PPs as modifiers are not in focus here. Examples are taken from the corpus, the English translation is given in the columns. Formulaic sequences were excluded from the analysis.
[2] The corpus is accessible via http://www.merlin-platform.eu/
[3] κ_{mean}=0.9 (IAA for 100 double annotated PPs)

DLEX[1]. In all, 1009 annotated PPs[2] were analyzed using target-like use analysis (Ellis/Barkhuizen, 2005, p. 74), computer-aided error analysis[3] (Dagneaux et al., 1998) and contrastive interlanguage analysis (Granger, 2015). Overall results show that the variable syntactic function affects accuracy and error type in the use of prepositions: 1) complements with specified prepositions (a) display a significantly higher error rate and thus pose greater problems for GFL learners than adjuncts (and adjunct-like complements) containing unspecified prepositions (b) and 2) certain error types[4] strongly correlate with syntactic function:

	(a)	(b)	Chi-Square-Test
Error rate (%)	28,1	19,7	p = 0,002
Omitted preposition (%)	50,8	22,9	p = 1,973e-05
Incorrect preposition (%)	32,8	59,6	p = 6.222e-05

The research questions can be answered as follows:
1. Complement-PPs show a lower accuracy rate regarding the use of prepositions than adjunct (and complement-like adjunct) PPs.
2./3. Dominant error types (omission and choice) differ according to and correlate with syntactic function of PPs.
PPs as complements contain specified prepositions whose semantic meaning is more abstract than e.g. those of unspecified prepositions in prototypical adjunct PPs (Duden, 2016, p. 825; Eisenberg, 2013, p. 183, see also Huddleston/Pullum, 2006, p. 647ff.) and this leads to greater uncertainty concerning the realization of prepositions in prepositional complements whereas the uncertainties in adjunct-PPs, not surprisingly, rather concern the correct choice of preposition. This tendency is in accordance with usage-based accounts in SLA, which stress the importance of (semantic) salience in language acquisition (Ellis, 2002, p. 175, see also Ellis, 2006). Extended results of the pilot study concerning single competence levels of the CEFR will be presented and discussed in the talk.
The impact of semantic salience of prepositions and its effect on prepositional use will be further explored in a follow-up corpus study targeting complement-PPs with less abstract prepositions (see Breindl 2006: 946). Results will be compared to those of the pilot study presented here. Further learner corpus studies on prepositions embarking on additional variables prominent in SLA research (and in particular within cognitive accounts of SLA) are planned.

References:

[1] The database is accessible via http://www.dlexdb.de/
[2] The 1009 PPs are distributed across 627 learner texts.
[3] For error tags and the error coding procedure see MERLIN project (2014) and Wisniewski et al. (2014).
[4] The table displays the two dominant error types *omission* and *choice*.

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,* 1(2), 111–126. Retrieved from http://www.ojs.unito.it/index.php/ricognizioni/article/view/702/677

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

Duden (2016). Die Grammatik 9., vollständig überarbeitete und aktualisierte Auflage. Edited by Wissenschaftlicher Rat der Dudenredaktion. Berlin: Dudenverlag.

dlexDB: Lexikalische Datenbank (University of Potsdam, Berlin-Brandenburg Academy of Sciences and Humanities). http://www.dlexdb.de/.

Eisenberg, P. (2013). *Grundriss der deutschen Grammatik*. Band 2: Der Satz. 4. Auflage. Stuttgart: J.B. Metzler.

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*, 4, 143-188.

Ellis, R. & Barkuhuizen, L. (2005). *Analysing Learner Language*. Oxford: Oxford UP.

Ellis, N. C. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics*, 27(2), 164-194.

Gibson, M., Hufeisen, B. & Libben, G. (2001). Learners of German as an L3 and their Production of German Prepositional Verbs. In J. Cenoz, B. Hufeisen & U. Jessner (Eds.), *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives* (pp. 138-148). Clevedon: Multilingual Matters.

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

Huddleston, R. & Pullum, G. K. (et al.) (2006). *The Cambridge Grammar of the English Language* (4th ed.). Cambridge: Cambridge: Cambridge UP.

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

MERLIN: MERLIN – Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context (Technische Universität, Dresden). www.merlin-platform.eu.

MERLIN project (2014). Annotation guidelines. Retrieved from www.merlin-platform.eu. (26.09.2016).

Wisniewski, K., Woldt, C., Schöne, K., Abel, A., Blaschitz, V., Štindlová, B. & Vodičková, K. (2014). The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language . Retrieved from www.merlin.platform.eu.

Weber, T. (2014). Verbvalenz und Rektion im Bereich Deutsch als Fremdsprache. Eine korpusgestützte Analyse zweier Verbgruppen (Master's Thesis TU Dortmund University). Retrieved from http://merlin-platform.eu/docs/Masterarbeit_Tassja_Weber.pdf

Weber, T. (2015). Verb Valency and Prepositional Complements in Learner Corpora: A Case Study in the German MERLIN Corpus. In P. de Haan (Ed.), LCR 2015 Book of Abstracts (pp. 164–166). Retrieved from http://www.ru.nl/publish/pages/765127/definitive_book_of_abstracts.pdf

Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment

Zarah Weiß, Detmar Meurers
University of Tübingen
zweiss@sfs.uni-tuebingen.de, dm@sfs.uni-tuebingen.de

This study investigates the applicability of diverse language features to German L2 proficiency assessment. In recent years, an abundance of features has been proposed to measure language proficiency, readability, and writing skills (cf. references). They differ in terms of language domain, specificity, and extraction complexity. Yet, it remains unclear to which extent the broad combination of features and the use of complex features are beneficial in predictive approaches, especially as more complex feature extraction procedures are more prone to errors when applied to non-standard data. We address this issue by i) comparing classification models with diverse and homogeneous feature sets, and ii) comparing feature performance on raw and normalized L2 data.

We extract data from 1,033 CEFR rated German L2 texts and their normalizations from the Merlin corpus. Unfortunately, Merlin is not balanced for CEFR scores: Learners at A1 and C1 account for only 10% of the data, while the other 90% are approximately evenly distributed among levels A2 to B2. Thus, we sampled 255 texts uniformly representing the five proficiency classes and worked with this data set throughout.

We analyze 398 language features chosen and implemented based on work from a broad range of perspectives (cf. references). They measure elaborateness and variability of measures from theoretical linguistics, language use, human language processing, and discourse and meaning. Feature extraction methods range from POS tagging to combinations of parsers. To our knowledge, this currently is the most extensive language complexity feature set for German.

We classify proficiency using the SMO algorithm in the Weka toolkit and rank features by information gain. We include test level and task id to account for task effects. For training and testing, we use 10-folds cross validation.

In the first experiment, we predict CEFR scores on the raw data. With the 100 most informative features, we obtain f1 scores of 75.0%. This clearly outperforms the majority baseline of 20.0%. It also performs significantly better than using only the 100 most informative syntactic (68.0%) or all 129 lexical features (69.3%). Examination of the feature ranking by information gain reveals that the selected features are in fact highly diverse: The ten most informative features include measures of lexical diversity, verb variation, length, clausal conjunction, coverage of deagentivation patterns, and verb mode. Among the top 50 features, there are instances of all feature groups.

The second experiment replicates experiment 1 on the normalized data. We obtain the highest f1 score of 74.1% using 150 features, which does not differ significantly from the results in experiment 1. It significantly outperforms the classifiers trained with other feature subsets. Interestingly, while the lexical model retains 69.9%, performance of the syntactic model increases significantly to 71.1%. The top ten features remain stable compared to the ranking from experiment 1. However, the top 20 features now include

varying word frequency measures as well as features measuring parse tree complexity. These are virtually irrelevant to the raw data, which indicates high sensitivity to non-standard data for these features. When regarding the top 50 features, psycholinguistic measures of total integration cost show to improve greatly, too. These rank increases are mostly at the expense of POS-based lexical and morphological features.
Overall, our experiments show that a large feature set covering a broad range of linguistic modeling is quantitatively and qualitatively beneficial for proficiency assessment. While non-standard data impacts the reliability of computationally more complex features, they mostly remain highly informative. Features directly based on parse accuracy or word form, such as tree complexity and word frequency measures, however, become mostly uninformative on non-standard data without normalization.

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(1), 111-126.

Barzilay, R., Lapata , M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics,* 34, 1-34.

Brown, C., Snodgrass, T., Kemper, S., Herman, R., Covington, M. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, 40(2), 540-545.

Crossley, S., Kyle, K., McNamara, D. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4), 1-11.

Crossley, S., Kyle, K., McNamara, D. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1-16.

Crossley, S., Salsbury, T., McNamara, D. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243-263.

Feng, L., Jansche, M. (2010). A Comparison of Features for Automatic Readability Assessment. *Coling 2010: Poster Volume*. Beijing, China, 276-284.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (Eds.). *Image, language, brain*. Cambridge: MIT Press, 95-12

Graesser, A., McNamara, D., Louwerse, M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behaviour Research Methods, Instruments, and Computers. 3(2), 193-202.

Hancke, J. (2013). Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language. *MA thesis*. Eberhard-Karls-Universität Tübingen.

Hancke, J., Vajjala, S., Meurers, D. (2012). Readability Classification for German using lexical, syntactic and morphological features. *Proceedings of COLING*. Mumbai, 1063-1080.

Henning, M., Niemann, R. (2013). Unpersönliches Schreiben in der Wissenschaft: Eine Bestandsaufnahme. *Informationen Deutsch als Fremdsprache*, 4, 439-455.

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. *PhD thesis*. Georgia State University.

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

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.

McNamara, D., Graesser, A., McCarthy, P., Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Camebridge University Press.

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.

Petersen, S., Ostendorf, M. (2010). A machine learning approach to reading level assessment. *Computer Speech and Language* 23, 86-106.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, 49-58.

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L., Bernhard, D. (2013). Coherence and cohesion for the assessment of text readability. *Natural Language Processing and Cognitive Science* 11, 11-19.

Vajjala, S. Meurers, D. (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada, 163-173.

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

Acquisition of tense and aspect in learner English: a cross-sectional perspective

Valentin Werner[1], Robert Fuchs[2]
University of Bamberg[1], Hong Kong Baptist University[2]
valentin.werner@uni-bamberg.de, rfuchs@hkbu.edu.hk

In the spirit of Housen (2002) and Myles (2015), this paper provides a corpus-based (re-)evaluation of two established principles in SLA research, (i) the order of acquisition of tense and aspect (OATA) and (ii) the Default Past Tense Hypothesis (DPTH).
The relevant literature on the OATA in learner English (see, e.g., Bardovi-Harlig 2000; Keck & Kim 2014) widely agrees on a development along the lines presented in (1).

(1) simple present/present progressive > simple past/past progressive > present perfect > present perfect progressive > past perfect > past perfect progressive

On a related note, the DPTH (Salaberry & Ayoun 2005), originally stated for learners of Romance languages, predicts that learners in early-intermediate stages will use a single morphological marker for past-time reference. For EFL learners, the most likely candidate is the simple past, as exemplified in (2) and (3).

(2) I watched a lot of good movies (ICCI-POL-746)
(3) I knew that it was my fault (ICCI-AUT-590)

The evidence supporting the OATA and DPTH largely relied on controlled environments and experimental techniques, such as elicitation tasks or observation of smaller learner groups (see, e.g., the overview in Bardovi-Harlig 2000: 206-210). Thus, there is a need to test these hypotheses on a broader empirical basis. In our study, we apply a corpus-based approach to test whether the said key constructs in SLA research can be traced quantitatively in the data. More specifically, (i) we tackle the issue whether the learner corpus data map the OATA and the DPTH in learners of English, (ii) we test the influence of other factors, such as morphosyntactic form (e.g. irregular vs. regular past tense) and complexity, the use of time adverbials, and verb frequency, and (iii) we assess the universal status of the OATA and the DPTH by providing a view across various learner samples with differing L1 backgrounds (cf. Collins 2002).
We establish a cross-sectional view of tense-aspect acquisition in (tutored) learner writing from the beginning to the advanced level. For beginning/intermediate data, we rely on the *International Corpus of Crosslinguistic Interlanguage* (Tono & Díez-Bedmar 2014), a collection of argumentative and descriptive school essays; for advanced data, we rely on the *International Corpus of Learner English* (Granger et al. 2009), a collection of timed and untimed essays by university students majoring in English. Both corpora come with extensive metainformation on the learners. In accordance with our research aims, we use four components of both corpora to map differences and similarities across a set of typologically different L1 language backgrounds (Germanic: German, Sinitic: Chinese, Slavic: Polish, Romance: Spanish).
Overall results are in accordance with the OATA and confirm previous findings in the area of tense-aspect acquisition (e.g. on the late emergence of the present perfect; Fuchs, Götz & Werner 2016), although two important qualifications regarding complex forms apply: (i) present (perfect) progressives seem to emerge at later stages than predicted; (ii) some

templates (such as the (past) perfect progressive) are hardly used even by advanced learners (also in comparison to native speakers; cf. Biber et al. 1999: 462), so that a general pattern "simple before complex" emerges. The results are further in line with the DPTH, finding a near-exclusive reliance on simple past forms for past-time reference in beginning-intermediate stages (grades 5 to 10), and thus extend its assertions to the EFL context. In addition, the data indicate that developmental patterns vary across learner samples with different L1 backgrounds, drawing attention to a factor sometimes neglected in SLA research (Shirai 2009). As far as secondary factors are concerned, we find that irregular verbs (except *be* and *have*) are avoided in the past tenses (compared to their use in the present) by younger learners. This group also disproportionately uses time adverbials in conjunction with verbs, compared to older learners.

The methodological aim of the paper is to show what can be gained by a cross-sectional design, including several learner populations with multiple language backgrounds and proficiency levels. We thus also seek to contribute to the continuum from small-scale, closely monitored learner settings to (by comparison) large-scale learner databases. We submit that both data types possess their inherent strengths and weaknesses, but that linking them leads to an overall more accurate picture of the tense-acquisition acquisition process in English, which may eventually inform the sequencing of EFL material (cf. Hahn 2007: 19).

References:

Bardovi-Harlig, K. (2000). Tense and Aspect in Second Language Acquisition: Form, Meaning, and Use. Malden: Blackwell.

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

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

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. Berlin: Mouton de Gruyter, 297-337.

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

Hahn, A. (2007). Learning and Teaching Processes: Teachers' Learning and Teaching Strategies for Tense and Aspect. Berlin: Langenscheidt.

Housen, A. (2002). A corpus-based study of the L2-acquisiton of the English verb system. In S. Granger, J. Hung & S. Petch-Tyson (Eds.). Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching. Amsterdam: Benjamins, 77-116.

Keck, C., & Kim, Y. (2014). Pedagogical Grammar. Amsterdam: Benjamins.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.). The Cambridge Handbook of Learner Corpus Research. Cambridge: CUP, 309-332.

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. San Diego: Academic Press, 1-33.

Shirai, Y. (2009). Temporality in first and second language acquisition. In W. Klein & P. Li (Eds.). The Expression of Time. Berlin: Mouton de Gruyter, 167-194.
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.

Direct quotes, paraphrases, and summaries in L2 academic assignments

Leonie Wiemeyer
University of Bremen
wiemeyer@uni-bremen.de

This corpus-based study explores intertextual strategies of advanced L2 writers in academic texts. Their use is compared quantitatively and qualitatively in reading reports, i.e. critical summaries of academic research articles, written by German learners of English from the *Corpus of Academic Learner English* (CALE; Callies & Zaytseva 2013). The purpose of the study is to trace the form and function of intertextuality in these assignments and identify aspects which require attention in teaching. Intertextuality is a characteristic feature of academic writing (Hyland 2004). It is typically created through direct quotes, paraphrases, and summaries of source text material. Direct quotation (see Example 2) implies repeating someone else's words verbatim, while paraphrases (Example 3) restate another author's ideas using different words and/or grammatical structures, and summaries (Example 4) represent the gist of a publication without going into detail.

(1) *Source text*: Listening to this group [...] reveals how Māori English helps to create and define Māori students' identity within the confines of physical and social spaces both on and off campus. The study shows that for these Māori students, Māori English functions as an important emblematic marker of their group identity. (King 1999: 36)

(2) *Direct quote*: Her final thought is that the interview with the focus group has shown that Maori students Maori English is **"an important emblematic marker of their group identity"** (King, 1999, 36) within the university. (CALE; RR1.G.HB.101)

(3) *Paraphrase*: In conclusion, the most important finding of King's research is that ME is a mechanism used to **create group identity in physical and social spheres**, especially where Māori constitute a minority. (CALE; RR1.G.HB.104)

(4) *Summary*: The research article by Jeanette King (University of Canterbury) "Talking Bro: Maori English in University Setting" is illustrates how Maori English is used by speakers in University, what does it means for their identity, their group mentality and how it effects their lives in Maori community. (CALE; RR1.G.HB.101)

Using sources in acceptable ways is challenging, especially in the L2 (Abasi & Akbari 2008). Novice L2 writers are often unaware of the functions of paraphrases beyond avoiding plagiarism, overrely on direct quotation, resort to patchwriting, and they generally lack knowledge of academic citing conventions (Davis 2013; Hirvela & Du 2013; Keck 2006; Verheijen 2015). Students who fail to incorporate sources effectively are likely to face accusations of plagiarism (Crocker & Shaw 2002; Pecorari 2003). In recent years, however, inappropriate source use has been reconceptualised as a developmental stage in academic literacy acquisition (Keck 2006). Despite the fact that student writers' competence is often measured against their ability to felicitously reference discipline-specific discourse (Shaw &

Pecorari 2013), there is a notable lack of corpus-linguistic research into intertextuality in L2 academic writing and of teaching resources with contextualised advice and practical examples (Keck 2010; 2015). This contribution aims to address this gap in research by exploring the use of direct quotes, paraphrases, and summaries in reading reports in linguistics. It addresses the following research questions:

- How do L2 writers quote, paraphrase, and summarise content/excerpts from source texts? To what extent do they rely on source text material, and which lexical and grammatical alterations do they make?
- How are intertextual passages documented and attributed in L2 academic writing?
- Which aspects of intertextuality require further attention in teaching?

The form, function, textual integration, and attribution of source material are investigated in order to clarify the strategies of source use employed by advanced German learners of English. For this purpose, instances of intertextuality are manually identified in fifty reading reports from the CALE. Drawing on existing taxonomies (Borg 2000; Campbell 1990; Keck 2006; Shi 2004; Verheijen 2015), they are classified based on closeness to source text, referencing, attribution, and reporting phrases. Intra-annotator agreement will be calculated to ensure reliability. Additional analyses focus on the rhetorical functions of citations and on preferences with regard to which chapters of the original text citations are based on. Inter-learner variability in intertextual practices will also be explored. The results are discussed in the light of findings from previous research into effective and ineffective strategies of source use in L1 and L2 writing.

References:

Abasi, Ali R. & Akbari, Nahal (2008): "Are we encouraging patchwriting? Reconsidering the role of the pedagogical context in ESL student writers' transgressive intertextuality". *English for Specific Purposes* 27, 267–284.

Borg, Erik (2000): "Citation practices in academic writing". In Thompson, Paul (ed.), *Patterns and perspectives: Insights into EAP writing practices*. Reading: University of Reading. 27–45.

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

Campbell, Cherry (1990): "Writing with others' words: Using background reading text in academic compositions". In Kroll, Barbara (ed.), *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University Press. 211–230.

Work in Progress Presentations

Tracking L2 language development through construction of a longitudinal spoken learner corpus

Mariko Abe[1], Yasuhiro Fujiwara[2], Yuichiro Kobayashi[3]
Chuo University[1], Meijo University[2], Nihon University[3]
abe.127@g.chuo-u.ac.jp, fujiwara@meijo-u.ac.jp,
kobayashi0721@gmail.com

The main purposes of this presentation are 1) to overview an innovative research project of compiling a longitudinal learner spoken corpus, 2) to share procedural problems and solutions related to transcribing learners' utterances from audio files, and adding required tags to the texts, and 3) to review initial findings from the corpus, discussing future possible applications of the corpus into learner corpus research (LCR) and second language acquisition (SLA).

As Meunier (2015) stated, even though there has been a dramatic increase in learner corpora in the last two decades, the majority of them are cross-sectional or pseudo-longitudinal in design. Thus, they fail to shed light on complex, and often times unpredictable, developmental patterns of learning and acquisition. For example, Abe (2007, 2014) investigated the largest spoken learner corpus in Japan, the NICT JLE corpus, of which the dialogue test the *Standard Speaking Test* is based on, and found that various types of linguistic features can be used to distinguish differences in oral proficiency levels. However, the NICT JLE corpus consists of cross-sectional data, and with such data it is impossible to see learning trajectories. In other words, it is impossible to see how each individual learner will progress or regress in their L2 learning and acquisition over time. Consequently, it is crucial to conduct a study using adequate amounts of longitudinal data (Larsen-Freeman & Cameron, 2008).

The newly-developed learner corpus of English in this current study is designed to directly grasp L2 developmental patterns in a literal sense, not only as a whole group, but also on an individual basis. This study will collect the same learners' task performances three times a year for the three consecutive years from 2016, creating nine data collection points in total. This will be, to our best knowledge, the largest longitudinal spoken corpora of beginners in the world, which will have the potential for new insights in regard to LCR and SLA research.

Samples will be collected from approximately 120 secondary school students. The students are asked to take a monologue speaking test, the "*Telephone Standard Speaking Test*," consisting of various tasks (e.g., description, comparison, and reasoning), and their utterances in L2 will be compiled to create our learner corpus. In a typical EFL context of Japan, they have hardly any opportunity to speak the target language inside and outside of the classroom. However, the participants of our research project are given sufficient speaking tasks to apply newly learned grammatical forms to real communication.

Along with the corpus development process, an abundance of relevant metadata will be collected and added to the texts to make full use of this new longitudinal spoken learner

corpus. With this design, we can gain new insights into learner language development. For example, what impact (a) individual differences such as motivation, personality, and learning style, (b) English use, (c) task type, and (d) oral proficiency may have on the speech of learners of English.

This project aims to investigate the language use of individual learners through large-scale data collection, and therefore the study will focus specifically on the quantity and speed of transcription and tag-annotations. The development of a spoken corpus requires the collection of all relevant information, and the maintaining of high levels of consistency (Thompson, 2005). Thus, clear guidelines for transcription and a thorough procedure for checking each transcription were established to reduce the risk of inconsistency. In order to ensure that the procedure is followed, each transcriber will be double-checked by another transcriber, and then the researchers will monitor the work of the transcribers. Regarding tag annotations, considering the interchangeability of the resource, the XML format is chosen for the mark-up of the transcribed texts. The tag sets specified in the guidelines cover the following components: utterances, pauses, fillers, repetitions, self-corrections, and so forth. To speed up and increase the accuracy of tag annotations, the transcripts are automatically annotated and then manually checked. Automated discourse tagging rather than automated part-of-speech tagging is also a newly developed technology in corpus linguistics. Accordingly, this study has the potential to provide new insights concerning automated speech tagging as well.

References:

Abe, M. (2007). A corpus-based investigation of errors across proficiency levels in L2 spoken production. *JACET Journal*, *44*, 1-14.

Abe, M. (2014). Frequency change patterns across proficiency levels in Japanese EFL learner speech. *Journal of Applied Language Studies, Special issue on "Learner language and learner corpora"*, *8*(3), 85-96.

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

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.). *The cambridge handbook of learner corpus research*. Cambridge: C UP, 378-400.

Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books, 59-70.

'Speaking in tongues': EFL learners' use of 'foreign words' in informal interviews

Sylvie De Cock
Université catholique de Louvain
sylvie.decock@uclouvain.be

The Louvain International Database of Spoken English Interlanguage (LINDSEI) contains informal interviews with intermediate to advanced level learners of English as a foreign language. The interviews follow the same set pattern and are made up of three main tasks: a personal narrative based on a set topic (an experience that taught them a lesson, a country that impressed them, or a film or play they liked/disliked), a free discussion mainly about university life, hobbies, foreign travel or plans for the future and a picture description. Although the interviews are all conducted in English, 'foreign' words ('FWs'), i.e. words from other languages than English, sometimes feature in the spoken productions. Foreign words have been specially marked up in the LINDSEI corpus (<foreign> WORD(S) </foreign) and can therefore be retrieved automatically using WordSmith Tools for example.

A previous study (De Cock 2015) explored the use of foreign words and their functions in five of the subcorpora included on the LINDSEI CD-ROM (Gilquin et al. 2010), namely LINDSEI_Dutch, LINDSEI_French, LINDSEI_German, LINDSEI_Italian and LINDSEI_Spanish. The study reveals that the frequency and the dispersion of foreign words varies quite markedly across the various subcorpora, with the French- and German-speaking learners using over twice as many FWs as the Spanish-speaking learners for example. The FWs, which come overwhelmingly but not exclusively from the learners' mother tongue, fall into four main functional categories:

(1) lexical bridges, which help learners bridge vocabulary/lexical gaps (words/expressions that appear to be unknown or inaccessible to them; e.g. 'cotizar', 'des algues', 'lasser'),

(2) cultural/institutional bridges, which denote aspects of the education system, events, folklore, places, etc. typically associated with some of the regions/countries mentioned in the set topic and free discussion parts of the interviews (e.g. 'Tour de France', 'Parco Nazionale del Gran Paradiso', 'Vlaamse Opera', 'Abitur', 'gilles de Binche'). This category clearly illustrates the impact of what is discussed on the use of FWs in the interviews,

(3) pragmatic/discourse bridges, which fulfil basic pragmatic/discourse functions in the learners' L1 (e.g. 'ja', 'allez', 'si', 'enfin', 'bueno'),

(4) FWs used in direct speech reporting or in metalinguistic discussions (e.g. 'all she could say was <foreign> ich liebe dich </foreign>' - LINDSEI_Dutch , 'in Spanish they they call it <foreign> chela </foreign>' - LINDSEI_Spanish.

The study shows that, while cultural/institutional bridges are the preferred functional category (with the largest proportion of FW tokens – around 40%) in LINDSEI_French, LINDSEI_German and LINDSEI_Spanish, pragmatic/discourse bridges and lexical bridges are the preferred categories in LINDSEI_Dutch (52%) and LINDSEI_Italian (44%) respectively.

This paper sets out to extend the investigation of FWs to the other six learner varieties included on the LINDSEI CD-ROM (i.e. LINDSEI_Bulgarian, LINDSEI_Chinese, LINDSEI_Greek, LINDSEI_Japanese, LINDSEI_Polish, LINDSEI_Swedish) and addresses the following main research question: how widespread is the use of FWs among EFL learner interviewees from a variety of mother tongue backgrounds? Frequency of use, FW lexical variation, dispersion, individual learner differences and preferred functional categories are examined and compared in the eleven learner varieties. The possible impact of interviewer variables such as status, mother tongue and knowledge of other foreign languages on the learner interviewees' use of non-English words is also analysed. The 2015 study showed that LINDSEI_Spanish contains the lowest number of FWs (compared with LINDSEI_Dutch, LINDSEI_French, LINDSEI_German and LINDSEI_Italian). It was suggested that the interviewer's status (i.e. whether or not the learner is familiar with / knows the interviewer) might affect learners' degree of use of FWs as LINDSEI_Spanish is the only subcorpus investigated in the 2015 study where the interviews were conducted either by an interviewer the learners did not know at all or by an interviewer who was labelled as only 'vaguely familiar' to the learners. This paper aims to further explore the possible impact of learners' level of familiarity with the interviewer by extending the analysis to other subcorpora that exhibit different degrees of familiarity (e.g. LINDSEI_Greek, LINDSEI_Bulgarian).

References:

De Cock, S. (2015) An exploration of the use of foreign words in interviews with EFL learners: a(n) (effective) communication strategy? Paper presented at LCR 2015, Nijmegen September 2015.

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

Annotating a German L1 Learner Corpus for Research on Orthography Acquisition

Stefanie Dipper, Anna Ehlert, Ronja Laarmann-Quante, Katrin Ortmann, Maurice Vogel
Ruhr-University Bochum, Germany
dipper@linguistics.rub.de, anna.ehlert@rub.de, laarmann-quante@linguistics.rub.de, katrin.ortmann@rub.de, maurice.vogel@rub.de

While learner corpus research is a flourishing discipline, the main focus is on L2 learners, and L1 learner corpora of written texts are still comparatively rare (but see for example Abel, Glaznieks, Nicolas, & Stemle, 2014, and Berkling et al., 2014, for German; Barbagli, Lucisano, Dell'Orletta, Montemagni, & Venturi, 2016, for Italian; Parr, 2010, for English). Yet, research on literacy acquisition of children can gain new insights with the help of corpus analyses (see, e.g., Fay, 2010, on the development of children's orthographic competence in freely written texts in German). We want to help filling this gap by providing a longitudinal corpus of freely written German texts produced by primary school children from grade 2-4. Between 2010 and 2012, children from 15 different classes in North Rhine-Westphalia/Germany, many of them with an immigration background, were asked to write down a story shown in a sequence of pictures at 10 different points in time. This yielded a corpus of roughly 2000 texts produced by about 250 children. On average, there are 7.4±2.1 texts per child.

The compilation of the corpus is part of a research project that is concerned with the role of implicit learning during literacy acquisition (see, e.g., Perruchet, 2008, for an overview). We want to investigate the relationship between spelling errors and the orthographic properties of words on different levels for good and poor writers. On the one hand, we want to examine surface properties of words such as n-gram frequencies, on the other hand we want to take orthographic phenomena into account, such as consonant doubling (<Kanne> 'pot') or vowel-lengthening <h> (<fahren> 'to drive'). One of the key hypotheses is that errors that good writers commit are more strongly correlated with the orthographic properties of German words than those of poor writers, who commit errors of rather arbitrary types.

We are currently in the process of transcribing the texts and providing a number of innovative annotations. Firstly, we create a target hypothesis (see, e.g., Reznicek et al., 2013 for L2 data) for each token which only corrects orthographic errors, disregarding grammatical errors. For this purpose, we developed comprehensive guidelines (Laarmann-Quante et al., 2017) which address difficult cases to achieve maximal consistency. For constructing the target hypotheses, we achieved word-based average agreement of 94.99% (SD: 2.12) between two annotators (four annotators in total).

Only correcting orthography may result in target tokens that are non-grammatical word forms in German. For instance, the form ~<treffte>[1] for <traf>, which is similar to ~<meeted> for <met> in English, is the result of an incorrect, but plausible, inflection. This is an error beyond orthography, hence it is not corrected in our orthographic target hypothesis. Instead, the word form is marked as non-existing in German.

At the time of submission, 1318 texts (143,202 tokens, i.e. 109 tokens per text on average) have been transcribed and annotated with a target hypothesis. 17.72% of the tokens contain one or more spelling errors and 0.46% of the target tokens are marked as non-existing word forms.

Besides the target hypothesis, each token is enriched with further annotations: for each target word, information about its phonemes, graphemes, syllables and morphemes will be provided. These are annotated automatically with the help of existing tools (Reichel, 2012). Within our project, further procedures have been implemented to (semi-)automatically annotate further details about each spelling error (Laarmann-Quante, 2016; Laarmann-Quante, Knichel, Dipper, & Betken, 2016):

- an error category (according to our own fine-grained annotation scheme)
- whether the error affects the pronunciation of the word
- whether orthographic knowledge of a related word form is necessary to arrive at the correct spelling ("morpheme constancy")
- whether the target word is a foreign word
- whether the misspelling resulted in another existing word
- whether the syllable structure of the word is violated

Some of the automatic annotations have already been evaluated with promising results and they are currently still being improved. The fully annotated corpus will not only be a basis for innovative analyses with regard to orthography acquisition, it can also be used by NLP applications, for instance as training data for a spell checker targeted at L1 learners (compare Flor & Futagi, 2012, on English L1/L2 learners). An extension to annotate grammatical errors as well is planned for the future.

References:

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 learner corpus for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2414–2421). Reykjavik, Iceland.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2016). CItA: An L1 Italian learner corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 88–95). Portorož, Slovenia.

Berkling, K., Fay, J., Ghayoomi, M., Hein, K., Lavalley, R., Linhuber, L., & Stüker, S. (2014). A database of freely written texts of German school students for the purpose of automatic spelling error classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1212–1217). Reykjavik, Iceland.

[1] ~ = ungrammatical word form

Fay, J. (2010). *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben: Eine empirische Untersuchung in Klasse 1 bis 4*. Frankfurt a. M.: Peter Lang.

Flor, M., & Futagi, Y. (2012). On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 105–115).

Laarmann-Quante, R. (2016): Automating multi-level annotations of orthographic properties of German words and children's spelling errors. In *Proceedings of the 2nd Language Teaching, Learning and Technology Workshop (LTLT)* (pp. 14-22). San Francisco, USA.

Laarmann-Quante, R., Knichel, L., Dipper, S., & Betken, C. (2016). Annotating spelling errors in German texts produced by primary school children. In A. Friedrich & K. Tomanek (Eds.), *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 32–42). Berlin, Germany.

Laarmann-Quante, R., Ortmann, K., Ehlert, A., Betken, C., Dipper, S. & Knichel, L. (2017). *Guidelines for the Manual Transcription and Orthographic Normalization of Handwritten German Texts Produced by Primary School Children*. Bochumer Linguistische Arbeitsberichte (BLA), Vol. 20.

Parr, J. M. (2010). A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research*, *2* (2), 129–150.

Perruchet, P. (2008). Implicit learning. In H. L. Roediger (Ed.), *Cognitive Psychology of Memory* (pp.597-621). Oxford, UK: Elsevier.

Reichel, U. (2012). PermA and Balloon: Tools for string alignment and text processing. In *Proc. Interspeech*. Portland, Oregon.

Reznicek, M.; Lüdeling, A. ,& Hirschmann, H. (2013). Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp.101-123). Amsterdam: John Benjamins.

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 in previous research (see e.g. Chen, 2006; Bolton et al. 2002). Granger & Tyson (1996) conducted a research on the 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; Chen, 2006; Heino, 2010; Martinez, 2004) found similar results for both ESL and EFL learners in that they have problems in the use of conjunctions. The findings mentioned have also been observed for Indonesian EFL learners (Swan & Smith, 2001; Ishak, 2002; Moehkardi, 2002; Marzuki & Zainal, 2004; Kurniyati, Prihadi & Rahayu, 2012; Antara, 2015). However, there is a general lack of corpus studies on Indonesian EFL learners because the corpus-based approach is not (yet) popular in Indonesia. Therefore, the aim of this study is to fill this research gap and examine Indonesian EFL learners' argumentative writing from a learner corpus perspective.

This work-in-progress report outlines the aims and the methodology of the project and presents a pilot study of 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), one of the largest learner corpora focusing on EFL learners from Asian countries. It currently includes 1.3 million tokens of argumentative essays produced by 2600 college students in Asian countries (including Indonesia). It also includes comparable writing from more than 200 English native speakers. The essays are based on two topics that are "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". 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 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?
3. How does Indonesian EFL learners' use of connectors compare to that of native English speakers in argumentative essay writing?

The method used in this study is Contrastive Interlanguage Analysis (CIA; Granger, 2015) which involves two types of comparison. The first is 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 the second one is to compare the use of connectors in argumentative essay produced by Indonesian EFL learners and English native speakers (ICNALE_ENS). For the analysis, the connectors will be 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 process is carried out by means of *UAM Corpus Tool* (O'Donnell, 2015). When analyzing the data, quantitative and qualitative approaches will be 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 preliminary results illustrate that the distribution of the different semantic categories is - surprisingly - nearly identical between the learners at the A.2. level and the English native-speaker students groups. Learner at the A.2 level show a tendency to use more linking adverbials as the English native-speaker-students, especially contrastive and resultative one, while the learners at the B.1.2 level employ additive and appositive devices much more than the learners at A.2 level and the native speakers.

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). 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, English, Stockholms 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. Volume 1. Kobe: Kobe University Press, 91-118.

Kurniyati, P & Rahayu (2012). Analisis Kesalahan Kohesi Dan Koherensi Paragraf Pada Karangan Siswa Kelas X Sma Negeri 3 Temanggung, e-Journal Universitas Negeri Yogyakarta, Vol 1, No 2.

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. 2nd Edition. Cambridge: Cambridge University Press.

Constrained Language Use: Using data-driven mixed methods to investigate the common ground between learner language and translated language

Ilmari Ivaska, Silvia Bernardini
University of Bologna
ilmari.ivaska@unibo.it, silvia.bernardini@unibo.it

Corpus research interested in universal tendencies and varietal typicalities is often faced with a data comparability paradox: on the one hand, comparable data should be used in order to account for the normal variation and to make sure the different studied varieties represent (a) comparable language use situation(s) or genre(s) (Granger 2015; Baker 2007). On the other hand, several typologically different languages should be studied parallelly to justifiably distinguish language-dependent and language-independent tendencies. Such representative multilingual data have mainly been collected in highly regulated and limited language use situations, most notably in the institutions of the European Union (e.g. Tiedemann 2012). The gain in comparability entails a loss in representativeness.

In this work-in-progress paper, we present a new research project that aims to contrast instances of language use where more languages than one are inherently present. Second or foreign language (L2) and translated language (L1T) can both be seen as such instances. Indeed, they have both been suggested to exhibit linguistic constraints or divergences from non-translated native language use (L1O), sometimes called learner/translation universals (Lanstyák & Heltai 2012). The project is located at the intersection of second language acquisition and translation studies, and the bottom-up data-driven approach adopted allows for novel, critical, and detailed definitions and understanding of the general, language-specific or variety-related typicalities, which are investigated simultaneously in three typologically diverging languages: English, Finnish, and Italian.

The paper explores and exemplifies the ways in which the use of existing large scale corpora can be combined with a novel highly controlled trilingual data set of L2, L1T, and L1O, together with acceptability judgement surveys and ethnographic interview data. The proposed mixed methods approach seeks to provide a solution to the data comparability paradox, and to take into account both systemic and individual facets of language use. More specifically, we apply a stepwise methodological procedure called key structure analysis (e.g Ivaska 2015) and use n-gram frequencies to reveal constructions that typically distinguish L1O, L1T and L2 in the three studied languages (cf. Baker 2004; Granger 2014). We then compare the findings across languages and use the results as a point of departure for a more detailed multifactorial statistical analysis (e.g. Gries & Deshors 2014) both in the large scale corpora and in a novel, highly comparable data set, to be collected as part of the project in the UK, Finland, and Italy. We complement this novel data set with experimental and more qualitative methodological approaches, including acceptability judgement surveys and ethnographic interviews. The exact test items in the surveys as well as the framework of the interviews will be partially defined based on the results of the key structure analysis. The data and the methodological procedure are depicted in figure 1. In contrast to earlier work on constrained language use (e.g. Kruger & Van Rooy 2012; Nisioi

et al. 2016), the present project will lead to a bottom-up definition of typical phenomena, effectively making it possible to evaluate the crosslinguistic and language-specific nature of the observed phenomena. It will also enable one to address questions regarding the existence and the nature of universal tendencies in constrained language use, and the relationship between translated language and learner language.

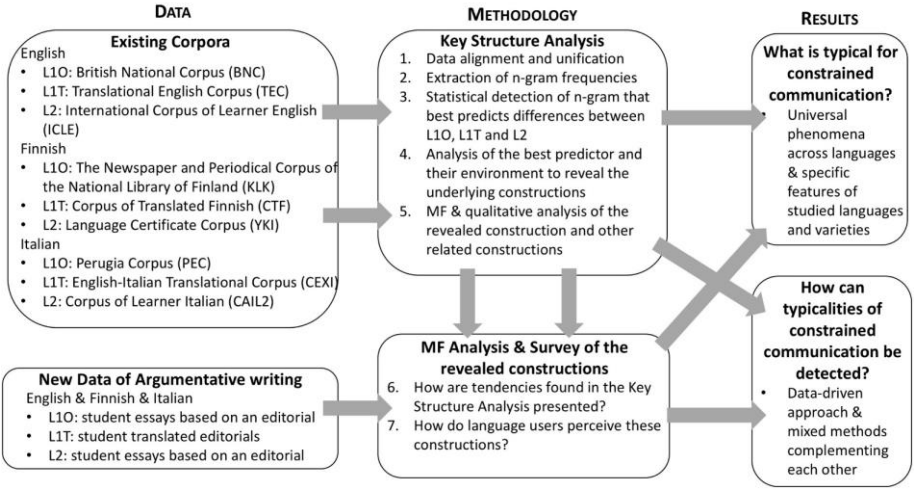


Figure 1. Data and methodological workflow of the project.

References:

Baker, M. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2), 167-193.

Baker, M. 2007. Patterns of idiomaticity in translated vs. non-translated English. *Belgian Journal of Linguistics* 21, 11-21.

Granger, S. 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14(1), 58-72.

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

Gries, S. & S. Deshors 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9, 109-136.

Ivaska, I. 2015. Longitudinal changes in academic learner Finnish: A key structure analysis. *International Journal of Learner Corpus Research* 1, 210-241.

Kruger, H. & B. van Rooy 2012. Register and features of translated language. *Across Languages and Cultures* 13, 33-65.

Lanstyák, I. & P. Heltai 2012. Universals in language contact and translation. *Across Languages and Cultures* 13, 99-121.

Nisioi, S., E. Rabinovich, L. Dinu, & S. Wintner 2016. A Corpus of Native, Non-native and Translated Texts. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4197-4201.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Ph.D. research: Can corpora be used effectively in English Language Teaching in Norway?

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

Breyer has written that there appears to be a 'consensus… that the impact of corpora on classroom practices and language education in general has remained limited… A small number of publications available on this subject appear to confirm this' (Breyer 2011: 119). My planned research contains three components: a survey of teachers' familiarity with corpora, a semester-long course in corpus linguistics for teachers, and case studies of how some of the participants in the course subsequently make use of corpora in or for the classroom.
Breyer interviewed teacher educators (Breyer 2011: 121, 145-154) and sought teacher trainees' views (Breyer 2009: 154). Farr (2008), Hüttner et al. (2008), Lenko-Szymanska (2014) and Zareva (2017) also evaluated the effectiveness of corpora for teacher trainees. My project offers a chance to look at the gap in this research, that is, how in-service teachers deal with corpora. We must make a distinction between teacher educator, in-service teacher, and pre-service teacher trainee. I propose to work with teachers, to solicit their views, to introduce them to corpora, and to follow their corpus work into the classroom.

Structure: three articles.

First article: The gathering of data through two surveys. First, a survey of English teachers. There are questionnaires in the literature (e.g. Breyer 2011: 161-2 and Lenko-Szymanska 2014: 268) that have been used to ascertain what respondents know about corpora, opening for example with the question, 'Have you ever heard the term corpus and do you know what it is?' (Lenko-Szymanska 2014: 268). Almost any English teacher can therefore be surveyed; I have communicated with the relevant organization in Norway about how this can be managed. A second set of data will be gathered through a seminar, where I introduce corpora to teachers. In-service teachers come to my institution for study points, as part of a programme called "Kompetanse for Kvalitet" ("KfK").

Second article: It will be necessary to make a group or groups of teachers familiar with corpora. I would incorporate corpus linguistics into a semester-long "KfK" course or courses. Teachers can be introduced to corpora, concordancing software, and uses for corpora in their teaching (indirect use in preparation, or direct use in the classroom). Establishing which material is the most user-friendly will require me to conduct in the first instance a pedagogical review of software. Teachers' assignments in corpus linguistics in the course can be collected as data (subject to their permission). There is precedence for the usefulness of such data: Hüttner et al. (2008) and Breyer (2011: 175-185) analysed corpus assignments by teacher trainees.

Third article: In-depth case studies conducted over some months. A small number of volunteers would be found from the teachers who receive the abovementioned introductory corpus education. These volunteer participants would then receive further support in corpora use. The corpus support would involve helping the volunteers to incorporate the use of corpora into their teaching. This means using a corpus or corpora in a way that is relevant to them and their pupils. The focus of the observation is how the teacher uses corpora and whether the teacher finds it beneficial to have done so. Observation is usually combined with interview, and in this case the interviews would be significant data. The main interview questions would be designed to find out if the teachers perceived whether or not there are any learning benefits for their pupils.

References:

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

Breyer, Yvonne. (2011), Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges. Peter Lang GmbH, Internationaler Verlag der Wissenschaften. Available at: <https://www.researchgate.net/publication/264499058_Corpora_in_Language_Teaching_and_Learning_Potential_Evaluation_Challenges>.

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

Hüttner, Julia, Ute Smit, and Barbara Mehlmauer-Larcher. (2008), 'ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis', System 37, 99-109.

Lenko-Szymariska (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.

Zareva, Alla (2017), 'Incorporating corpus literacy skills into TESOL teacher training', ELT Journal 71/1, 69-79.

Do Estonian ELF speakers follow similar patterns of article use as identified for English as a lingua franca?

Merli Kirsimäe, Jane Klavan
University of Tartu
merli.kirsimae@gmail.com, jane.klavan@gmail.com

Over the past decade, research into English as a lingua franca (ELF) has steadily developed into a thriving field. ELF has been studied from the perspective of pronunciation, lexis, grammar and pragmatics (Dewey, 2007b; Önen, 2014; Seidlhofer, 2011; Ur, 2010; Walker, 2010). Our research proceeds from two basic assumptions: first, that English does not *belong* to native speakers, since there are now more non-native than native speakers of English (Cogo & Dewey, 2006; Seidlhofer, 2011) and second, that ELF is not a *language* as such, but rather "a means of communication not tied to particular countries and ethnicities, a linguistic resource that is not contained in, or constrained by, traditional (and notoriously tendentious) ideas of what constitutes 'a language'" (Seidlhofer, 2011: 81). Although ELF has obtained world-wide recognition, no excessive research on it has been carried out in Estonia. The specific aim of our study is to take stock of the use of articles by Estonian ELF speakers. Our aim is to test whether the predictions put forward on the use of articles in previous ELF studies (e.g. Dewey, 2007a,b; 2009; Seidlhofer, 2011) are also present in the Estonian ELF data, taking into consideration the fact that Estonian lacks a system of articles. In addition to pinpointing the general trends, we are interested in finding out the potential factors behind the usage patterns attested in the data. It has been claimed, for example, that the selection of an article does not depend on the nature of the noun (i.e. its inherent qualities); instead the use of articles is seen as a resource which is used as a means of giving additional prominence to a referent (Dewey 2007a: 341). Another prominent characteristic is the tendency of ELF speakers to use the zero article in contexts where ENL predicts the use of the definite article (Dewey, 2009: 63). Our aim is to use both qualitative and quantitative methods to test the viability of these claims.

For our research purposes, 9 semi-structured interviews (approximately 85 minutes of speech altogether) were recorded with Estonian speakers of English at B1-C2 level (all university students). All participants (8 of whom were women) reported to be native speakers of Estonian (average age 22.4 years). The interviews were conducted by a Polish native speaker who did not speak Estonian. The interviews were transcribed using the free software EXMARaLDA and the VOICE conventions for transcription (https://www.univie.ac.at/voice/). The transcribed data were manually annotated for relevant variables by two independent annotators. The annotation schema includes a selection of semantic, morpho-syntactic and discourse-related variables (e.g. the use of article, referent, context of use, and the speaker, among others).

The preliminary qualitative analysis shows a tendency to omit definite articles in places where the articles are made redundant by the uniqueness of the words they should be preceding (such as ordinals and superlatives) (cf. Dewey, 2007a: 341). However, other instances of article omission also occur. As we expect to see variation tied to both speakers and the individual referents, the influence of these variables is levelled by using state-of-

the art modelling techniques. A mixed-effects regression model (Pinheiro and Bates, 2002) is fitted to the data with the use of article as the dependent factor and the various semantic, morpho-syntactic and discourse variables as fixed effects. In addition, subjects and referents are included in the model as random effects. Using quantitative methods enables us to see how much of the variation attested in the use of articles by ELF speakers can be attributed to individual speakers and contexts - something that may prove difficult with a qualitative method.

Implications for both ELF theory and methodology will be discussed. By providing the data of Estonian ELF speakers, our study makes a crucial contribution towards validating the generality of the proposed characteristics of ELF. As to the practical outcomes, our study is one of the cornerstones in the development of a larger Estonian corpus of English as a lingua franca. The corpus will serve as a valuable awareness-raising reference material for teachers of English in Estonia.

References:

Cogo, A & M. Dewey (2006). Efficiency in ELF communication: From pragmatic motives to lexico-grammatical innovation. *Nordic Journal of English Studies* 5(2), 59-93.

Dewey, M (2007a). English as a lingua franca and globalization: an interconnected perspective. *International Journal of Applied Linguistics* 17(3), 332-354.

Dewey, M (2007b). *English as a lingua franca: an empirical study of innovation in lexis and grammar* (Unpublished doctoral dissertation). King's College London, London.

Dewey, M (2009). English as a lingua franca: Heightened variability and theoretical implications. In A. Mauranen & E. Ranta (Eds). *English as a Lingua Franca: Studies and Findings* (60 – 83). Newcastle upon Tyne: Cambridge Scholars Publishing.

Önen, S (2014). *Lexico-grammatical features of English as a lingua franca: A corpus-based study of spoken interactions* (Unpublished doctoral dissertation). Istanbul University, Istanbul.

Pinheiro, J. C. & D. M Bates (2002). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Seidlhofer, B (2004). Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics* 24, 209-239.

Seidlhofer, B (2011). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Ur, P (2010). English as a Lingua Franca: A Teacher's Perspective. *Cadernos de Letras (UFRJ)* 27, 85-92.

Walker, R (2010). *Teaching the Pronunciation of English as a Lingua Franca*. Oxford: Oxford University Press.

The use of tense and aspect in English texts written by monolingual and bilingual learners of English as a foreign language

Eliane Lorenz
University of Hamburg
eliane.lorenz@uni-hamburg.de

Globalization and migration movements changed society: Europe has become an area where multilingualism is normal and constantly increasing (cf. Gogolin et al. 2013, Meyer 2008). This development results also in a change of the German classroom. The former monolingual learning environment includes now children from a variety of linguistic backgrounds. Researchers found out that third language acquisition (L3A) and second language acquisition (L2A) differ considerably (De Angelis 2007; Bardel & Falk 2007; Falk and Bardel 2011; Rothman 2011; Siemund 2017). We can assume that monolinguals transfer from their native language when acquiring a foreign language; bilingual or multilingual learners possess two or more potential resources for both positive and negative transfer when acquiring an additional foreign language (Siemund 2017). This should result in multilinguals having an advantage over monolinguals when acquiring another foreign language in school. Yet, in the current literature, we find contrasting models concerning L3A (cf. the L1 Factor Model (Na Ranong & Leung 2009); the Cumulative Enhancement Model (Flynn et al. 2004); the Typological Primacy Model (Rothman 2011)). One theory argues the L2 to be the language that is mainly transferred from in L3A (Bardel & Falk 2007; Falk & Bardel 2011; Rothman 2011). However, Cummins' points out that the heritage language "can be a powerful intellectual resource" (302: 2013). Yet, the student's awareness of their resources needs to be stressed, in order to make such cross-linguistic connections and to profit from them (Cummins 2007). Based on this, the present paper follows the aforementioned theory and examines the assumption that the heritage language can also be a source for transfer. If this is true, is this transfer positive or negative? Can multilingualism be a positive resource for studying another language and does it put multilingual students in a beneficial situation?
The languages under investigation are German and Russian as native languages and English as the foreign language currently acquired. Two groups of the participants in this study are students in school year 7 and 9 growing up in Germany: L1 German (n=40) and L1 Russian/ L2 German (n=40). The third group of participants are L1 Russian speakers (n=20). Their task was to write an English text to a picture story and to fill in a questionnaire asking for personal information. With the help of this sample, I intend to provide evidence that not only the main language of the country (i.e. German, the language of the environment) but also the first native language (Russian) can influence the performance in the additional foreign language. The three languages differ substantially in their morphology: on a continuum with analytic languages on one end and fusional languages on the other end of the spectrum, English belongs more to the group of analytic languages, whereas German and Russian belong to the group of fusional languages (Siemund 2017). Hence, the

representation of grammatical tenses and the expression of aspectual meaning differs crucially in the languages focused here. The bilingual participants are expected to produce significant differences in the use of tense and aspect from their Russian or German peers. Results reveal, for example, a difference in the use of the progressive aspect: the bilingual students appear to be somewhere in between the monolingual Germans and monolingual Russians, when considering formal correctness and the target-like meaning of the progressive. The cover term 'correct usage' was separated into these two categories (adopted from Bardovi-Harlig 1992) to differentiate between form (i.e. auxiliary present, subject verb agreement) and use (verb commonly used in the progressive or not). The Germans produced more formally correct progressives than the Russians. It is the other way around with the number of progressives that express target-like meanings. Here, the Russian students scored a higher number than the German students. An explanation for this outcome is the following: in German, we do not find a fully grammaticalized system of the progressive (cf. Siemund 2013) but several lexical items that correspond to the English progressive aspect (König & Gast 2012: 92-93). Yet, we do find complex tenses, similar to English (cf. König & Gast 2012). Russian, on the other hand, differentiates between different aspectual concepts, as opposed to German. However, the formal representation of tense and aspect differs to English: in Russian, we do not find auxiliaries but affixes attached to the verb (cf. Wade 1992).

References:

De Angelis, G. (2007). *Third or Additional Language Acquisition*. Clevedon: Multilingual Matters.

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research*, 23(4), 459-484.

Bardovi-Harlig, K (1992). The relationship of form and meaning: A cross-sectional study of tense and aspect in the interlanguage of learners of English as a second language. *Applied Psycholinguistics*, 13, 253-278.

Cummins, J. (2007). Rethinking monolingual instructional strategies in multilingual classrooms. *Canadian Journal of Applied Linguistics*, 10, 221–240.

Cummins, J. (2013). Current research on language transfer: Implications for language teaching policy and practice. In: P. Siemund, I. Gogolin, M. Schulz & J. Davydova (Eds.). *Multilingualism and Language Contact in Urban Areas. Acquisition – Identities – Space – Education*. Amsterdam: Benjamins, 289-304.

Falk, Y., & Bardel, C. (2011). Object pronouns in German L3 syntax. Evidence for the L2 status factor. *Second Language Research*, 27(1), 59-82.

Gogolin, I., Siemund, P., Schulz, M., & Davydova, J. (2013). Multilingualism, language contact, and urban areas. An introduction. In P. Siemund, I. Gogolin, M. Schulz & J. Davydova (Eds.). *Multilingualism and Language Contact in Urban Areas. Acquisition – Identities – Space – Education*. Amsterdam: Benjamins, 1-15.

König, E., & Volker G. (2012). *Understanding English-German Contrasts.* 3rd Edition. Berlin: Erich Schmidt Verlag.

LiMA, Linguistic Diversity Management in Urban Areas-LiPS, LiMA Panel Study. (2009-2013). *Projektkoordination LiPS: Prof. Dr. Dr. H. C. Ingrid Gogolin; ©LiMA-LiPS 2013.* Hamburg: LiMA.

Meyer, B. (2008). Nutzung der Mehrsprachigkeit von Menschen mit Migrationshintergrund. Expertise für das Bundesamt für Migration und Flüchtlinge. Universität Hamburg.
Rothman, J. (2011). L3 syntactic transfer selectivity and typological determinacy. The typological primacy model. *Second Language Research*, 27(2), 107-127.
Siemund, P. (2013). *Varieties of English. A typological approach.* Cambridge: Cambridge University Press.
Siemund, P. (2017). Englisch als weitere Sprache im Kontext herkunftsbedingter Mehrsprachigkeit. In J. Duarte, I. Gogolin, T. Klinger, B. Schnoor & M. Trebbels (Eds.). *Sprachentwicklung im Kontext von Mehrsprachigkeit – Hypothesen, Methoden, Forschungsperspektiven*. Wiesbaden: Springer VS.
Wade, T.L.B. (1992). *A comprehensive Russian Grammar.* Oxford: Blackwell.

English Prosody of Advanced Learners: A Contrastive Interlanguage Analysis with Language-Pedagogical Implications

Karin Puga,
Justus Liebig University Giessen,
Karin.Puga@anglistik.uni-giessen.de,

Prosodic deviances in nonnative speech can contribute to a perceived foreign accent and/or impede communication, intelligibility, and comprehensibility (cf. Jilka 2000, 2007; Mennen et al. 2014; Trofimovich & Baker 2007). A lot of meaning, which cannot be inferred by grammar or lexis, can be conveyed by intonation and is, therefore, an important aspect of learning and speaking different languages. However, even advanced learners still deviate from native-like intonation patterns (cf. Bongaerts et al. 1997; Scovel 2000). Previous studies on L2 prosody (e.g. Gut 2009 (German); Ramírez Verdugo 2002 (Spanish)) reported that learners overuse rises and replace falls with rises and vice versa. Although the general interest in L2 prosody has grown (cf. Mennen & de Leeuw 2014; Li & Post 2014; Mennen et al. 2014), there is still a demand for more exhaustive approaches, especially those that adopt a language-pedagogical angle.

In response to this need, the present study adopts a corpus-based approach examining native and interlanguage data. Thus, this study sets out to characterize the intonational features produced by three L2 English learner groups and investigates the extent to which the learners adopt native values of the target language. The study focuses on the following research questions:

1. What are the structural and functional features of prosody in the spoken interlanguage of advanced learners of English with different L1-backgrounds? Can universal features be observed across different language families?
2. To which extent do these learners diverge from the native-speakers' prosodic patterns or adopt language-appropriate values in spontaneous speech?

By answering these research questions, this study not only provides valuable insights to understanding the concepts of nonnative intonation patterns, it also has implications for Second Language Acquisition research and teaching L2 intonation.

Within the autosegmental-framework this paper reports on a study on L2 learners' intonational deviances in spontaneous monologic and dialogic speech derived from a "Contrastive Interlanguage Analysis" (CIA) (cf. Granger 1998) and "Contrastive Analysis" (Granger 1998) of Czech, German, and Spanish. Through quantitative and qualitative analyses, the structure of intonational phrases and the frequency and use of pitch and phrase accents, and boundary tones are compared. Additionally, the interlanguages based on the Czech, German, and Spanish components of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010) are compared to English native speech with prosodically annotated versions of the *Louvain Corpus of Native English Conversation* (LOCNEC; cf. De Cock 2004), representing British English, and the *Charlotte Narratives and Conversation Collection* (CNCC; cf. Atkins 2017) corpus, representing

American English. Since neither of the corpora include prosodic annotations, the manual annotation had been performed with the *Tone and Break Indices* (ToBI) (Silverman et al. 1992) system. In total the corpus consists of 225 annotated files (corresponding to 225 speakers, á 45 per group), which contain similar spontaneous speech on similar topics (e.g. a country travelled to) produced in an interview situation, the same age group (18-30), and same length of speech (≈550 tokens = 1-2 minutes) consisting mostly of statements, to ensure data comparability. For the analysis of the files, information was extracted and significance tests were conducted by the help of Praat- (Boersma & Weenink 2016) and R-scripts (R Core Team 2015).

The analysis of a smaller subset of the data (n=30) indicates that German, Spanish and British speakers of English, deviate from each other in their intonational phrasing and pitch heights at utterance-final and -medial position. For instance, the learners broke their utterances into considerably more intermediate phrases (ip) and intonation phrases (IP) than native-speakers. While the British native-speakers (n=10) mainly stick to the usage of falls within IPs and ips, the German learner's (n=10) speech was characterized by a frequent usage of rising tones and the Spanish learners (n=10) overuse falling tones. This deviant usage of pitch found in the pilot study can be attributed to many different factors such as L1-transfer and/or developmental factors, the relationship between the interviewer and the interviewee, genre-dependent differences (dialogue vs. monologue), learner variables (age, gender, years of English, stays abroad, etc.), pragmatic functions such as turn-taking, influences of regional accents (cf. Grabe et al. 2000), other native, foreign, or second languages that have been learned, etc. Finally, further explanations are suggested for the differing intonation patterns by learners and language-pedagogical implications derived from the pilot study are discussed.

References:

Atkins Library (2017). *New South Voices Collection*. Atkins Library, University of North Carolina at Charlotte.

Boersma, P. & D. Weenink (2016): *Praat: doing phonetics by computer*. (Version: 5.3.69). [Computer Software]. 29.03.2016. < http://www.fon.hum.uva.nl/praat/>.

Bongaerts, T., C. Van Summeren, B. Planken & E. Schils (1997): "Age and ultimate attainment in the pronunciation of a foreign language", *Studies in Second Language Acquisition* 19, 447-465.

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

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

Grabe, E., B. Post, F. Nolan & K. Farrar (2000): "Pitch accent realization in four varieties of British English", *Journal of Phonetics* 28, 161-185.

Granger, S. (1998): "The computer learner corpus: a versatile new source of data for SLA research", *Learner English on Computer*, ed. S. Granger. London: Longman. 3-18.

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

Jilka, M. (2000): *The Contribution of Intonation to the Perception of Foreign Accent*. PhD thesis, Institute of Natural Language Processing, University of Stuttgart.

Jilka, M. (2007): “Different manifestations and perceptions of foreign accent in intonation”, *Non-Native Prosody – phonetic description and teaching practice*, eds. J. Trouvain & U. Gut. Berlin: Mouton de Gruyter. 77-96.

Li, A. & B. Post (2014): “L2 Acquisition of prosodic properties of speech rhythm – Evidence from L1 Mandarin and German Learners of English”, *Studies in Second Language Acquisition* 36 (2), 223-255.

Mennen, I. & E. de Leeuw (2014): “Beyond Segments: Prosody in SLA”, *Studies in Second Language Acquisition* 36, 183-194.

Mennen, I., F. Schaeffler & C. Dickie (2014): “Second language acquisition of pitch range in German learners of English”, *Studies in Second Language Acquisition* 36, 303-329.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scovel, T. (2000): “A critical review of the critical period research”, *Annual Review of Applied Linguistics* 20, 213-223.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert & J. Hirschberg (1992). “ToBI: A standard scheme for labeling prosody”, *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, 867-879.

Trofimovich, P. & W. Baker (2007): “Learning prosody and fluency characteristics of L2 speech: The effect of experience on child learners’ acquisition of five suprasegmentals”, *Applied Psycholinguistics* 28, 251-276.

Annotation of cohesion in learner corpora. Insights from an in-depth analysis of a longitudinal data set of German L2

Carola Strobl
Ghent University
carola.strobl@ugent.be

In this presentation, the results of an in-depth analysis of cohesion development in the writing of four students of German over a period of five collegiate semesters will be presented. This analysis serves as a pilot study for a planned annotation and contrastive interlanguage analysis (CIA) of cohesion based on a large corpus of L1 and advanced L2 writing in German, the Falko summary corpus (Reznicek, Lüdeling, & Schwantuschke, 2012). The long-term goal of this project is to create an instrument for appropriate pedagogical support with regard to cohesion in L2 academic writing.

Cohesion in written learner language is an important, yet underresearched area of investigation (Lee, 2002). The current pedagogical approach towards cohesion in language curricula is characterised by a focus on grammatical accuracy, rather than on stylistic appropriateness, which has been identified as harmful, because it causes L2 writers to making non-nativelike use of cohesive features (Gilquin & Pacquot, 2007; Granger & Tyson, 1996). L2 writers´ cohesion problems persist even at advanced stages of acquisition, when starting to write academic text genres like summaries and essays (Hinkel, 2002; Reid, 1992; Segev-Miller, 2004).

To adequately address the cohesion problems of L2 writers, first, their use of non-target like features has to be investigated through a CIA based on comparable corpora of native and non-native speakers. CIA research for German L2 has started to receive more attention in the last decade. Yet, cohesion has rarely been included in earlier studies (Belz, 2005; Walter, 2007; Walter & Schmidt, 2008). A comprehensive analysis of cohesion in German L2 writing, covering the range of cohesive devices described in Halliday and Hasan (1976), listing coreference, lexical cohesion, ellipsis, connectives, and substitution, has not been undertaken to date.

The theoretical basis for this project is a comprehensive description of cohesive devices in German that was elaborated in the context of a multilingual corpus analysis project (Kunz, Maksymski, & Steiner, 2009; Lapshinova-Koltunski & Kunz, 2014). This description served as a point of departure for the in-depth analysis of cohesion in a longitudinal corpus of learner German collected at a Belgian University. A second theoretical-methodological source of inspiration is Langlotz (2014) who exemplified how the degree of integration of connecting propositions (as opposed to mere aggregation) can be used as a predictor of writing proficiency development.

The corpus consists of short texts (100 - 280 words) that were produced without auxiliary means under timed conditions (20 minutes). They were written in response to general prompts about student life topics. Four texts were collected of each participant (n=23) over a period of five semesters of their bachelor study in Applied Linguistics. Their proficiency levels range from A2-B1 of the Common European Framework in the first occasion of data collection to B2-C1 in the fourth occasion. For this in-depth analysis, four representative

case studies were selected based on their initial proficiency level and the destination of their study abroad semester.
The results show little development over time with regard to coreference strategies, as well as to lexical cohesion strategies which rarely exceeded recurrence. The most interesting results were found in the areas of ellipsis and connectivity. In all four students, a development towards diversification of connector use over time was observed, which, however, did not necessarily coincide with an increased integration of connecting propositions. Overall, students displayed a continued preference for aggregative connecting strategies. Lower-proficient students made non-targetlike use of connectives in terms of syntactic embedding. At a higher proficiency level, especially after spending a semester abroad in a German-speaking country, students tended to confuse formal and non-formal language registers, leading them to use ellipses and connective structures in a way that is typical for the (semi-)oral register of written chat.
The conclusions that can be drawn from this pilot study in view of the annotation of a larger corpus for a contrastive analysis of cohesion strategies are that (a) ellipsis and connectivity are two areas that deserve specific attention and (b) with regard to the latter, it is important to annotate the degree of syntactic integration of connected propositions. Based on Langlotz´ (2014) results, it is expected that the degree of connective integration marks a difference between L1 and L2 writers.

References:

Belz, J. A. (2005). Corpus-driven characterizations of pronominal da-compound use by learners and native speakers of German. Die Unterrichtspraxis / Teaching German, 38(1), 44-60.

Gilquin, G., & Pacquot, M. (2007). Spoken features in learner academic writing: identification, explanation and solution. Paper presented at the Proceedings of the Fourth Corpus Linguistics Conference CL2007, University of Birmingham.

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. doi: 10.1111/j.1467-971X.1996.tb00089.x

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longman.

Hinkel, E. (2002). Second Language Writers´ Text. Linguistic and Rhetorical Features. Mahwah, N.J.: Erlbaum.

Kunz, K., Maksymski, K., & Steiner, E. (2009). Suggestions for a corpuslinguistic analysis of cohesion (Deliverable No. 3). Deliverables of CroCo project Uni Saarland. Universität des Saarlandes. Retrieved from http://fr46.uni-saarland.de/uploads/media/GECo_AP3.pdf

Langlotz, M. (2014). Junktion und Schreibentwicklung: Eine empirische Untersuchung narrativer und argumentativer Schülertexte. Berlin, Boston: De Gruyter.

Lapshinova-Koltunski, E., & Kunz, K. (2014). Annotating cohesion for multilingual analysis. Paper presented at the Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, Reykjavik.

Lee, I. (2002). Teaching coherence to ESL students: a classroom inquiry. Journal of Second Language Writing, 11(2), 135-159. doi: 10.1016/s1060-3743(02)00065-6

Reid, J. (1992). A computer text analysis of four cohesion devices in english discourse by native and nonnative writers. Journal of Second Language Writing, 1(2), 79-107. doi: 10.1016/1060-3743(92)90010-m

Reznicek, M., Lüdeling, A., & Schwantuschke, F. (2012). Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.01. Humboldt-Universität zu Berlin. Institut für deutsche Sprache und Linguistik - Korpuslinguistik. Berlin. Retrieved from https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01

Segev-Miller, R. (2004). Writing from Sources: The Effect of Explicit Instruction on College Students' Processes and Products. L1-Educational Studies in Language and Literature, 4(1), 5-33. doi: 10.1023/B:ESLL.0000033847.00732.af

Walter, M. (2007). Hier wird die Wahl schwer, aber entscheidend. Konnektorenkontraste im Deutschen. In H.-J. Krumm (Ed.), Theorie und Praxis - Österreichische Beiträge zu Deutsch als Fremdsprache (Vol. 10, pp. 145-161). Innsbruck, Wien, Bozen: StudienVerlag.

Walter, M., & Schmidt, K. (2008). Und das ist auch gut so! Der Gebrauch des satzinitialen "und" bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. In B. Ahrenholz, U. Bredel, W. Klein, M. Rost-Roth & R. Skiba (Eds.), Empirische Forschung und Theoriebildung. Beiträge aus Soziolinguistik, Gesprochene-Sprache- und Zweitspracherwerbsforschung. Festschrift für Norbert Dittmar zum 65. Geburtstag. (pp. 331-342). Frankfurt / Main: Peter Lang.

Poster Presentations

Corpus-aided Error Analysis (CEA) of Accuracy and Proficiency in Learner Finnish

Sisko Brunni[1], Jarmo H. Jantunen[2], Valtteri Airaksinen[1]
University of Oulu[1], University of Jyväskylä [2]
sisko.brunni@oulu.fi, jarmo.jantunen@jyu.fi, valtteri.airaksinen@oulu.fi

The present study focuses on errors at the different proficiency levels in the Common European Framework of Reference for Languages (CEFR; Council of Europe 2011). Although error analysis is a debated method (see e.g. Ellis & Barkhuizen 2005), the procedure of error identification, classification and explanation is still an essential method in learner corpus research (see e.g. Dagneaux et al. 1998, Thewissen 2015). For example, recent learner corpus studies on English (e.g. Thewissen 2015) have shown that error-tagged learner corpora are valuable sources of accurate information on the development of language proficiency.

In the present study, we argue that in CEA focusing on development and proficiency levels, it is useful to utilize learner corpora of morphologically rich languages such as Finnish. The rich and varying morphology of Finnish – caused, for example, by inflection – produces partly dissimilar problems in language learning compared to the morphologies of less agglutinative languages and creates challenges for error categorization and analysis. The error analysis in this study thus has methodological implications in the sense that it develops an error taxonomy that is suitable for Finnish.

The data come from the International Corpus of Learner Finnish (ICLFI), which consists of pseudolongitudinal learner data. At the moment, the error-annotated sub-corpus of ICLFI contains 1,182 texts, with a total of 184,000 tokens. The error classification system of ICLFI is hierarchical, and covers different components of language. The nine main error categories are orthographic, phonological, morphophonological, morphological, morphosyntactic, syntactic, lexical, phraseological classes (and class unexplainable). Most of these contain several subcategories, bringing the total number of error categories to 32. During the error annotation process, tagging is conducted by several people. Annotation is based on an error-tagging manual and error classifications from previous EA studies. Each text is error-annotated by one annotator, but problematic cases are negotiated with other annotators to reach agreement (Brunni, Lehto, Jantunen & Airaksinen 2015.) To acquire an overall picture of the erroneous forms, the errors are first normed per tokens and the shares of main error categories are then counted. However, since the counting and norming of errors per total dataset gives only a rather vague picture of the errors and quality of the texts, the present analysis utilizes potential occasion analysis (see Thewissen 2015: 143–145), in which errors are counted in all of the cases they could potentially occur. Depending on the error category, the denominator used in analysis can be a specially created denominator (e.g. part-of-speech), the amount of sentences per text or the amount of tokens per text.

In the analysis, we concentrate mainly on the CEFR levels A2–B2, since they are the largest subcorpora in the ICLFI. The errors per token at A2–B2 proficiency levels are as follows: A2: 7,282 errors / 27,196 tokens; B1: 20,600 errors / 88,810 tokens; B2: 7,774 errors / 54,896

tokens. The initial comparison focuses on the nine main error categories, but those subcategories that are frequent and essential in describing the differences between levels are also discussed.

The results show that the overall number of errors decreases as the proficiency level grows. However, despite the change in level, the shares of the different main error categories remain, surprisingly, rather similar. Potential occasion analysis reveals three tendencies in how the number of errors develops as proficiency level grows: first, the number of errors decrease as the proficiency level grows (A2 > B1 > B2); second, the number of errors is highest on the B1 level and lowest on the B2 level (A2 < B1 > B2); and third, the number of errors is highest on the B1 level and lowest on the A2 level (A2 < B1 > B2 (>A2). Similar tendencies, improvement and regression (along with stabilisation) were found in Thewissen's (2015: 272) study of levels B1–C2.

The first tendency is the most common and describes the general tendency in language learning. The second type indicates that as a learner's language skills grow, the more the learner tests new structures and more complex grammar (e.g. modifying structures), which, in turn, generates new types of errors. In that regard the tendency of a seeming regression can also be seen as a signpost of development (Thewissen 2015: 273). The last tendency suggests that there are also some features that cannot be easily mastered even at the higher proficiency levels, such as phraseology. The qualitative analysis of error types reveals the actual erroneous forms that learners produce and provides more accurate information on language development.

References:

Brunni, S., Lehto, L.-M., Jantunen, J. H. & Airaksinen, V (2015). How to annotate morphologically rich learner language: principles, problems and solutions. In A.-K. Helland Gujord, S. Nacey, S. Ragnhildsveit (Eds.) *Learner Corpus Research: LCR2013 Conference Proceedings*. Bergen Language and Linguistic Studies.

Council of Europe (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-Aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics*, 26(2), 163.

Ellis, R. & Barkhuizen, G (2005). *Analysing learner language*. OUP.

Thewissen, J. (2015). Accuracy across proficiency levels. A learner corpus approach. Presses Universitaires de Louvain.

Tracking Written Learner Language (TRAWL): A longitudinal corpus of Norwegian pupils' written texts in second/foreign languages

Hildegunn Dirdal[1], Eli-Marie Danbolt Drange[2], Anne-Line Graedler[3], Tale M. Guldal[2], Ingrid Kristine Hasund[2], Susan Lee Nacey[3], Sylvi Rørvik[3]
University of Oslo[1], University of Agder[2], Inland Norway University of Applied Sciences[3]
hildegunn.dirdal@ilos.uio.no, eli.m.drange@uia.no, anneline.graedler@inn.no, tale.guldal@uia.no, kristine.hasund@uia.no, susan.nacey@inn.no, sylvi.rorvik@inn.no

TRAWL is a research project where the primary objective is to explore and describe how Norwegian pupils develop writing skills in second and foreign languages throughout their education journey. Texts are being collected longitudinally from pupils at Norwegian schools and compiled into a searchable corpus, which will remain accessible as a resource for researchers, teachers and teacher educators after the end of the project. Our poster will present the TRAWL corpus and describe its design, the transcription and annotation of the texts, the research aims of the project group and other potential applications for the corpus.

Design:
Learner corpora containing data from the early stages of SLA are scarce (Tono et al. 2012: 8), as are longitudinal learner corpora with data collected from the same learners over time. Longitudinal data are essential since the time perspective is a crucial aspect of the language learning process and development in general (Granger 2012: 11). Also, more data from a Norwegian context are needed. The CORYL corpus contains error-tagged texts by Norwegian pupils in years 7, 10 and 11 in the National tests of English writing from 2004–2005. It is, however, a small corpus, containing cross-sectional data only. English L2 learner texts have also been collected for Norwegian components of three international corpora initiated at the Centre for English Corpus Linguistics (CECL) in Belgium: the written corpora ICLE and VESPA, and the spoken corpus LINDSEI. These corpora are cross-sectional and contain texts written at university level.
The TRAWL corpus supplements these corpora with longitudinal data from younger learners: in years 5–13 for English and years 8–13 for French, German and Spanish, the most common foreign languages that pupils can select to study in addition to English. Data are collected from all school years from the start and will continue for at least three years to allow for both pseudo-longitudinal and truly longitudinal studies. Some pupils are also asked to contribute texts written in Norwegian to enable comparisons of L1 and L2 writing development. The collected texts have been written as part of the pupils' regular class work.

The corpus contains metadata describing pupils (age, gender, language background, etc.) and texts (format, task prompts, task conditions etc.), as well as teachers' written assessment.

Transcription and annotation:
At lower levels, most texts are written by hand. The first stage of data processing is therefore transcription of these texts, without any changes to spelling, grammar or punctuation. Since spelling variation makes automatic searches difficult, corrected versions will be linked up with the primary transcriptions. It will also be possible to view pdf-versions of the (anonymized) original texts.
The texts are annotated using macros and Perl scripts originally created for the British Academic Written English (BAWE) corpus and adjusted for VESPA by Alois Heuboeck and further for TRAWL by Jarle Ebeling (see Ebeling and Heuboeck 2007, Paquot et al. 2015). The annotation follows the TEI conventions and includes sentence, paragraph and various other text divisions, formatting, lists, tables, figures and quotes/mentioned items, as well as the metadata described above.

Research goals:
In addition to the primary objective of the TRAWL research project mentioned above, the secondary objectives are 1) to map grammatical, lexical and text coherence features that characterize learner language at various stages and age levels, 2) to research factors that may affect learner L2 development. The first stage of data collection will enable analysis of cross-sectional data by comparing texts from different levels. At the second stage the data will be genuinely longitudinal, and will contribute with unique empirical evidence for second language (L2) proficiency development.

Other applications:
The TRAWL corpus will remain accessible for researchers, teachers and teacher educators after the end of the project. In addition to providing data for further studies, including master's theses, the corpus will be useful in teacher training. In courses that use corpus data, students currently study language by learners at the same level as themselves. With TRAWL, they can investigate learner language from younger pupils as well. The corpus will also function as a useful source of examples in the development of teaching materials.

References:

Ebeling, S. O. & Heuboeck, A. (2007). Encoding document information in a corpus of student writing: The British Academic Written English Corpus. *Corpora*, 2(2), 241–256.

Granger, S. (2012). How to Use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.). *Research Methods in Second Language Acquisition: A Practical Guide*. London: Wiley-Blackwell, 7–29.

Paquot, M., Ebeling, S. O., Heuboeck, A. & Valentin, L. (2015). The VESPA tagging manual. Version 2.3. Centre for English Corpus Linguistics, Université catholique de Louvain.

Tono, Y., Kawaguchi, Y. & Minegishi, M. (2012). *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins.

Error annotation by means of the Scope – Substance Error Taxonomy

Nikola Dobrić, Günther Sigott, Hermann Cesnik
Alpen-Adria-Universität Klagenfurt
nikola.dobric@aau.at, sigott.guenther@aau.at, hermann.cesnik@aau.at

Since the beginning of error analysis as an applied linguistics discipline a variety of error taxonomies have been proposed. Broadly speaking, they are based either on linguistic categories or on surface-structure descriptions (George 1972; Dulay, Burt & Krashen 1982; James 1998; Havranek 2002; Ellis & Barkhuizen 2005; Dobric and Sigott 2014; Pibal, Sigott & Cesnik , forthcoming). One of the problems inherent in these taxonomies is the high degree of subjectivity and the resulting lack of annotator agreement (Landis & Koch 1977; Carletta 1996; Gwet 2001; Viera & Garrett 2005; Díaz-Negrillo & Fernández-Domínguez 2006). The Scope – Substance error taxonomy takes an alternative approach to error description by using the concepts of scope and substance to describe errors. In principle, this distinction was already suggested by Lennon (1991), who used 'extent' to refer to scope and 'domain' to refer to substance. Scope refers to the linguistic and extralinguistic context that needs to be taken into account for an error to be noticed whereas substance designates the size of the linguistic structure that needs to be changed in order for the error to disappear. Levels of scope and substance are described in terms of word, phrase, clause, sentence and text (Quirk et al. 1985). For instance, if we look at the example below and try and locate the error, we can see that the error only becomes noticeable when one extends the context of observation (scope) beyond the individual words and beyond the verb phrase to the level of the clause.

[Example 1] There was it beautiful and very interesting.

Only when we look at the entire clause does it become clear that clause structure rules of English have been violated. In cases like this, we will say that the scope of the error is clause. The substance of the error is clause as well, because rectifying it involves changing the clause structure by changing the word order.

[Example 2] There is a lot of evidence that body art was used three to five thousand years BC, and it is believed that the first one of them was made by accident.

On the other hand, Example 2 shows a sentence consisting of two coordinated clauses. When considered in isolation, neither of them violates any grammatical rules of English. However, when they are combined, it becomes obvious that 'body art', being an uncountable noun, cannot serve as an antecedent for 'one of them', which presupposes a countable antecedent. So in this case it is not enough to consider clause-level context, but it is necessary to widen the scope to the level of the sentence. In cases like this we will say that the scope of the error is sentence while the substance is phrase.

This kind of logic enables errors to be described in terms of fourteen error types resulting from combinations of scope and substance into the following categories:

1) scope TEXT substance PUNCTUATION;
2) scope TEXT substance PHRASE;
3) scope TEXT substance CLAUSE;
4) scope TEXT substance SENTENCE;
5) scope TEXT substance TEXT;
6) scope SENTENCE substance PUNCTUATION;
7) scope SENTENCE substance PHRASE;
8) scope SENTENCE substance CLAUSE;
9) scope SENTENCE substance SENTENCE;
10) scope CLAUSE substance PUNCTUATION;
11) scope CLAUSE substance PHRASE;
12) scope CLAUSE substance CLAUSE;
13) scope PHRASE substance PUNCTUATION; and
14) scope PHRASE substance PHRASE.

We contend that this approach to error description should leave less room for individual interpretation because the model of grammatical analysis that it is based on is made explicit. In order to investigate annotator agreement reached on the basis of the taxonomy, a first pilot study with thirteen trained student annotators was conducted in 2014. All students had been introduced to the principles of Quirk et al. (1985) in one of their Introduction to Linguistics classes. The experiment was conducted in a 2 hours per week 4th semester BA course on corpus linguistics taught at the University of Klagenfurt in the summer semester of 2014 by two of the authors. In this course, the students were required to deepen their understanding of the principles of the Comprehensive Grammar of the English Language (Quirk et al. 1985) They were given ample practice in sentence analysis in order to ensure that they would be able to apply the model of analysis to novel text material in a confident way. In particular, they were trained in identifying phrases, clauses and sentences and in recognizing the hierarchical syntactic relationships among these in texts. The annotators were then given electronic access to five texts chosen by the authors as the basis for this study. The choice was made with the intention of representing writing typical of the material available. The annotators were instructed to mark both substance and scope in the five learner texts in the framework of a computer platform developed for large-scale annotation in the future. They were asked to analyse the five texts, all within the A2 CEFR band, as part of their course requirements. Each word and each punctuation mark in the five texts was considered a unit of observation. The annotators were instructed to code each unit of observation in terms of absence or presence of error. Errors were coded using the fourteen errors types resulting from the error taxonomy. Agreement was expressed in terms of Error Location Density Indices developed for this purpose (Sigott, Cesnik and Dobric 2016). The first pilot study has shown that while the approach has potential, annotator agreement was still low. This was attributed to a lack of detail in the instructions provided for annotators. Consequently, the guidelines for the application of the taxonomy were refined. They now contain instructions for setting up an authoritative reconstruction of the learner text by applying the principle of minimal correction. In the case of competing authoritative reconstructions, preference is to be given to the one which involves the smaller change in

substance (cf. Siemen et al. 2006). Furthermore, instructions for dealing with unitary constituency, nested error, multi-level error and multiple error were formulated. The refined guidelines also contain instructions for dealing with missing constituents and superfluous constituents. Moreover, guidelines for dealing with punctuation errors have been added.
Currently, a second pilot study of the taxonomy, which includes the refined guidelines, is in progress. The same learner texts will be used. The study will involve a group of student annotators comparable to the ones in the first pilot study, and a group of university teaching staff, both equally trained on the basis of the updated guidelines and a redesigned training procedure. This will make it possible to investigate the effect of the new guidelines as well as the possible effects of differences in the annotators' language proficiency on annotator agreement. The results will be available later in the year and will be reported in the presentation.

References:

Carletta, J. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic." Computational Linguistics 22 (2): 249-254.

Díaz-Negrillo, A. & J. Fernández-Domínguez 2006. "Error Tagging Systems for Learner Corpora. " RESLA 19: 83-102.

Dobric, N. & G. Sigott. 2014. "Towards an Error Taxonomy for Student Writing. " Zeitschrift für interkulturellen Fremdsprachenunterricht 19 (2) : 111-118.

Dulay, H., B. Marina & S. Krashen 1982. Language Two. Oxford: Oxford University Press.

Ellis, R. & G. Barkhuizen 2005. Analyzing Learner Language. Oxford: Oxford University Press.

George, V. 1972. Common Errors in Language Learning: Insights from English. Rowley: Newbury House.

Gwet, K. 2001. Handbook of Inter-rater Reliability. Maryland: STATAXIS Publishing Company.

Havranek, G. 2002. Die Rolle der Korrektur beim Fremdsprachenlernen. Frankfurt/Main: Peter Lang.

James, C. 1998. Errors in Language and Use: Exploring Error Analysis. London: Longman.

Landis, R. & G. Koch 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33: 159-174.

Lennon, P. 1991. "Error: Some Problems of Definition, Identification and Distinction." Applied Linguistics 12 (2): 180-196.

Pibal, F. , G. Sigott & H. Cesnik. (forthcoming). The Role of Error in Assessing English Writing in the National Educational Standards Baseline Test. In : Sigott, G. (Ed.) Language Testing in Austria: Taking Stock. / Sprachtesten in Österreich: Eine Bestandsaufnahme. Frankfurt / M. et al.: Peter Lang.

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

Siemen, P., A. Lüdeling & F.H. Müller 2006. FALKO – Ein fehlerannotiertes Lernerkorpus des Deutschen. Proceedings of Konvens 2006.

Sigott, G., H. Cesnik, and N. Dobrić. 2016. "Refining the Scope – Substance Error Taxonomy: A Closer Look at Substance.V In : Dobrić, N., Graf, E. and Onysko, A. (Eds.) Corpora in Applied Linguistics - Current Approaches. Newcastle upon Tyne: Cambridge Scholars Publishing : 79-94.

Viera, A. & G. Joanne 2005. "Understanding Interobserver Agreement: The Kappa Statistic. " Family Medicine 37 (5): 360-363.

Learning Italian verb-noun collocations through corpora: a pilot study

Luciana Forti
Università per Stranieri di Perugia
luciana.forti@unistrapg.it

The last few years have seen a rise in the construction of Italian native and non-native corpora. However, despite some sound attempts to investigate the potential of using corpus data in an Italian L2 teaching and learning context (Chiari, 2011; Ducati & Leone, 2009; Kennedy & Miceli, 2010; Corino, 2014), empirical research into its effectiveness has so far been quite scarse, as opposed to EFL learning contexts (Johns, 1990; Sinclair, 2004; Leńko-Szymańska & Boulton, 2015; Cobb & Boulton, 2015). Collocations in learner language and learner acquisition processes have also a solid research tradition, though again mainly grounded into investigating phenomena related to the English language (Durrant, 2008; Nesselhauf, 2005; Granger, 1998; Gyllstad & Wolter, 2016).
This poster presents the findings of a pilot study aimed to evaluate the effectiveness of using corpus data to aid the acquisition of Italian verb-noun collocations. The study is based on a longitudinal controlled design, with four data collection points over a time span of 12 weeks. The data was collected through a collocational proficiency test, aimed at eliciting both definitional and trasferable knowledge of the collocations.
The corpus data used to inform learning materials and the creation of the collocational proficiency test was extracted from PEC (*Perugia Corpus*) (Spina, 2014), a native Italian reference corpus, and from LoCCLI (*Longitudinal Corpus of Chinese Learners of Italian*), an Italian learner corpus.
The learners were Chinese mothertongue speakers, with a pre-intermediate proficiency level. They were all attending general Italian language courses at the University for Foreigners of Perugia as part of the Marco Polo and Turandot programs, with the intention of pursuing academic studies in Italy upon successful completion of the course. The classes were randomly assigned to the experimental and control conditions. The first were exposed to paper-based concordance materials, while the second to traditional materials. Both received a 1-hour lesson per week for eight weeks, focusing on the same collocational learning aims.
The overall set of verb-noun collocational learning aims was made of 64 collocations. The list was created by integrating the verb-noun collocations found in LoCCLI that contained errors, with the most significant ones found in PEC and included in the DICI-A project (Spina, 2016). The list of 64 collocations was then divided into 8 sets of thematically linked collocations, containing 8 collocations each which were established as weekly learning aims. Both the experimental and the control series of lessons were taught by the experimenter.
The design of the experimental activities draws upon guided discovery techniques and is theoretically informed by notions related to lexical priming and pattern grammar (Hoey, 2005; Hunston & Francis, 2000; Sinclair, 2003).

In summary, the main questions of the pilot study are: 1. How does concordance-based learning affect the acquisition of Italian verb-noun collocations compared to traditional activities? 2. Are there any significant learning differences across collocation categories? The results of the pilot study will also be considered in regards to possible modifications to the main study.

References:

Chiari, I. (2011). Teaching Language Variation Using Italian Corpora. *Corpora, Language, Teaching, and Resources: From Theory to Practice*. Ed. N. Kuebler. Berlino: P.I.E. Peter Lang, 301–322.

Cobb, T., & Boulton, A. (2015). Classroom Applications of Corpus Analysis. In Biber, D. & Reppen, R. (Eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 478–497.

Corino, E. (2014). Didattica delle lingue corpus-based. *EL.LE*, *3*(2), 231–258.

Ducati, R., & Leone, P. (2009). Corpora e apprendimento del lessico. Risorse per docenti dai programmi nazionali. Programma Operativo Nazionale 2007-2013 'Competenze per lo Sviluppo' cofinanziato dal Fondo Sociale Europeo.

Durrant, P. (2008). *High frequency collocations and second language learning. Unpublished PhD dissertation*. Nottingham: The University of Nottingham.

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.

Gyllstad, H., & Wolter, B. (2016). Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter?: Collocational Processing and Semantic Transparency. *Language Learning*, *66*(2), 296–323.

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

Hunston, S., & Francis, G. (2000). *Pattern Grammar*. Amsterdam - New York: Benjamins.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28–44.

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

Sinclair, J. M. (2003). *Reading Concordances*. London: Pearson.

Spina, S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014* (Vol. 1). Pisa: Pisa University Press, 354-359.

Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In B.Sanromán Vilas (Ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching* (Mémoires de la Société Néophilologique de Helsinki C). Helsinki: Société Néophilologique, 219-244.

We Agreed to Disagree: Agreement Patterns in Learner English

Lenka Garshol
University of Agder
lenka.garshol@uia.no

This poster presents results of a Ph.D. project with a focus on English as a second language in Norway. One of the goals of this project is to categorize and analyze subject-verb agreement error patterns occurring in English interlanguage produced by young Norwegian learners. Previous studies of this type have focused on advanced learners studying English at the university level (Johansson, 2008; Thagg Fisher, 1985) and offered only cross-sectional perspective. The current project looks at younger learners who are followed for one year, which offers a closer look at the developing interlanguage.

Subject-verb agreement errors are quite common in this learner population despite the relatively high fluency and complexity of their texts (usually B1-B2 on the CEFR scale). The data used in this poster consist of written texts of 198 Norwegian high school students, each followed for one school year. The texts are compiled into a corpus (approx. 400 000 words) which is screened for subject-verb agreement errors. Found instances of the erroneous agreement are further analyzed based on the type of subject (pronominal, full NP with/without post-head material, coordinated NPs, clause as subject) and the type of verb (BE or other) to uncover possible error patterns in the interlanguage.

Learners of English as a second language often have problems with the marking of the third person singular in the present tense (Cook, 2008). Young children acquiring English as their first language also acquire the third person -s as one of the last inflectional morphemes (Radford, 1990). However, both of these learner groups normally omit the morpheme in the contexts where it is required before they start producing the standard forms. The Norwegian learners consistently over-produce the third person -s overgeneralizing this verbal pattern into all persons in both singular and plural. Out of the agreement errors detected in this learner corpus, the majority are occasions of plural subjects or first and second person singular subjects combined with verbs with the third person singular morpheme -s. Furthermore, this overgeneralization often prevails even after ten years of English instruction, and there is no noticeable improvement in the error scores of the students during their last year of instruction, suggesting that such use is already fossilized in their interlanguage. Recent paper on grammaticality judgements by young Norwegian students (Jensen, Westergaard, & Slabakova, 2017) found similar patterns of erroneous judgements, which indicates that such errors are not slips, but reflect the state of the interlanguage of this student population.

Recurring patterns of non-standard syntax can have several explanations. First, they could be part of the normal acquisition process; in which case, similar patterns should also be detected among learners of other languages at a similar proficiency level. Second, they could be signaling transfer from the first language; in which case, similar patterns should be detected among learners whose L1 is closely related to Norwegian, such as Swedish or Danish. Thagg Fisher (1985, p. 189) explains some of the plural NP + singular V errors as "-s

preservation" arguing for a process of copying of the -s morpheme from a plural head noun onto the following verb. However, this would not explain the errors where the plural noun is not directly followed by the verb or where the plural noun does not form a regular plural with the -s morpheme. The current author argues instead for a cross-linguistic transfer as the dominant cause of the erroneous overgeneralization. The Norwegian learners may be influenced by the verbal pattern in their first language and use it as a null hypothesis in their L2 learning. Norwegian uses the suffix -r for all persons in the present tense, while the suffix-less verb form is only allowed in infinitive constructions. The young learners' English production in this study seems to follow this pattern.

References

Cook, V. J. (2008). *Second language learning and language teaching* (4th ed.). London: Hodder Education.

Jensen, I. N., Westergaard, M., & Slabakova, R. (2017). *The Bottleneck Hypothesis in L2 acquisition: Norwegian L1 speakers' knowledge of syntax and morphology in English L2*. Paper presented at the GASLA 14, University of Southampton.

Johansson, S. (2008). *Contrastive analysis and learner language: A corpus-based approach*. University of Oslo. Retrieved from http://www.hf.uio.no/ilos/forskning/grupper/Corpus_Linguistics_Group/papers/contrastive-analysis-and-learner-language_learner-language-part.pdf

Radford, A. (1990). *Syntactic theory and the acquisition of English syntax : the nature of early child grammars of English*. Oxford: Blackwell.

Thagg Fisher, U. (1985). *The sweet sound of concord: a study of Swedish learners' concord problems in English.* CWK Gleerup, Lund.

Inter-Annotator Agreement Measures for Error Annotation in Learner Corpus Linguistics

Lucie Gillová
Charles University, Prague
luckagillova@gmail.com

Corpora are tagged for various features and different systems of annotation are used. As some of these features and tagging systems are dependent on annotators' subjective decisions, it is important to have measures that would evaluate and improve such systems and decision making processes. Since learner corpora are often tagged for errors and deciding what an error is can be very subjective, these measures are necessary.
Inter-annotator agreement (IAA) is a measure that expresses how well two or more annotators can make a decision concerning the annotation of certain cat egory. For some annotation systems it is quite easy to calculate the IAA because every word or sentence is tagged for a certain category and therefore only the agreement on category or subcategory needs to be taken into account. However, calculating IAA for error annotation is more complicated because apart from agreement on category the annotators have to decide which part of the corpus to annotate (meaning not every word or sentence needs to have a tag).
There are several measures of inter-annotator agreement. Probably the most common one is the measure used by Carletta (1996). She works with kappa coefficient K = (P(A) - P(E)) / (1 - P(E)) "where P(A) is the proportion of times that the coders agree and P(E) is the proportion of times that we would expect them to agree by chance". There has not been agreement on the base value of this measure. Reidsma (2008) claims there are fields of research that are more difficult to annotate and this should be taken into account.
As for other measures, Brants (2000) uses F-score that works with recall (number of identical nodes in annotation X and Y / number of nodes in X) and precision (number of identical nodes in annotation X and Y / number of nodes in Y). He uses this measure for structural annotation but we can replace nodes with errors. Artstein and Poesio (2008) provide an overview of methods used for computing agreement used in this study.
The aim of this study is to review the existing IAA measures suitable for error annotation and to test them on the Czech part of LINDSEI. LINDSEI-CZ was tagged for errors using a modified error tagging system used originally for ICLE. This system is currently being modified to be more suitable for spoken language data but for the purposes of this study, the existing annotation was sufficient. Firstly, the corpus was tagged for errors, the annotators received exactly the same instructions and based on these annotations, measures of inter-annotator agreement were tested. The testing was done initially on a smaller sample of data to see more clearly what influences the value of the various measures and what this value expresses when a specific measure is used for error annotation. This initial testing has shown that most of the results are comparable with those found in the literature when annotators marked either same places in the text or at least a similar number of places (e.g. kappa around 0.6 etc.). However, when there is a considerable difference in the tagged parts of the text (e.g. annotator A marks 21 errors

and 6 of them are phrases not marked by annotator B), the results start to be problematic. To explore this further, artificial dataset was created and manipulated to explore the ways the results change when different parameters are changed. This method provides interesting results, showing that kappa coefficient is very suitable for situations when every word is tagged and annotators do not have to choose what to annotate but it could be useful to add other measures in other situations. Therefore, a measure that would take into account different importance of these two parameters is proposed. This allows us to distinguish between situations when there is a possible problem in the annotation system and situations when annotators do not mark the same phrase.

Since this study is a work-in-progress, IAA for the whole corpus is currently being calculated and the results are not available yet but the calculation will be finished soon. A comparable corpus of intermediate English learners is being compiled, the measures will be tested on a random selection of annotated data from this corpus.

References:

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555-596.

Brants, T. (2000). Inter-annotator Agreement for a German Newspaper Corpus. Presented at LREC conference. Available at https://pdfs.semanticscholar.org/d630/9b4cbaf24f65dd9582d21397e5661567fcab.pdf.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254.

Passonneau, R., Habash, N., & Rambow, O. (2006). Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC), 1951-1956.

Reidsma, D. (2008). Annotations and Subjective Machines of Annotators, Embodied Agents, Users, and Other Humans. PhD thesis series No. 08–121. Univ. of Twente.

Learner corpus compilation: steps for a tidy format

Andressa Rodrigues Gomide[1], Deise Prina Dutra[2]
Lancaster University[1], Universidade Federal de Minas Gerais[2]
a.rodriguesgomide@lancaster.ac.uk, deisepdutra@gmail.com

Analysis of learner language has attracted the attention of linguists (e.g. Selinker 1972) for a long time. Some investigations are part of the second language acquisition (SLA) field (Mitchell & Myles 2004; Myles et al. 2013) and others side with Corpus Linguistics (CL) (e.g. Granger 1998, Granger et al. 2015). The combination of both areas has become more frequent as Myles (2015) and Meunier (2015) show. In this paper we focus on how CL can shed light into SLA issues by presenting how a growing learner corpus (Corpus do Inglês sem Fronteiras - CorIsF) has been compiled and organized in a systematic way, so that it can used in cross-sectional and longitudinal studies. At the time of the analysis presented in this paper the corpus was composed of texts totalling 145,043 words. These texts, written by university students majoring in 58 disciplines at four different institutions, were submitted through Google Forms under one of the two conditions: test in language labs or classroom activity. This study consists of making the processing of CorIsF replicable as the procedure was carried using scripts in R, a free software environment for statistical computing and graphics. This procedure was divided in four parts: dataset compilation and pre-processing; dataset processing; extraction of the key features; and data visualization. Variables, such as discipline, type of task and text genre can be chosen for the analysis. This paper has the goal to describe the compilation steps and corpus organization. We, then, present result examples from a cross-sectional perspective, comparing them with the British Academic Written English corpus (BAWE). The first compilation step refers to the cleaning process, such as eliminating unwanted data, and keeping the relevant ones. This was accomplished through the creation of functions to delete information from learners who do not wish to participate in the research, to anonymize the collaborators, and to delete irrelevant information from the dataset. These functions made the cleaning process automatic, so that the data can be continuously cleaned as it grows. In the following step, CorIsF was subset in five small corpora covering two different learner profiles (learners from courses with a high and with a low demand), two different tasks (dependent or independent), and a specific genre (summary). The third step was to have the subcorpora annotated with the Apache OpenNLP part-of speech (POS) tagger (Apache Software Foundation, 2004). The following step was an exploratory investigation to check for within subcorpora variability (e.g. POS, type and token frequency and n-grams). In the final step some exploratory data visualizations were performed with the creation and analysis of plots and wordclouds. After the preparation of the data, the language used in each subcorpora was contrasted and some observations were made. For instance, a high frequency of the word “possible” was observed, in contrast to a low usage of the modal verb “may”, when this corpus was compared to a subcorpus of BAWE. An analysis of the KWIC confirmed that the second language learners tend to overuse the construction “it is possible” when ‘may’ could also be adopted. A main contribution of this work was to set a framework to collect and keep learner data in a tidy format. In this way, once the data goes

through the cleaning process, it can easily be subset according to the research needs. Making the extraction of subcorpora from the dataset before applying the investigation techniques has proved to reduce processing time. Once the subcorpora are set, the investigation functions here developed can then be applied. Additionally, these batches can also be extracted as txt file, so that they can be analysed with more user-friendly interfaces such as AntConc and WordSmiths Tool. The analyses were restricted due to the small size of the corpus. However, as the data grows, new data analytics can be implemented in order to assist further investigations.

References:

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

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

Meunier, F. (2015). Developmental patters in learner corpora. In S. Granger, G. Gilquin, F. Meunier (Eds.). *Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 379-400.

Mitchell, R.; Myles, F. (2004). *Second language learning theories* (2nd ed.). London: Hodder Arnold.

Myles, F., Marsden, E., Mitchell, R., & Babb-Rosensfeld, L. (2013). *Second Language Learning Theories*. New York: Routledge.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, F. Meunier (Eds.). *Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 309-331.

Selinker, L, (1972). Interlanguage. *International Review of Applied Linguistics*, 10(1), 201-231

Preposition selection errors made by Spanish learners of English: exploring the root causes of the errors

Patricia González Díaz
Universidad Autónoma de Madrid
glezdiaz.p@gmail.com

Within the process of acquiring a Second Language, Spanish Learners must frequently deal with specific linguistic aspects that pose problems they find utterly complicated to overcome. Prepositions, due to their heterogeneous nature in each different language, are one of these troublesome aspects.

This poster presents the research on the use of English preposition by Spanish students of English as a Second Language carried on as part of ALEGRO project (Adaptive Learning of English Grammar Online).

ALEGRO's goal is to develop adaptive learning software which will be able to adapt to the Learner's needs during their training in the L2. It will also facilitate a more autonomous acquisition of the language.

The task of this part of the project consists in recognizing the Spanish Learners' misusage of English prepositions, determining and classifying which of these units are more difficult for them, and, also, identifying the linguistic reasons behind these errors. This information will be provided to the learning program to help it to lead the learner through the acquisition of the most critical prepositions.

The study is based on the use of a corpus of texts written by Spanish Learners of English as a second language (L2). The corpus is a 75,000 word subset of the Wricle corpus (Rollinson & Mendikoetxea 2010). Each text is also associated with the proficiency level of the student at the time of writing, in terms of CEFR levels, derived using the Oxford Qucik Placement Test (UCLES, 2001). Corpus annotation was done using UAM CorpusTool (O'Donnell, 2008), a tool for manual corpus annotation.

To limit the scope, the study focuses on prepositions functioning as head of a prepositional phrase. Prepositions functioning in phrasal verbs and prepositional verbs were ignored.

In terms of corpus methodology, each preposition in the corpus was automatically located, and where it was part of a prepositional phrase, a segment was tagged. The study makes use of earlier work which identified all prepositional errors in the corpus, recording as well the preposition that should have been used (Murcia Bielsa & MacDonald, 2013). Where the preposition had not been tagged as an error in the previous study, it was tagged as "correct". Otherwise it was tagged as "incorrect" and an additional tag was provided for the correct preposition.

An additional phase of work involved the current author "back-translating" each clause containing the preposition (deriving the probable Spanish expression which the learner was working from), to identify the most probable source preposition (if it was such in Spanish). This process was supervised and double-checked by Dr Susana Murcia Bielsa, Lecturer in the Department of English Philology at Universidad Autónoma de Madrid, and it was carried out following the standard procedure of "back-translating", that is to say, by means of consulting different glossaries, dictionaries, and text corpora in order to be as accurate as this task demanded.

The rationale behind this step was pedagogic, as my central interest is to identify which Spanish prepositions are most problematic for Spanish learners to express in English. For instance, while it is interesting to know that “in” is the most frequent erroneous preposition produced by the learners, it is more interesting to know that Spanish speakers, translating into English, have a choice between a number of prepositions, which include “in”, but also “on”, “into”, “for” and others, depending on the context.
The next stage of the work will identify the specific semantic contexts which condition the translation of selected Spanish prepositions, with a goal of identifying exactly which of these contexts lead to most interlanguage errors. On the basis of this, materials will be prepared for the online system to help the learners avoid these problems.

References:

Murcia Bielsa, S. and MacDonald, P. 2013. "The TREACLE project: Profiling learner proficiency using error and syntactic analysis". In S. Granger, G. Gilquin and F. Meunier (eds.) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1. Louvain-la-Neuve: Presses universitaires de Louvain, 335-344. ISBN 978-2-87558-199-0.

O'Donnell, M. 2008. "The UAM CorpusTool: Software for corpus annotation and exploration". In Bretones Callejas, Carmen M. et al. (eds) Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Almería: Universidad de Almería. Pp. 1433-1447.

Rollinson, Paul & Mendikoetxea, Amaya 2010. Learner Corpora and Second Language Acquisition: Introducing WriCLE. In Bueno Alonso, Jorge L. et al. (eds) Analizar datos: Describir variación / Analysing Data: Describing Variation. Vigo: Universidad de Vigo, 1–12.

UCLES (2001). Quick Placement Test (Paper and pencil version). Oxford: Oxford University Press.

Aachen Corpus of Academic Writing (ACAW): A Multilingual Corpus of First and Second Language Writing

Elma Kerz, Marcus Stroebel
RWTH Aachen University
elma.kerz@ifaar.rwth-aachen.de, marcus.stroebel@rwth-aachen.de

The present paper introduces the Aachen Corpus of Academic Writing (ACAW), a multilingual learner corpus of first (German) and second language (English) academic writing. Each student who has contributed to the compilation of this corpus has submitted one L1 sample of academic writing and one L2 sample of academic writing produced at a comparable point in time, i.e. either the undergraduate or at the graduate level. Moreover, meta-data about learner variables were gathered through a self-report questionnaire, including age, gender, knowledge of other foreign language, reading exposure and time spend in English-speaking country. All students contributing to ACAW can be classified as either having a Common European Framework (CEF) English proficiency level of upper intermediate (CEF=B2) or lower advanced (CEF = C1) based on their institutional status (cf. also Callies 2009, p. 116f). Several key design features of ACAW can contribute to widening the scope of Learner Corpus Research. One of its key features is that it consists of paired L1-L2 texts allowing studies based on a within-subject design. Such studies can contribute to the investigation of transfer effects not only at the group but also at an individual level (see, Jarvis, 2000). Moreover, such studies can explain the role of native language proficiency on second language proficiency (cf. Stroebel, Kerz, & Wiechmann, 2017). The insights gained by these studies would have strong implications for the assessment of proficiency in a second language (cf. Hulstijn, 2015). Another key feature – also relevant for learner corpus studies based on within-subject design is the average text length of L1 and L2 writing samples. While the vast majority of available learner corpora typically contain relatively short (500-1000 words) texts written by higher intermediate to advanced learners of English, ACAW contains longer stretches of text per learner (mean length L2 = 5,084 words, SD = 2,019; mean length L1 = 4,650 words, SD = 1,695), making the assessment of L2 learner proficiency at the individual level statistically more robust (cf. Stroebel, Kerz, Wiechmann, & Neumann, 2016, on the automatic assessment of grammatical and lexical complexity in L2 writing using a sliding-window technique which makes it possible to generate several measurements per individual text). Finally, the inclusion of both undergraduate and graduate writing can contribute to research on the development of both L1 and L2 academic writing (cf. also a recent paper by Penris & Verspoor, 2017). In its current form, ACAW consists of 80 texts of L1 academic writing and L2 academic writing. All data are annotated using the components of the Stanford CoreNLP toolkit (Manning et al., 2014): (1) Tokenization (TokenizerAnnotator), (2) Sentence splitting (WordToSentenceAnnotator), (3) Part-of-Speech (POS) tagging (PostTaggerAnnotator), (4) Lemmatization (MorphAnnotator) and (5) Syntactic Parsing (ParserAnnotator). The ACAW L2 English component has been parsed with the PCFG Parser (Klein & Manning, 2003) and its L1 German component has been parsed with its adaptation for German (Rafferty & Manning, 2008).

References:

Callies, M. (2009). Information Highlighting in Advanced Learner English: The Syntax-Pragmatics Interface in Second Language Acquisition. Amsterdam: Benjamins.

Hulstijn, J. H. (2015). Language Proficiency in Native and Non-Native Speakers: Theory and Research. Amsterdam: John Benjamins.

Klein, D. & Manning, C. (2003). Accurate unlexicalized parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics (pp. 423-430).

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

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D. (2014), The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics. Baltimore, Maryland (pp. 55-60).

Penris, W. & Verspoor, M. (2017). Academic writing development: a complex, dynamic process. In Pfenniger, S. & Navracsisc, J. (Eds). Future Directions for Applied Linguistics. Bristol: Multilingual Matters Ltd (pp. 215-242).

Rafferty, A. & Manning, C. (2008). Parsing three German Treebanks: Lexicalized and unlexicalized baselines. Proceedings of the ACL-08: HLT Workshop on Parsing German. Columbus, Ohio: Association for Computational Linguistics (pp. 40-46).

Stroebel, M., Kerz, E., Wiechmann, D., & Neumann, S. (2016). CoCoGen-Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-Window Technique. CL4LC 2016, 23.

Stroebel, M., Kerz, E., Wiechmann, D. (2017). First Language Proficiency Predicts Second Language Proficiency: An Investigation of Linguistic Complexity in L1 and L2 Academic Writing. Paper presented at LCR 2017, Bozen, Oct. 5-7.

Approaches to automated English essay evaluation in Russian students' learner corpus[1]

Olga Lyashevskaya[1, 2], Olga Vinogradova[1],
School of Linguistics, Higher School of Economics, Moscow[1], Vinogradov Institute of the Russian Language RAS, Moscow[2]
olesar@yandex.ru, olgavinogr@gmail.com

Corpus linguistics community is well familiar with reviews of the suggestions and descriptions of the computing tools applied to working with language corpora (Crossley et el. 2011, Graesser et al. 2011, Graesser & McNamara 2012, Roscoe et al. 2012). Reports on achievements in applying computing instruments to learner corpora have not lagged behind – starting with Biber 1999 to Landauer 2007, Folse 2011, Sawaki et al. 2013, Lozano & Mendikoetxea 2013, Pickering & Garrod 2013, Lavallée & McDonough 2015, and many others. The paper presents some new tools that have been developed for the needs of a specific learner corpus and discusses the benefits they bring to both learners of English and EFL instructors.

REALEC (Vinogradova, 2016) is the first in the open access collection of English texts (mainly essays) written by students with Russian as their native language who are learning English at the university. At the end of their 2nd year in the Bachelor programme all students take a final EFL examination which assesses, among other skills, their writing proficiency. REALEC stands for Russian Error-Annotated Learner English Corpus, and examination essays can be seen at http://realec.org/hse/#/data_4_staff/IELTS/IELTS2015/. Errors in English essays are annotated in REALEC by experts (EFL instructors, as a rule). REALEC can be considered a new type of learner corpus as in addition to pointing out for students the mistakes in their essays, the corpus in fact serves the purpose of preliminary evaluation of student writing. Besides, it also presents material for linguistic observations. The project team working with the corpus over the last two years have been developing computational tools to make the use of REALEC efficient for both students and their English instructors in preparation for the university EFL examination. This paper considers four tools designed to enhance corpus-mediated work in the classroom:

- easy access to the statistics of student errors in one text, in all texts written by the same author, or in all texts in a current folder, which provides for on-the-spot feedback on the quality of the text uploaded to the corpus;
- automated evaluation of lexical proficiency (Vinogradova et al. 2017), which includes commonly used features such as length of words; length of sentences; distribution of words across the Common European Framework scale levels (A1-C2); use of academic vocabulary compared with one of the two lists - the

1 The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 — 2017(grant №16-05-0057) and the Russian Academic Excellence Project «5-100».

Coxhead Academic Word List and in the Corpus of Contemporary American English; number of repetitions; use of linking words; use of collocations (as attested by the comparison with the Pearson academic collocation list);

- automated test-maker, which extracts sentences from the corpus and turns them into questions for placement and progress testing purposes;
- automated evaluation of syntactic complexity of the text which takes into account features such as mean sentence depth and the average number of relative and adverbial clauses. Dependency parsing is performed using UDpipe (Straka et al. 2016).

The last two of these tools are in the focus of this paper.

The test-making tool for REALEC was first introduced in 2016 (Kustova 2016). Since then, the convenience for both EFL instructors designing tests and for students taking those tests has been significantly improved. The interface was set up in Moodle (Modular Object-Oriented Dynamic Learning Environment), the open source learning platform (https://moodle.org/), which provides a user-friendly environment for preparation and administration of tests and delivers the statistics for students' performance after the test. The tests can be seen at http://webcorpora.net/realec (authorization is necessary). Each test focuses either on one single area (grammar, punctuation, or discourse), or on a few - up to five - different areas. Work with tests can be offered automatically in case a student makes more errors in his/her writing than the threshold number of errors, or it can be administered by the instructor as a part of the curriculum. For the instructor who makes a test, there are options of choosing the types of testing questions, whether or not to differentiate the level of each question's sophistication, and also of the mode of administration of the test and the system of scoring. The variety of options makes it possible for the instructor to provide custom-made tests for any purpose – placement tests, progress tests, diagnostic tests, tests as a form of additional practice, revision tests.

The option to get automated evaluation of the variety of syntactic means used in a student text is an important feature for both instructors and learners. Our analysis of the IELTS treebank suggests that the use of adverbial clauses highly correlates with the grades assigned by the experts: these constructions allow students to produce complex explanations and express a variety of temporal, causal, and other relations for the propositions. The average number of adverbial clauses per essay is 5.41±1.07(CI 95%) for the texts scored 75% and over, which is three times as high as that of the essays scored 30% and lower (1.86±0.5(CI 95%)). Another example is the index of syntactic depth in argumentative essays. Indeed, the more successful learners are more likely to vary the inventory of structures than those students who got the worst grades.

References:

Biber, D., Conrad, S., & Reppen, R. (1999). Corpus Linguistics: Investigating Language Structure and Use. *International Journal of Corpus Linguistics,* 4(1), 185–188.

Crossley, S. A., Salsbury, T., & McNamara, D. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2).

Folse, K. (2011). Applying L2 lexical research findings in ESL teaching. *TESOL Quarterly*, 45(2), 362–369.

Graesser, A. C., & McNamara, D. (2012). Automated analysis of essays and open-ended verbal responses**.** H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, K. J. Sher (Eds*). APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* Washington, DC: American Psychological Association, 307-325.

Graesser, A. C., McNamara, D., & Louwerse, M. M. (2011) Methods of Automated Text Analysis. M. X. Kamil, P. D. Pearson, E. B. Moje, & P. Afflerbach (Eds.). *Handbook of Reading Research 4,* Routledge

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

Kustova, M. (2016). Design of corpus-generated EFL placement and progress pests for university students*. Proceedings of the Conference The Future of Education. libreriauniversitaria.it Edizioni,* 148-149.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lavallée M., & McDonough, K. (2015). Comparing the Lexical Features of EAP Students' Essays by Prompt and Rating. *Revue TESL du Canada 43*, 32 (2), 30-44.

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and Second Language Acquisition: The design and collection of CEDEL2.1. A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.). *Automatic Treatment and Analysis of Learner Corpus Data.* Amsterdam: John Benjamins, 65-100**.**

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension*. Behavioral and Brain Sciences 36,* Cambridge: CUP, 329-392.

Roscoe, R. D., Kugler, D., Crossley, S. A., Weston, J. L., & McNamara, D. S (2012). Developing Pedagogically-Guided Threshold Algorithms for Intelligent Automated Essay Feedback*. Proceedings of FLAIRS-25,* 466-471.

Sawaki, Y., Quinlan, T., & Lee Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task*. Language Assessment Quarterly*, 10(1), 73–95.

Straka, M., Hajic., J., & Strakova, J. (2016). Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Slovenia**.**

Vinogradova, O. (2016). THE ROLE AND APPLICATIONS OF EXPERT ERROR ANNOTATION IN A CORPUS OF ENGLISH LEARNER TEXTS*. Computational Linguistics and Intellectual Technologies* 15(22), 740-751.

Vinogradova, O., Lyashevskaya, O., & Panteleeva, I. (2017). MULTI-LEVEL STUDENT ESSAY FEEDBACK IN A LEARNER CORPUS *Computational Linguistics and Intellectual Technologies* 16, v.1, 382-396

Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers*. English for Specific* Purposes, 28 (1), 33–41.

Using learner corpora and language testing to evaluate relative difficulty of linguistic features

Mick O'Donnell
Universidad Autónoma de Madrid
michael.odonnell@uam

In recent years, there has been increasing attention given to the task of using learner corpora to assign linguistic features to proficiency levels (e.g., Hawkins & Buttery 2010; Hawkins & Filipoviç 2012). However, our studies on learner corpora (within the TREACLE project, Murcia Bielsa & MacDonald 2013) reveal that linguistic features are not clearly critical to any one proficiency level, but are rather acquired gradually over a number of proficiency levels (see: O'Donnell 2013). For this work, two learner corpora were used, the WriCLE corpus (Rollinson and Mendikoetxea, 2010) and the *UPV Learner Corpus* (Andreu *et al*, 201), together consisting of around 700,000 words of text over 1,334 learner productions.

To overcome this problem, recent work by the author (O'Donnell 2013, 2014) changed the objective: rather than assign linguistic features to levels, the goal was to order the features in terms of their average acquisition difficulty. One might not be able to justifiably say that learners acquire feature 1 at B1 and feature 2 at B2, but one can say that, over a large number of learners, feature 1 is usually acquired before feature 2. By ordering all linguistic features relative to each other in this way, one produces a general order of acquisition of the features. For teaching purposes, these features could then be split up into sub-lists, the easiest assigned to A1, the next to A2, and so on.

In this work, I explored the changing levels of use of linguistic features at each proficiency level to determine their relative difficulty: exploring the onset of use of the feature (at which proficiency level a majority of students started the use the feature), and also whether usage rises or falls with increasing proficiency (e.g., falling usage is sometimes explained where a feature is easily transferred but learners then learn alternative strategies to realise the same meaning).

Increasingly though, I have found that the explanation of the changing patterns of usage of a feature (with increasing proficiency) need to be explained by two types of learning: firstly, learning HOW to produce the structure, and secondly, learning WHEN the structure should be appropriately used. For instance, Spanish learners of English learn fairly quickly how to produce the present perfect tense (e.g., *I have seen John)* as they have a similar structure in Spanish. However, within the two languages, the tense is appropriately used in different contexts, and the learner thus can spend years learning when to use this tense. I use the term 'contracting contexts of use' where a learner stops using the structure in contexts which are appropriate in their mother tongue but not in the L2, and 'expanding contexts of use' where a learner starts using the structure in contexts not used in their mother tongue.

Patterns of changing usage over increasing proficiency can involve all three of these factors: learning how to produce the form, learning to use the form in new contexts, and also learning not to use the form in some contexts.

As a result, levels of usage of a linguistic feature by themselves are not that useful to determine when the feature is acquired or its relative acquisition difficulty. One needs more information as to whether or not students at each level have difficulty in forming the structure, and also in which contexts of use the learners are using the structure at each level.

In consequence, this talk will outline a number of studies we have done which use more delicate coding of learner corpora to identify the contexts of use of particular linguistic features (articles, tenses, quantifiers and prepositions) in an attempt to better understand the acquisition patterns of these structures. Additionally, where the learner corpus does not provide sufficient information, the use of targeted correctness tests (asking whether a number of sentences are correct or not) allow us to identify more clearly the patterns of development.

References:

Andreu, M., Astor, A., Boquera, M, MacDonald, P., Montero, B. & Pérez, C. (2010). Analysing EFL Learner Output in the MiLC Project: An Error It's*, but Which Tag?. In M. Campoy, B. Belles-Fortuno, M. Gea-Valtor (Eds.) *Corpus-based Approaches to English Language Teaching*. London: Continuum, 167-179.

Hawkins J. & Buttery, P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. English Profile Journal 1/1, 1–23.

Hawkins, J. and Filipoviç, L. (2012). Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework. Cambridge: Cambridge University Press.

O'Donnell, M. (2013). From Learner Corpora to Curriculum Design: an Empirical Approach to Staging the Teaching of Grammatical Concepts. Proceedings of the V International Conference on Corpus Linguistics (CILC2013). Procedia – Social and Behavioral Sciences, 571–580.

O'Donnell, M. (2015). Using learner corpora to order linguistic structures in terms of apparent difficulty. In E. Castello, K. Ackerley & F. Coccetta (Eds.) *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*. Peter Lang. Pp71-85.

Murcia Bielsa, S. & MacDonald, P. (2013). The TREACLE project: Profiling learner proficiency using error and syntactic analysis. In S. Granger, G. Gilquin & F. Meunier (Eds.) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1. Louvain-la-Neuve: Presses universitaires de Louvain.

Rollinson, P. & Mendikoetxea, A. (2010). Learner Corpora and Second Language Acquisition: Introducing WriCLE. In J.L. Bueno Alonso et al. (Eds.) *Analizar datos: Describir variación/Analysing data: Describing variation*. Vigo: Universidade de Vigo, 1-12.

Watch the puppet: an exploratory corpus of primary-school learner English

John Osborne[1], **Heather Hilton**[2]
Université Savoie Mont Blanc[1], Université de Lyon[2]
john.osborne@univ-smb.fr, heather.hilton@univ-lyon2.fr

The teaching of foreign languages at primary school level has developed considerably in recent years throughout Europe (see Enever, 2011). In France, pupils now receive one and a half hours of language teaching per week from the age of 6 onwards. Although it has been noted that speaking in the early language learning classroom is usually reproductive and imitative (Becker & Roos, 2016), there is still relatively little direct linguistic data available to analyse what actually takes place in these classes. What kind of language is produced by the teacher and by the learners? What interplay is there between L1 and L2? What kinds of interaction take place during a lesson? How is language production linked to actions, manipulation of objects etc? What differences are there between classes of different age groups? The exploratory study reported on here aims to provide data to further the analysis of such questions.

Within an approach that views the emergence of new language as the result of complex interactions between numerous factors, the present phase of the project focuses on external methodological and institutional variables which influence language learning in a six to eight year old primary school population. The aim is therefore to capture as much linguistic and non-linguistic information as possible during language learning sessions in the classroom. The study is based on a small-scale, richly annotated corpus, an approach which was preferred principally because of the aims and exploratory nature of the project, but also because the data collection involved is relatively demanding in terms of time and human resources (school liaison, sound and video recording in a complex environment, transcription and annotation of recordings). The recorded data come from two primary schools in the Seine et Marne area just east of Paris. Two different classes were followed during a school year; the first is a preparatory level [CP] class of 25 six year old learners; the other a second-level elementary [CE2] class of 30 eight year olds. For each class, a week of English lessons (two or three 30-40 minute sessions per week) was recorded at three periods during the school year: in December, in February and in May. The audio and video recordings are transcribed in EXMARaLDA (Schmidt 2013), which uses a musical-score format allowing for multiple tiers of transcription, annotation and description. For the transcription of classroom sessions, three transcription tiers are used: one for the teacher, one for individual learner productions and one for collective learner productions (e.g. choral responses). Annotation tiers were added for interaction type, errors, L1 use and speakers' non-linguistic actions, while four description tiers (i.e. not linked to a specific speaker) respectively recorded phases in the lesson structure, theme, materials used and teacher/learner activity.

The classroom data are complemented by recorded interviews with the two teachers and a series of individual oral reception and production tasks for the learners. These tasks will be used notably for examining the phonological characteristics of learners' L2 production,

since the classroom recording conditions (background noise, a single boom microphone) are not favourable for close analysis of pronunciation.
The presentation will concentrate on features of the classroom corpus. Although both of the classes follow a similar "communicative-active" methodology, involving contextualised presentation of new words and structures, role play, games, use of *realia*, picture cards and puppets, there are interesting differences in the types of interaction, lexical richness, the length of utterances, the frequency of key structures and in the way the teachers contextualise language items and elicit language responses from the learners, with corresponding differences in the language productions of the learners and in their recourse to the L1. It is these contextual differences in language behaviour which will be presented here, with a particular focus on the presentation, practice and production of similar target vocabulary items in thematically comparable lessons.

References:

Becker, C. & Roos, J. (2016). An approach to creative speaking activities in the young learners' classroom. *Education Inquiry*, 7(1), 9-26.

Enever, J. (2011). (Ed.). *ELLiE: Early Language Learning in Europe*. London:British Council.

Schmidt T. (2013). *EXMARaLDA*: *EXtensible MARkup Language for Discourse Annotation.* University of Hamburg Zentrum für Sprachkorpora: http://www.exmaralda.org

Effects of input on written proficiency in L2 English and Dutch: CLIL and non-CLIL learners in French-speaking Belgium

Luk Van Mensel[1], Amélie Bulon[2], Isa Hendrikx[2], Fanny Meunier[2], Kristel Van Goethem[2]
Université de Namur[1], Université catholique de Louvain[2]
luk.vanmensel@unamur.be, amelie.bulon@uclouvain.be, isa.hendrikx@uclouvain.be, fanny.meunier@uclouvain.be, kristel.vangoethem@uclouvain.be

The present study falls within the framework of an interdisciplinary project on Content and Language Integrated Learning (CLIL) in French-speaking Belgium (Hiligsmann et al, forthc.). It explores the effects of input on written proficiency in immersive (CLIL) and non-immersive (non-CLIL) language learning settings using various input measures.
Several studies (e.g. Zydatiß 2007; Lorenzo & Moore 2010; Jexenflicker & Dalton-Puffer 2010; Gené-Gil, Juan Garau & Salazar-Noguera 2015; Martínez 2015; Bulon, Hendrikx, Meunier & Van Goethem forthcoming) compare the language proficiency of learners in CLIL and non-CLIL settings using global measures of complexity, accuracy and/or fluency, typically referred to as CAF (Housen, Kuiken & Vedder 2012; Norris & Ortega 2009). However, few studies on CLIL have controlled for the possible effect of L2 exposure (Saladrigues & Llanes 2014), while time spent on learning a second/foreign language has been recognized as one of the most important factors for successful acquisition. The amount and quality of target language (TL) input has become a focus of interest for many SLA researchers (e.g. Kinsella 2009 and Moyer 2009 on formal and informal contact; Llanes & Muñoz 2009 and Pérez-Vidal & Juan-Garau 2011 on study abroad; Long 1983 on formal instruction; Johnstone 2007 on length of instruction).
The present study seeks to investigate 1) the written proficiency of CLIL and non-CLIL secondary school pupils and 2) the contribution/impact of various input measures to explain potential differences in proficiency between immersive and non-immersive pupils. Since CLIL programs aim to provide more target-like and input-rich environments than non-CLIL programs, we can expect a more native-like acquisition of the target language, i.e. a more native-like level of syntactic complexity and lexical diversity in the writing of CLIL pupils.
The analysis is based on a corpus of written data in the form of 378 e-mails on similar topics. The participants are 5th year French-speaking secondary school pupils in CLIL and non-CLIL settings, learning Dutch (CLIL n=124; non-CLIL n=85) or English (CLIL n=86; non-CLIL n=83) as a foreign language. Besides the effect of CLIL education, the input measures used to investigate the potential impact of L2 input on the learners' written proficiency are based on Muñoz's work (2011; 2014) and are derived from extensive questionnaires: 1) length of target language (TL) exposure in years and 2) current informal contact with TL, a composite measure consisting of frequency of internet use in the TL, frequency of TL (productive and receptive) use outside school and frequency of contact with native speakers outside school. The texts were analyzed in terms of morpho-syntactic complexity

and lexical diversity using different software programs. In addition, the pupils' IQ (Raven test-score) is included in our analysis to detangle the effect of this cognitive variable from the impact of CLIL and the input measures.

A comparison of the complexity scores found for the texts written by CLIL and non-CLIL pupils showed that the Dutch texts written by immersive pupils were more complex for all measures but type-token ration (TTR), a measure of lexical diversity. As for the English texts, the immersive pupils wrote more words, more sentences and obtained higher MTLD scores (measures for textual lexical diversity), while non-immersive pupils wrote longer words. So, while there is a more clear-cut difference in written proficiency between CLIL and non-CLIL learners of Dutch, this is less the case for learners of English. At the time of writing the present abstract, the various analyses involving the impact of the input measures are being performed but preliminary results indicate that for Dutch, CLIL remains a significant predictor for all the complexity measures selected. Also the pupils' IQ shows a significant relationship with lexical diversity and morphological and syntactic complexity. This is probably due to our sample; we found greater differences in IQ between the CLIL/non-CLIL learners in Dutch compared to English, which can also be related to a possible selection bias (see for instance Thomas & Collier (2002)). Furthermore, current informal contact with the target language is significantly related to lexical diversity and text length (in number of words).

For English we found that CLIL and the number of years spent learning the target language are significant predictors for morpho-syntactic and lexical complexity. However, current informal TL contact does not appear to have an impact on the pupils' writing complexity in English. A potential explanation is that since English is an international – and omnipresent – language, both CLIL and non-CLIL learners of English have a frequent contact with the language, in consequence this variable is less distinctive. Dutch, on the other hand, as a non-international language, is probably less frequently used in informal contexts (e.g. on the internet), therefore the non-CLIL learners are exposed less to Dutch, and CLIL has a greater impact on the acquisition of L2 Dutch.

References:

Bulon, A., Hendrikx, I., Meunier, F. & Van Goethem, K. (forthcoming). Using global complexity measures to assess second language proficiency: Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. Proceedings from the 2016 Linguists' Day of the Linguistic Society of Belgium. In *Travaux du CBL*.

Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Development of EFL writing over three years in secondary education: CLIL and non-CLIL settings. *The Language Learning Journal*, *43*(3), 286-303.

Hiligsmann, P., Van Mensel, L., Galand, B., Mettewie, L., Meunier, F., Szmalec, A., Van Goethem, K.,Bulon, A., De Smet, A., Hendrikx, I., & Simonis M. (forthcoming). Assessing Content and Language Integrated Learning in French-speaking Belgium: linguistic, cognitive and educational perspectives. *Cahiers du GIRSEF.*

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency*:* Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. John Benjamins, Amsterdam/Philadelphia, 1-20.

Jexenflicker, S., & Dalton-Puffer, C. (2010). The CLIL differential: Comparing the writing of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms*. John Benjamins, Amsterdam, 169-190.

Johnstone, R. (2007). Nationally-sponsored innovations at school in Scotland: issues of evidence, generalizability and sustainability. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 111-128.

Kinsella, C. (2009). *An investigation into the proficiency of successful late learners of French.* Unpublished doctoral thesis, Trinity College, Dublin.

Llanes, A. & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System*, *37*(3), 353-365.

Long, M. H. (1983). Does second language instruction make a distinction? A review of research. *TESOL quarterly*, *17*(3), 359-382.

Lorenzo, F., & Moore, P. (2010). On the natural emergence of language structures in CLIL: Towards a theory of European educational bilingualism. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms*. John Benjamins, Amsterdam, 23-38.

Martínez, A. C. L. (2015). Analysis of the Written Competence of Secondary Education Students in Bilingual and Non-Bilingual Programmes. In *Conference proceedings. ICT for language learning*. libreriauniversitaria.it edizioni, 499-503.

Moyer, A. (2009). Input as a critical means to an end: Quantity and quality of experience in L2 phonological attainment. In T. Piske & M. Young-Scholten (Eds.), *Input Matters in SLA*, Multilingual Matters, Bristol, 159–174.

Muñoz, C. (2011). Input and long-term effects of starting age in foreign language learning. *IRAL-International Review of Applied Linguistics in Language Teaching*, 49 (2), 113-133.

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, 35 (4), 463-482.

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.

Pérez-Vidal, C. & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A Study Abroad perspective. *VIAL*, *Vigo international journal of applied linguistics*, *4*, 157-185.

Saladrigues, G. & Llanes, À. (2014). Examining the impact of amount of exposure on L2 development with CLIL and non-CLIL teenage students. *Sintagma*, *26*, 133-147.

Thomas, W.P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Santa Cruz, CA: University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.

Zydatiß, W. (2007). *Deutsch-Englische Züge in Berlin (DEZIBEL). Eine Evaluation des bilingualen Sachfachunterrichts in Gymnasien: Kontext, Kompetenzen, Konsequenzen.* Peter Lang, Frankfurt am Main.

Software Demonstration

#LancsBox: A new corpus tool for the study of learner language

Vaclav Brezina , Dana Gablasova
Lancaster University
v.brezina@lancaster.ac.uk, d.gablasova@lancaster.ac.uk

Corpus linguistics has become a major methodological approach in a number of areas including language learning research. There are a variety of software that can be used by researchers and teachers for a corpus analysis of language produced by language learners or the target users of the language. In this software demonstration, we introduce #LancsBox, a new desktop software package for the analysis of language data and corpora, which was developed at Lancaster University.

In recent years, researchers have been critically re-evaluating the existing procedures in studies that use corpus methods and have proposed more rigorous approaches to data analysis (e.g. Kilgarriff 2005, 2012; Gries 2006, 2013; Lijffijt et al. 2014; Brezina & Meyerhoff 2014; Brezina et al. 2015; Gablasova et al. 2017). Thus combining statistical sophistication and accessibility is the main challenge that needs to be met by corpus linguistics software tools; the tools should encourage a multi-dimensional view of data, easy comparison, and effective visualization. #LancsBox and its features have been developed as a response to the recent debate in the field of corpus-based studies on the nature of corpus tools. The software tool incorporates a number of existing analytical techniques and adds new innovative methods that enable more efficient and sophisticated exploration of the data. #LancsBox takes plain text or XML file input and processes data automatically adding part-of-speech annotation using the Tree-tagger. It also seeks to provide sophisticated statistical analyses, while enabling the user to navigate through large amounts of data with ease. As a result it can be use by researchers interested in studying learner language as well as by teachers or students themselves in the classroom. It is free to use for non-commercial purposes and works with any major operating system.

In particular, #LancsBox:

- Searches, sorts and filters examples of language use.
- Compares frequency of words and phrases in multiple corpora and subcorpora.
- Identifies and visualises collocations.
- Uses a simple but powerful interface.
- Supports a number of advanced features such as customisable statistical measures.

Given the increasing importance of collocations in the studies of learner language, particular attention will be paid to a particular module within #LancsBox – GraphColl. GraphColl (Brezina et al. 2015) is a tool that can be used to identify collocations using all available association measures (AMs). Moreover, it provides visualisations of the collocational relationships between words. GraphColl (that stands for 'graphical collocations') was developed and included in #LancsBox with the specific aim of allowing

users to easily apply dozens of different AMs while supporting high transparency and replicability through explicit access to the equation used in each AM. In addition to existing AMs, it allows users to define their own collocational measure.

The main aim of the software demonstration is to provide a practical introduction to #LancsBox, highlighting the innovative features of the new tool and demonstrating their use with real learner data. The data used to demonstrate the tool are taken from the Trinity Lancaster Corpus, a large corpus of spoken L2 English that is being developed at Lancaster University in collaboration with Trinity College London (Gablasova et al, 2015). Currently the corpus contains over 4 million words from interviews with 1900 L2 speakers of English, at three proficiency levels and a variety of linguistic/cultural backgrounds.

References:

Aijmer, K. (ed.). (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Brezina, V., & Meyerhoff, M. (2014). Significant or random. A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1-28.

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.

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task type and speaker style. *Applied Linguistics* (Advance Access).

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Language Learning*.

Gries, S. Th. (2013). *Statistics for linguistics with R: a practical introduction*. Berlin: Walter de Gruyter.

Gries, S. Th. (2006). Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 191-202.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2), 263-276.

Kilgarriff, A. (2012). Getting to know your corpus. In Sojka, P., Horák, A., Kopecek, I. & Pala, K. *Text, Speech and Dialogue* (pp. 3-15). Berlin: Springer.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, advanced access.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Index of Names

http://lcr2017.eurac.edu
lcr2017@eurac.edu