

upon the recent development of academic written English corpora such as the British Academic Written English (BAWE) the number of supporters of the use of corpora in language teaching. The very fact that TaLC edition only confirms that the applied use of corpora in the language classroom is here to stay. is still a long way to go before corpora can be understood and used by language teachers encounter on their first hands-on contact with corpora and proposes a task-based, consciousness-raising app are not corpus linguists) understand the basics of corpora. With a limited number of language teachers using . With a limited number of language teachers using corpora, it comes as no surprise that there do suggest that language teachers who are new to corpora may find it difficult to grasp that corpora corpora may find it difficult to grasp that corpora do not work in the same way as at helping language teachers gauge different types of corpora and discern which ones are best suited to Aston Several proposals for the pedagogic use of corpora have suggested that learners can usefully be involve learners can usefully be involved in constructing the corpora to be used. Such proposals have included: * use area of translator training, where constructing specialised corpora which can be analysed to identify domain-specific proposals we find three main ideas: * that such corpora may be more appropriate for learning contexts than acquire insight into how to use and select corpora appropriately, and the ability where necessary to co of work involved and the probability that such corpora will be less carefully edited and less comprehensive use of subcorpora defined within larger pre-compiled corpora. A version of SARA (originally designed for use gap between the use of standard pre-compiled corpora and that of learner-constructed ones, with benefits poor accessibility of language resources (e.g. text corpora) for French, which is being taught in our suggested that one of the main functions of corpora in language pedagogy is that of being able

TaLC/2

20-23 July 2016

Teaching and Language Corpora Conference

Contents

Plenaries	1
Towards corpus literacy in language teacher education (<i>Marcus Callies</i>) . . .	1
One norm to rule them all? Describing and evaluating learners' usages in learner corpus research (<i>Gaëtanelle Gilquin</i>)	2
Collaborative native-non-native student translation tasks in CMC: A corpus-based study (<i>Barbara Lewandowska-Tomaszczyk</i>)	4
Morphological productivity in spoken and written learner German (<i>Anke Lüdeling</i>)	4
A new look at learner language - the Trinity Lancaster Corpus (<i>Tony McEnery</i>)	6
Full papers	7
Grammatical errors across proficiency levels in L2 spoken and written English (<i>Mariko Abe</i>)	7
Have intermediate level students learned how to use the corpus approach for learning general verbs in the classroom? (<i>Awatif Alruwaili</i>)	9
Using a learner corpus for peer tutor training (<i>Melanie Andresen</i>)	10
Evaluating the effectiveness of prototypical text detection in teaching and research: New developments and applications of <i>ProtAnt</i> (<i>Laurence Anthony, Paul Baker</i>)	12
Complexity and qualitative lexical knowledge – A corpus-based study on the use of <i>take</i> in German learner English (<i>Albert Biel</i>)	13
Stance-taking in advanced spoken L2 English: The effect of speaker role (<i>Vaclav Brezina, Dana Gablasova</i>)	14
A meta-analysis of DDL research 1: Rationale, methodology and outcomes (<i>Alex Boulton, Tom Cobb</i>)	15
All tooled up: Corpus-assisted editing for academic writers (<i>Maggie Charles</i>)	17
A quantitative corpus-based study on English prepositions: Conceptual contiguity and its pedagogical implications (<i>Alvin Cheng-Hsien Chen</i>)	18
What are the suitable corpus data to support L1 teaching (not only) in Czech? (<i>Lucie Chlumská, Anna Čermáková</i>)	20
Modifying corpora authenticity to benefit beginner level EFL students: An update on SCoRE (<i>Kiyomi Chujo, Kathryn Oghigian, Shiro Akasegawa</i>)	22
A meta-analysis of DDL research 2: Variation, good practice and future work (<i>Tom Cobb, Alex Boulton</i>)	23

Cohesion or coesione? L1 Italian learners' use of linking adjuncts in argumentative essays (<i>Meredith D'Arienzo</i>)	25
Dutch and German NN compounds in translation (<i>Hinde De Metsenaere</i>) .	26
Using specialized corpora in the ESP classroom to explore corporate pragmatic strategy (<i>Bridgit C. Fastrich, Christoph Wolk</i>)	28
How inter-annotator agreement helps to improve error annotation schemes in learner corpora (<i>Alena Fenogenova, Elizaveta Kuzmenko, Olga Vinogradova</i>)	30
The research grant proposal genre: corpus-based findings and applications for data-driven learning (<i>Lynne Flowerdew</i>)	34
Hands-on use of corpora by trainee translators (<i>Ana Frankenberg-Garcia</i>) .	36
Disagreement in L2 spoken English: From learner corpus research to corpus-based teaching materials (<i>Dana Gablasova, Vaclav Brezina</i>)	37
Anticipatory <i>it</i> patterns and rhetorical moves in the critique genre family (<i>James Robert Garner</i>)	39
The corpus as a writing tool in first-year composition (<i>James Robert Garner</i>)	41
Language for specific purposes corpora and tailor made concordancer: (Semi-)big data corpora and flexible open source software for writing centres (<i>Tobias Gärtner</i>)	42
Speech rate revisited – the effect of task design on speech rate (<i>Tomáš Gráf</i>)	43
Building a large scale corpus of academic written English in a Brazilian university context: An academic learner corpus (<i>Annallena de Souza Guedes, Bárbara Malveira Orfanò, Jessica Ceritello Alves</i>)	46
Extensive reading, false-beginners, and vocabulary development: Investigating the impact of data-driven learning (<i>Gregory Hadley, Maggie Charles</i>)	47
Using a corpus-based collocation network explorer to detect/correct L2 learners' collocations (<i>Ping-Yu Huang, Nai-Lung Tsao</i>)	48
Get in line and show me what you've got: Exploring concordance line ranking and selection methods for different language learning goals. (<i>Stephen Jeaco</i>)	50
An open educational resource: The SOURCe project (<i>Fryni Kakoyianni-Doa, Eleni Tziafa</i>)	52
Word semantics in terms of suffix combinability: L2 acquisition and a specialized electronic corpus (<i>Stela Manova, Bartosz Brzoza</i>)	53
Teaching language variation in a Spanish L2 class: Analyzing classroom interactions and students' perceptions (<i>Nausica Marcos Miguel</i>) . . .	55
Introducing DDL vocabulary files in an ESP context (<i>Sanja Marinov</i>) . . .	56
Nominal complexity in learner writing (<i>Jillian Nahm</i>)	57
Evaluating learner language complexity: Traditional corpus methodology versus online analysers. A case study using the ICCI. (<i>Pascual Pérez-Paredes, María Belén Díez-Bedmar</i>)	59
Exploring, teaching and learning phraseology with Phrime (<i>Piotr Pezik</i>) .	60
English intonation of advanced learners: A contrastive interlanguage analysis (<i>Karin Puga</i>)	62

Linguistic accessibility (SDH and AD) in the university programs for teaching AVT: the AVLA project (<i>Juan Pedro Rica Peromingo, Ángela Sáenz Herrero</i>)	64
L2 learners' developing knowledge of English verb constructions: Usage-based views and implications for teaching (<i>Ute Römer, Cynthia M. Berger & Nick C. Ellis</i>)	65
The use of smallwords in the speech of German learners of English: a corpus-based study of the factors of instruction and natural exposure (<i>Anna Rosen</i>)	67
The effects of corpus and dictionary use: Error correction in L2 writing (<i>Yoshiho Satake</i>)	69
Which corpus management systems for non-specialists in computer science? The case of French trainee teachers of German. (<i>Eva Schaeffer-Lacroix</i>)	70
The influence of task-based factors on L2 assignment writing – Corpus-based evidence from ESP learner language (<i>Christine S. Sing</i>)	72
Corpus construction and specialist vocabulary learning by Chinese and Finnish EAP students (<i>Simon Smith, Nicole Keng</i>)	74
Investigating Linguistic Characteristics of L2 Writing Development through the Examination for the Certificate of Proficiency in English: A Multi-Dimensional Analysis (<i>Shelley Staples, Xun Yan</i>)	76
Developing a first year composition L2 writing corpus and repository (<i>Aleksandra Swatek, Hadi Banat, Shelley Staples</i>)	78
How learners make sense of corpus data (<i>Kuei-Ju Tsai</i>)	79
Developing spelling with data-driven learning in L1 French (<i>Henry Tyne</i>) .	80
Instructed second language acquisition and longitudinal learner corpus research: The case of lexical and syntactic complexity (<i>Nina Vyatkina, Hagen Hirschmann, Felix Golcher</i>)	81
The EAP teaching implications of a corpus investigation into the semantic patterning of grammatical keywords in undergraduate History and PIR essays (<i>Karin Whiteside</i>)	83
Direct quotation in second language writing: A corpus-based study of intertextuality in academic learner English (<i>Leonie Wiemeyer</i>)	85
Cognitive changes after inductive data-driven language learning (<i>Kiyomi Yatabe, Yasuyo Minagawa</i>)	87
Posters	89
KoLaS – Commented Learner Corpus of Academic German (<i>Melanie Andresen, Dagmar Knorr</i>)	89
Corpora, collocations, classroom kids: Teaching collocations to young learners at primary school (<i>Martina Bredenbröcker</i>)	90
Building the Trinity Lancaster Corpus of L2 spoken production (<i>Vaclav Brezina, Dana Gablasova, Tony McEnery, Elaine Boyd</i>)	92
Guangwai Lancaster Chinese Learner Corpus: Filling the gap in learner corpus research (<i>Vaclav Brezina, Hai Xu</i>)	93

Introducing data-driven learning to PhD students for research writing purposes (<i>John Flowerdew, Meilin Chen</i>)	94
Developing didactic principles for annotation based learning (<i>Julia Hübner, Marc Schutzeichel</i>)	96
Exploring the semantic prosody of attributive verbs in academic texts (<i>Hsueh-chao Marcella Hu & Shu-jung Wu</i>)	97
The DGS-L2C: Creation of a German Sign Language learner corpus at the University of Cologne (<i>Thomas Kaul, Alejandro Oviedo, Reiner Griebel, Leonid Klinner & Monika Krumpfen</i>)	99
Towards a <i>Discipline-Specific Vocabulary Core</i> for postgraduate Law students: reporting on a pilot study (<i>Jenny Kemp</i>)	101
Corpora for engineers writing in a foreign language: Methods and applications for Language for Specific Purposes corpora (<i>Marina Kogan, Tobias Gärtner, Sigrun Schroth-Wiechert</i>)	102
Building a corpus of written academic texts in Portuguese (<i>Tanara Zingano Kuhn, José Pedro Ferreira</i>)	103
CollGram profiles and n-gram frequencies as gauges of phraseological competence in EFL learners at different proficiency levels (<i>Agnieszka Leńko-Szymańska</i>)	104
Developing a culturally annotated corpus for foreign language teaching (<i>Çalışkan Nihal</i>)	106
Translation and Do-it-Yourself Corpora used as pedagogical tools in EAP activities (<i>Paula Tavares Pinto, Anna Luisa Lopes Alves</i>)	108
Just another brick – Reflections from PAC-ToE and ICE-IPAC in the teaching/learning of oral English for non-native learners. (<i>Anne Przewozny-Desriaux, Nadine Herry-Bénit</i>)	109
Novice academic writing in Norwegian and English: a new corpus project (<i>Sylvi Rørvik, Marte Monsen</i>)	112
Investigating the use of cohesive devices by advanced learners of German through contrastive interlanguage analysis (<i>Carola Strobl</i>)	113
Geo- and sociolinguistics of plural acquisition by German pre-schoolers with and without immigration background (<i>Eugen Zaretsky, Benjamin P. Lange</i>)	115

Author Index

Plenaries

Towards corpus literacy in language teacher education

Marcus Callies

University of Bremen

Mukherjee's (2004) survey among language teachers identified a large gap between corpus linguists' enthusiasm about the language-pedagogical potential of corpus resources and tools on the one hand, and the reality of English language teaching in Germany on the other. He advocated a concerted effort to popularise the language-pedagogical potential of corpus linguistics and to train teachers to acquire corpus literacy. Mukherjee emphasized that first and foremost it was the teachers who needed to be trained and convinced of the usefulness of corpus data to face challenges in their own teaching before they could be expected to implement more advanced, learner-autonomous activities. Similarly, Römer's (2011) survey article concluded that the practice of English language teaching seemed to be only marginally affected by the advances of corpus research, and comparatively few teachers and learners know about the availability of useful resources and get their hands on corpus computers or concordances themselves (2011: 206). However, there is by now a still relatively small but increasing number of studies that report on the integration of corpus-linguistic content into curricula for language teachers, and examine and evaluate the process of initiating teachers into corpus literacy (e.g. Farr 2008; Breyer 2009; Heather & Helt 2012; Leńko-Szymańska 2014a, 2014b). These studies report first encouraging results.

Against this background, the aim of this talk is to contribute to research and practice on the integration of corpus linguistics into curricula for foreign language teachers. I will first discuss the concept of corpus literacy, a multicomponential set of complex skills, broadly defined as "the ability to use the technology of corpus linguistics to investigate language and enhance the language development of students" (Heather & Helt 2012: 417). To update Mukherjee's (2004) survey, I will then present some first findings of a very recent large-scale survey among secondary-school teachers of English at schools throughout Germany and report on the integration of corpus-linguistic content into the curriculum for English language teachers at the University of Bremen. The focus will be on a description of a corpus activity on the acquisition of intensifying adverbs that confronts teacher trainees with a learner corpus containing data that are close to their own future teaching reality, i.e. written data collected from young learners ranging from primary to secondary school pupils across several different regions/countries from the *International Corpus of Crosslinguistic Interlanguage* (ICCI; Tono & Díez-Bedmar 2014).

References

- Breyer, Yvone, 2009: "Learning and teaching with corpora: reflections by student teachers". *Computer Assisted Language Learning* 22(2), 153-172.
- 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.
- Heather, Julian & Helt, Marie, 2012: "Evaluating corpus literacy training for pre-service language teachers: Six case studies". *Journal of Technology and Teacher Education* 20(4), 415-440.
- Leńko-Szymańska, Agnieszka, 2014a: "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.
- Leńko-Szymańska, Agnieszka, 2014b: "A teacher-training course on the use of corpora in language education: Perspectives of the students". In: Turula, Anna / Mikołajewska, Beata / Stanulewicz, Danuta (eds.): *Insights into Technology-enhanced Language Pedagogy*. Peter Lang: Frankfurt am Main, 129-144.
- Mukherjee, Joybrato, 2004: "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany". In: Connor, Ulla & Upton, Thomas A. (eds.): *Applied Corpus Linguistics. A Multi-dimensional Perspective*. Rodopi: Amsterdam, 239-250.
- Römer, Ute, 2011: "Corpus research applications in second language teaching". *Annual Review of Applied Linguistics* 31, 205-225.
- Tono, Yukio / Díez-Bedmar, Maria Belen, 2014: "Focus on learner writing at the beginning and intermediate stages. The ICCI corpus". *International Journal of Corpus Linguistics* 19(2), 163-177.

One norm to rule them all? Describing and evaluating learners' usages in learner corpus research

Gaëtanelle Gilquin

Université catholique de Louvain

In his seminal paper on interlanguage, Selinker (1972: 212) refers to "those adults who 'succeed' in learning a second language so that they achieve native-speaker 'competence'". As suggested by this quotation, the native speaker is often used as a model for learners and as a reference against which to evaluate their proficiency (see, e.g., Cook 1997: 38, Cummins and Davison 2007: 8). This was true in the era of Error Analysis (Corder 1981), but this is also the case, to a large extent, in learner corpus research. Thus, Flowerdew (2015: 469) notes that "[a] key facet of learner corpus research is that the learner corpus is usually compared with a native-speaker control corpus". In fact, one of the most popular methodologies in learner corpus research, Contrastive Interlanguage Analysis (CIA, Granger 1996), involves as one of its components a comparison between learner data and native speaker data.

In this talk, I will discuss various issues relating to the question of the norm in learner corpus research. These will include the issue of whether a norm is acceptable at all (see Bley-Vroman's (1983) comparative fallacy) and what the options are to

dispense with a norm. I will also consider whether the norm should necessarily be native (cf. the Quirk-Kachru controversy, as exemplified by Quirk 1990 and Kachru 1991) and what problems a native norm may present. Using authentic examples, I will illustrate that learners who behave differently from native speakers do not necessarily use language incorrectly. As an alternative to a unique, native norm, I will show that a range of norms are available (“reference language varieties” in Granger’s (2015: 17) revised version of the CIA method), but that again some of these norms may be problematic if they are not selected carefully (depending on the learner corpus, the purpose of the comparison, etc.) and handled cautiously. It will be demonstrated that different choices of norms may produce different results and thus lead to different conclusions with respect to learners’ usages (e.g. Chen 2013). Finally, the pedagogical implications of such choices will be examined, with particular emphasis on whether all differences between the learner corpus and the reference corpus should be targeted for teaching intervention.

References

- Bley-Vroman, Robert. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33: 1-17.
- Chen, Meilin. 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.
- Cook, Vivian. 1997. Monolingual bias in second language acquisition research. *Revista Canaria de Estudios Ingleses* 34: 35-50.
- Corder, Stephen Pit. 1981. *Error Analysis and Interlanguage*. Oxford: Oxford University Press.
- Cummins, Jim and Chris Davison. 2007. The global scope and politics of ELT: Critiquing current policies and programs. In Jim Cummins and Chris Davison (eds), *International Handbook of English Language Teaching, Part I*. Dordrecht: Springer, pp. 3-12.
- Flowerdew, Lynne. 2015. Learner corpora and language for academic and specific purposes. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier (eds), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 465-484.
- Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds), *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press, pp. 37-51.
- Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7-24.
- Kachru, Braj B. 1991. Liberation linguistics and the Quirk Concern. *English Today* 7(1): 3-13.
- Quirk, Randolph. 1990. Language varieties and standard language. *English Today* 6(1): 3-10.

Collaborative native-non-native student translation tasks in CMC: A corpus-based study

Barbara Lewandowska-Tomaszczyk

State University of Applied Sciences in Konin; University of Lodz

The present paper focuses on the processes and effects of online collaborative translation tasks in native - non-native student pairs for improving written production and translation skills. The analysis and discussion of the learner data aim to create a corpus-informed translation teaching approach, develop adequate didactic materials, and foster discussion between practitioners and theorists in the field of translator education. The purpose of this study is also to add to the growing body of research on asynchronous computer-mediated communication (CMC), and its potential to promote interactional and cognitive growth of student knowledge and skills in translation tasks by combining the processes of CMC interaction and conscious metalinguistic reflection.

The study exploits contacts between Polish MA students of English and translation at the University of Lodz and the State University of Applied Sciences in Konin and students of technical and engineering subjects at North Dakota State University within a Trans-Atlantic Pacific Project (TAPP) to study the students' progress in the interaction. The analysis of collected language data (Source Language-Target Language consecutive versions, email exchanges and a final survey) gives information on longer-term translation learning processes and thus helps justify the collaborative model of language/translation competence development.

Morphological productivity in spoken and written learner German

Anke Lüdeling

Humboldt University Berlin

This talk will investigate qualitative and quantitative aspects of morphological productivity in written and spoken learner corpora of German as a Foreign Language (GFL).

Native speakers of a language are able to form new words in order to fill lexical gaps or condense information (see, among many others, Baayen 1992, Plag 1999, Bauer 2001, Lüdeling & Evert 2005). Such productively formed new words follow regular word formation patterns. These can be constrained on many linguistic levels (phonology, argument structure, part of speech, semantics, pragmatics, etc.). Since the constraints are often difficult to analyze and formulate, they cannot be taught categorially. Moreover, morphological productivity is a quantitative phenomenon: A given pattern is more or less likely to form a new word. There are different measures

that deal with different aspects of productivity. Most of them rely on the observation that productive patterns – since they form new words – have rare types. This is modelled by type token distribution in corpora (cf. Baayen 2001, see Zeldes 2012 for a comprehensive overview of productivity measures). Morphological productivity must therefore be acquired primarily through the linguistic input, and learners of a language obviously have less linguistic experience of that language than native speakers. The (very few) previous studies on the acquisition of morphological productivity (Zeldes 2013, Lüdeling, Hirschmann & Shadrova, submitted) suggest that learners understand that there are productive patterns but have problems acquiring the constraints.

I will look at qualitative and quantitative aspects of different word formation patterns in learner corpora produced by advanced learners of GFL in order to understand whether learners are able to see how the productivity of word formation patterns are acquired.

References

- Baayen, R. Harald (1992) Quantitative Aspects of Morphological Productivity. In: *Yearbook of Morphology 1991*. Kluwer, Dordrecht, 109-159
- Baayen, R. Harald (2001) *Word Frequency Distributions*. Kluwer, Dordrecht.
- Bauer, Laurie (2001) *Morphological Productivity*. Cambridge University Press, Cambridge.
- Lüdeling, Anke & Evert, Stefan (2005) Productive Non-Medical *-itis*. Corpus Evidence and Qualitative Analysis. In: Stephan Kepser & Marga Reis (eds.) *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Mouton de Gruyter, Berlin, 351–370.
- Lüdeling, Anke; Hirschmann, Hagen & Shadrova, Anna (submitted) Linguistic Models, Acquisition Theories, and Learner Corpora: Morphological Productivity in SLA Research Exemplified by Complex Verbs in German
- Plag, Ingo (1999) *Morphological Productivity. Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.
- Zeldes, Amir (2012) *Productivity in Argument Selection. From Morphology to Syntax*. De Gruyter Mouton, Berlin.
- Zeldes, Amir (2013) Komposition als Konstruktionsnetzwerk im fortgeschrittenen L2-Deutsch In: *Zeitschrift für germanistische Linguistik* 41(2), 240-276.
-

A new look at learner language - the Trinity Lancaster Corpus

Tony McEnery
Lancaster University

Does cultural and linguistic background affect learner speech - and if so how? What impact may age have on learner production? Is gender a linguistically important feature when exploring the speech of learners of English? How does learner language production vary by task type? Is learner language different when a learner is leading an interaction as opposed to being led through an interaction by a person who is proficient in the language?

Questions such as these have been addressed regularly in the literature on learner language. However, until recently it was difficult to explore these questions in learner speech. Using a new, multi-million word, corpus being developed at Lancaster University with Trinity College London we can start to address these issues. By exploiting this large, orthographically transcribed, corpus of learner speech, amply provided with plentiful relevant metadata, we can gain fresh insights into learner speech.

In this talk I will overview the construction of the Trinity Lancaster Corpus, discussing the tasks the speakers engaged in and the range of metadata we have available for those speakers. Following from that I will review some initial findings from the corpus. The findings will use a range of metadata to show how, when considered singly and in groups, that metadata can give us answers to questions such as those outlined.

Full papers

Grammatical errors across proficiency levels in L2 spoken and written English

Mariko Abe

Chuo University

Computer-aided Error Analysis (CEA), advocated by Dagneaux, Denness, and Granger (1998) enabled researchers to examine what problems learners encounter during the process of language learning quantitatively. Researchers come to use learner corpora to investigate how learner language develops from the viewpoint of errors (e.g., Götz, 2015; Granger, 1999; Kaneko, 2004; Thewissen, 2013; Tono, 2013).

In a series of CEA study, Abe (2007a) analysed L2 written production from (a) part-of-speech, (b) error types (misformation, missing, unnecessary), and (c) school year. It described the tendencies for each error type to be related to part-of-speech and how its frequency has changed. Misformation errors, for example, were related to nouns and verbs, and their frequency decreased as one's school year advanced. As another study, Abe (2007b) examined L2 spoken performance, which contains oral proficiency information. It focused on the association of errors with oral proficiency, and found that some errors (e.g., verbal agreement, verbal aspect, and nominal inflection) have the potential to disappear as language learning progresses. Following these studies, Abe (2007c) compared L2 written and spoken production to identify various types of linguistic items that can discriminate learner language between production modes and proficiency groups. She found that verb-related errors (e.g., agreement, aspect) were more likely to be made by novice learners and noun-related errors (e.g., nominal case, nominal vocabulary) by advanced learners in both production modes.

These cross-sectional studies did not differ in terms of targeted linguistic features, but have led to the awareness that to describe learner language in a more unified and comprehensive way. Thus, the present study used L2 written and spoken production data from the same task completed by the same Japanese EFL learners (in total 143). It focused on how production modes affect the accuracy of L2 performance across the proficiency levels. The error rate of linguistic features related to (a) tenses and aspects, (b) agreement and inflection, and (c) verbal and nominal lexical choices were examined. Part-of-speech tags were added using the Brill tagger, and error tags were manually inserted following the error tagging guidelines of the NICT JLE corpus (Izumi, Uchimoto, & Isahara, 2004).

As a result, linguistic features apart from the verbal agreement showed significant differences in accuracy rate between written and spoken modes. The proportion size decreased in the following order: (a) nominal inflection, (b) verbal and nominal lexical choices, and (c) tenses and aspects. The accuracy rates of these items were significantly improved in written mode (except for noun plural *-s* which had a higher accuracy rate in spoken mode). The verbal agreement error is problematic in that its accuracy rate did not improve even in a less time-pressured written mode. However, its accuracy rate in both production modes significantly increased as proficiency rose, which supports the results of Abe (2007b). Consequently, it can be assumed that verbal agreement can be used as a predictor of spoken and written proficiency levels. These detailed examinations of error categories suggested that some errors have common developmental patterns, while others vary considerably across proficiency levels. Additionally, the findings supported the assumption that errors can distinguish learners' linguistic competence and production modes.

References

- Abe, Mariko. 2007a. An overview of part-of-speech error in JEFLL corpus. In Yukio Tono (ed.), *JEFLL corpus, a corpus of 10,000 Japanese EFL learners: Analysing written composition of junior and senior high school student*, 146-158. Tokyo: Shogakkan.
- Abe, Mariko. 2007b. A corpus-based investigation of errors across proficiency levels in L2 spoken production. *JACET Journal* 44. 1-14.
- Abe, Mariko. 2007c. Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics* 49. 117-129.
- Dagneaux, Estelle., Denness, Sharon., & Granger, Sylviane. 1998. Computer-aided error analysis. *System: An International Journal of Educational Technology and Applied Linguistics* 26. 163-174. doi:10.1016/S0346-251X(98)00001-3
- Götz, Sandra. 2015. Tense and aspect errors in spoken learner English: Implications for language testing and assessment. In Marcus Callies & Sandra Götz (eds.), *Learner corpora in language testing and assessment*, 191-216. Amsterdam: Benjamins.
- Granger, Sylviane. 1999. Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In Hilde Hasselgård & Signe Oksefjell (eds.), *Out of Corpora: Linguistic information from computer text corpora*, 191-202. Amsterdam: Rodopi.
- Izumi, Emi., Uchimoto, Kiyotaka., & Isahara, Hitoshi. 2004. *A speaking corpus of 1200 Japanese learners of English*. Tokyo: ALC Press Inc.
- Kaneko, Tomoko. 2004. The use of past tense forms by Japanese learners of English. In Junsaku Nakamura, Nagayuki Inoue, & Tomoji Tabata (eds.), *English corpora under Japanese eyes*, 215-230. Amsterdam: Rodopi.
- Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97. 77-101. doi:10.1111/j.1540-4781.2012.01422.x
- Tono, Yukio. 2013. Criterial feature extraction using parallel learner corpora and

machine learning. In Ana. Díaz-Negrillo, Nicolas Baillier, & Paul Thompson (eds.), *Automatic treatment and analysis of learner corpus data*, 169-203. Amsterdam: John Benjamins.

Have intermediate level students learned how to use the corpus approach for learning general verbs in the classroom?

Awatif Alruwaili

University of Nottingham

Corpus-based learning has been the subject of increased interest in recent years, in terms of its application in language education to facilitate learners' language acquisition (Bernardini, 2004). However, despite developments in corpus linguistics (CL) and the potential for applying corpus-based learning in language pedagogy, Römer (2006:121) declares that it still has not 'fully arrived'. Moreover, regardless of the wide recognition of corpus benefits, there is still, as she points out, 'hesitation' about using corpus approaches in teaching. Although there is growing interest in the use of corpora in the classroom, relatively little attention has been paid to investigating learners' actual use of corpora in learning collocations.

The main aim of this study is to examine the use of a corpus approach in the classroom through a case study of general verbs (GVs). I have first operationalised GV concepts in the classroom by designing an approach for selecting targeted items. This multi-level approach was designed with three main concerns in mind: frequency, phraseology and pedagogy. I have set up specific criteria for each level, in a manner that is suitable with regard to the context of the study (English as foreign language/EFL). The six GV's used in the study are *make, have, do, take, give* and *get*.

This study is longitudinal in nature, and it was thus necessary to modify and build on existing approaches and procedures for teaching collocations in the classroom, so as to best suit the context of the study.

The study runs for five weeks, including three training sessions and two testing sessions. The participants are 66 low-intermediate students who enrolled in the preparatory year. The first session is introductory, concerning corpus linguistics and corpus analysis tools. It also addresses steps and procedures for using a corpus analysis tool (AntConc). The second and third sessions are training sessions on how to use a concordancer to identify patterns in GV's. Sinclair's model (2003) is used for reading and analysing concordance lines. The fourth and fifth sessions test the effectiveness of using the corpus approach to teach GV patterns. A software tracker is used to track participants' actual uses and behaviours during the sessions. The study addresses two main questions:

1. How can GV's be defined and identified within a corpus approach as an operationalisable concept in the classroom?
2. How can GV's be taught as part of a corpus approach in an EFL classroom?
 - (a) Is training in using a corpus approach successful? If so, how?

- (b) Is there a link between the participants' use of a corpus approach and the successful completion of the tasks? If so, how?

The preliminary results show improvement in the participants' performance in terms of the time spent in investigating GVs in the training and the testing sessions. This improvement implies that the participants performed the analysis faster in the testing sessions.

References

- Bernardini, S. (2004) Corpora in the classroom: an overview and some reflections on future developments. In: Sinclair, J.(ed.) *How to use corpora in language teaching*. Amsterdam: John Benjamins Publishing company, pp.15-38.
- Sinclair, J. (2003) Reading concordances: An introduction. Harlow: Pearson Longman.

Using a learner corpus for peer tutor training

Melanie Andresen
University of Hamburg

In this talk, we present a corpus-based lesson on the use of the pronoun 'I' in academic writing that was part of the peer tutor training at the *Multilingualism Writing Centre* ('Schreibwerkstatt Mehrsprachigkeit', Knorr & Neumann 2014) at the University of Hamburg. Students are often insecure whether or not to use 'I' in academic writing (e. g. Honegger & Sieber 2012:39). As this question is also highly recurrent at the writing centre, we aim at providing our peer tutors with comprehensive knowledge regarding self-reference. For this purpose, we used data-driven learning (Johns 1991) and had our students apply a theoretical model to corpus data.

Steinhoff (2007) distinguishes three types of 'I' (for German texts): the Author-I, which comments on the text, the Researcher-I, which makes claims and expresses the author's position, and the Narrator-I, which gives autobiographic information. The last type occurs in learner texts only and is considered unacademic by expert raters (Steinhoff 2007:23). While this model is quite helpful in explaining when to use 'I', teaching it theoretically neither enables our tutors to apply it to real examples nor to give constructive feedback on problematic use of 'I'.

To enhance their understanding of the appropriate use of 'I', we used a corpus of journal articles from educational science and the newly released learner corpus KoLaS¹ ('Kommentiertes Lernendenkorpus akademisches Schreiben', Commented Learner Corpus of Academic German). KoLaS was compiled at the *Multilingualism Writing Centre* and comprises 453 authentic German texts written by students for assignments.

The structure of the lesson was based on the concept by Thompson & Tribble (2001) and comprised the following steps:

¹www.uhh.de/uk-kolas

1. Repetition of the model for usage of ‘I’ (Steinhoff 2007) by the students with a focus on surface features that enable them to categorize a given sentence.
2. Analysis of expert sentences by applying the model. The students found ambiguities between Researcher- and Author-I and some instances of the non-academic Narrator-I, showing that also expert corpora contain sentences of controversial appropriateness.
3. Next, sentences from the learner corpus KoLaS were analysed in the same way. As expected, several clearly non-academic instances of the Narrator-I were found. Furthermore, in many cases the student writers seem to intend a Researcher-I but make very subjective claims.
4. At the end of the lesson, our students chose one of the sentences they considered inappropriate and wrote a short feedback to the student in which they explained why the use of ‘I’ was not correct and suggested changes.
5. During the following week, the tutors were asked to write a reflective blog entry following our guiding questions. Students stressed that authentic sentences trigger real situations they may encounter in tutoring and thus give them the possibility of exploring what kind of reaction is most helpful.

In summary, we regard data-driven learning in peer tutor training as an excellent means to encourage reflection on models. Moreover, the process of commenting on authentic examples and texts enables the participants of such training to prepare for their role as peer tutors.

References

- Honegger, Monique & Peter Sieber. 2012. Schreibkompetenz von mehrsprachigen Lehramtsstudierenden: Die Schulsprache als Knackpunkt. In Dagmar Knorr & Annette Verhein-Jarren (eds.), *Schreiben unter Bedingungen von Mehrsprachigkeit*, 35–49. (Textproduktion und Medium 12). Frankfurt am Main: Lang.
- Johns, Tim. 1991. Should you be persuaded. Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*(4). 1–16.
- Knorr, Dagmar & Ursula Neumann. 2014. Die Schreibwerkstatt Mehrsprachigkeit – (Lehramts-)Studierende mit Migrationshintergrund der Universität Hamburg schreiben. In Dagmar Knorr & Ursula Neumann (eds.), *Mehrsprachige Lehramtsstudierende schreiben: Schreibwerkstätten an deutschen Hochschulen*, 119–144. (FörMig-Edition 10). Münster: Waxmann.
- Steinhoff, Torsten. 2007. Zum ich-Gebrauch in Wissenschaftstexten. *Zeitschrift für germanistische Linguistik* 35(1-2). 1–26.
- Thompson, Paul & Chris Tribble. 2001. Looking at citations: Using Corpora in English for Academic Purposes. *Language Learning and Technology* 5(3). 91–105.

Evaluating the effectiveness of prototypical text detection in teaching and research: New developments and applications of *ProtAnt*

Laurence Anthony¹, Paul Baker²

¹Waseda University, ²Lancaster University

Evaluating and ranking texts in terms of their prototypicality is useful in a wide range of teaching and research applications. For example, teachers of writing classes might rank a set of target texts in terms of prototypicality so they can immediately select good exemplars for use as in-class models. Corpus linguists, on the other hand, can use a prototypicality ranking of their corpus texts as a basis for selecting texts for close reading, such as in critical discourse and genre analysis studies.

To date, many studies in the area of natural language processing (NLP) have focused on clustering or grouping texts, and this leads to possible ways of ranking texts using a distance measure. However, interpreting these rankings can be difficult, especially when the NLP methods use complex learning algorithms. Also, very few of the current algorithms have been packaged as standalone, ready-to-use tools. As a result, many teachers and researchers have to rely solely on subjective judgements when ranking texts. This leads to a criticism that their choices are 'cherry-picked' or biased in some way.

In this presentation, we will first introduce and explain recent advances in the development of our *ProtAnt* corpus analysis tool. *ProtAnt* is an easy-to-use, standalone, freeware software application that allows teachers and researchers to quickly analyze a corpus of texts in terms of their characteristic features and rank the texts in various ways based on their prototypicality. The current version of *ProtAnt* uses a keywords approach to select characteristic features. In the presentation, we will explain this method and also introduce other feature selection procedures that do not rely on an external reference corpus as a basis for comparison. Next, we will present the results of a recent study that evaluates the effectiveness of *ProtAnt* at finding model student essays for use as in-class exemplars and tests its ability to identify potential cases of plagiarism in student writing. Finally, we will discuss potential applications of *ProtAnt* in non-teaching settings, such as literary studies, political and social science studies, and legal discourse studies.

References

- Anthony, L. and Baker, P. (2015). *ProtAnt*: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics*, 20(3), 273-292. DOI: 10.1075/ijcl.20.3.01ant.
-

Complexity and qualitative lexical knowledge – A corpus-based study on the use of *take* in German learner English

Albert Biel

University of Bonn

This paper presents research that investigated differences in the use of the high-frequency verb *take* between German learners of English and American English native speakers. The study aimed to contribute to the relatively small body of research on qualitative lexical knowledge, and considers possible implications for teachers of German learners of English.

Over the course of about 25 years, large corpora have been used to identify patterns in the use of English by non-native speakers and to contrast these with those of native speaker varieties. Inspired by Lewis, who states that “[l]anguage consists of grammaticalised lexis, not lexicalised grammar” (1993: iv), some corpus linguists have explored quantitative aspects of learners' vocabulary, though only a few have explored the topic of vocabulary in combination with grammar. Furthermore, the research in this area has, to date, tended to focus mostly on Asian and European varieties (though not German) of English and their use of high frequency verbs, most prominently *make*.

The research reported in this paper examines 724 instances of *take* and their respective lexical-syntactic structures in the German part of the ICLE and a comparable American English speaker corpus (LOCNESS). The sentences are classified according to a modified framework of complexity of lexical-syntactic structures adopted from Liu and Shaw (2001). The results of this research suggest that there are no significant differences in usage between both groups. The general conception that non-native speakers overuse simple verbs due to restricted vocabulary does not hold true for *take* in the German learner data in this study. This seems to be congruent with what Altenberg and Granger's study (2001) revealed about Swedish learners of English. The study further suggests that in both corpora, the frequency decreases with higher complexity. In other words, the verb in question is being used more frequently in easier constructions than in complex ones.

Despite the fact that the German learners make some mistakes when the structures become more complex, they still use these structures as frequently as native speakers of American English. As previous studies have shown, the German learners' use of English seems to be very similar to that of native speakers. Surprisingly, the study showed that German learners use a broader range of phrasal verb combinations in comparison to the American native speakers. Mistakes in the German learner data tend to be related to complex phrasal verb constructions. Therefore, English teachers in Germany could usefully strengthen their focus on phrasal verbs and complex phrasal constructions. In particular, the results of the study show that prepositions are easily confused by German learners of English, and therefore, that structures like *take a look at* or *take care of* should be taught as whole phrases and not only as individual constituents.

References

- Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics* 22 (2). 173-195.
- Lewis, Michael. 1993. *The lexical approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.
- Liu, Eric & Philip Shaw. 2001. Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *IRAL* 39. 171-194.
-

Stance-taking in advanced spoken L2 English: The effect of speaker role

Vaclav Brezina, Dana Gablasova
Lancaster University

Introduction

Epistemic stance-taking is an important aspect of communicative skills, whether in one's native or non-native language. It plays an essential role in conveying the epistemic perspective of the speaker (i.e. his/her certainty-related evaluation of what is said) as well as in managing and negotiating interpersonal relationships between speakers (Kärkkäinen 2006; Hunston & Thompson 2000). So far, there has been only a limited number of studies that address this issue in second language spoken production (e.g. Aijmer 2004; Fung and Carter 2007; Mortensen 2012). This study therefore aims to contribute to our understanding of this area by exploring how epistemic stance is expressed in the context of a spoken English exam by two groups of speakers – the (exam) candidates (advanced L2 speakers of English) and examiners (L1 speakers of English). The research was guided by the following two questions:

RQ 1: Is there a difference in the number of certainty and uncertainty adverbial markers used by the two groups of speakers across different tasks?

RQ 2: Is there a difference in the functions of the certainty markers used by the two groups of speakers?

Method

Corpus. The data were taken from a new, growing corpus of L2 spoken production - the Trinity Lancaster Corpus (TLC). The corpus is based on examinations of spoken English conducted by the Trinity College London, a major international examination board, and contains interactions between exam candidates (L2 speakers of English) and examiners. In this study, we used the advanced subsection of the TLC containing approximately 0.45 million words from 132 L2 speakers (exam candidates) and 66 L1 speakers (examiners). The L2 speakers in this subcorpus are advanced users of English, their proficiency corresponding to C1 and C2 levels of CEFR. Speech from

each candidate was elicited in four speaking situations – one monologic and three highly interactive tasks.

Procedure. We combined automatic corpus searches with manual analysis of the data.

Results & Conclusion

RQ1: In all tasks the candidates used on average more uncertainty markers than the examiners with the difference being statistically significant in all cases. No statistically significant differences were found between the two groups with respect to certainty.

RQ2: The adverbial certainty markers were classified according to the type of certainty they expressed (i.e. subjective or intersubjective). The results showed both differences and similarities in the use of adverbial certainty markers between examiners and candidates. The expression of certainty appeared strongly linked to the speaker role that the exam candidates (L2 users) had in a particular task.

These findings show that there is no clear-cut difference between how L1 and advanced L2 speakers express certainty; rather, L2 speakers modify their epistemic stance-taking according to the interactional setting and their speaker role. These findings show that when studying L2 spoken production it is important to go beyond characterising the interlocutors as ‘native’ or ‘non-native’ speakers of a language. Whereas the fact of being a ‘native user’ or a ‘non-native user’ can indeed be part of the speaker role and speaker identity, there are other equally important factors that arise from the context of the exchange.

A meta-analysis of DDL research 1: Rationale, methodology and outcomes

Alex Boulton¹, Tom Cobb²

¹Atilf, CNRS & Université de Lorraine, ² Université du Québec à Montréal

Please note that this paper is the first of two submitted conjointly. See Cobb & Boulton on page 23.

Corpora have a long history of use as a learning aid and reference resource for foreign and second language (L2) learners, allegedly going back to the 1960s (McEnery & Wilson 1997). The first research papers date from the early 1980s (McKay 1980), but the concept is largely associated with Johns (1990) who coined the term ‘data-driven learning’ (DDL). Since then, it has generated considerable research interest; the question now is how to make sense of it all. The most common form of synthesis is the literature review, which relies on expert judgement to identify and interpret previous research but often fails to an extent on both counts. A more rigorous attempt can be conducted in the form of a meta-analysis (cf. Norris & Ortega 2000), which aims to (a) collect all relevant work in a given field, and (b) interpret results in the same impartial way for all studies. This effect size (Cohen’s *d*) avoids a number of the disadvantages of null hypothesis significance testing (see Plonsky & Oswald 2014), providing a picture of ‘what works’ (and what doesn’t).

Having aggregated the data in a field as a whole, the same measures can then be used to break the data down to analyse specific questions within those studies; in other words, to go from what works and how much, to isolating moderator variables that account for the variation found.

Following a preliminary meta-analysis outlined at TaLC in Lancaster in 2014, our research is accordingly based on three main research questions: (1) How much DDL research is there? (2) How effective/efficient is DDL? (3) How can we best account for any variation observed? The presentation is planned in two parts. The present paper will look at the first two questions, covering the rationale (and limitations), methodology and overall effects obtained. The second paper provides greater granularity in its analysis and discussion of the moderator variables, but also critiques research methodology to date supplemented with suggestions for good practice, and outlines areas in need of future work for a more coherent research programme.

- RQ1. How much DDL research is there? Extensive and principled trawls from a variety of databases up to June 2014 located 205 individual publications (journal articles, book chapters, PhDs and other published texts) that furnish some kind of evaluation of DDL as here defined. Inclusion criteria reduced this to 64 separate studies and 88 unique samples. These were coded by the two researchers, who then extracted the data necessary for the meta-analysis itself (minimally N, M and SD).
- RQ2. How effective/efficient is DDL? Cohen's unbiased d (i.e. weighted for sample size) was calculated for each study: the within-groups (pre/post-test design) effect size is 1.50 ($k=71$; $SD=.91$; 95% $CI=1.28$ to 1.71); the between-groups (control/experimental design) effect size is .95 ($k=50$; $SD=.99$; 95% $CI=.64$ to 1.22). Plonsky and Oswald (2014) provide empirically-based criteria for interpreting these against 91 other meta-analyses in SLA: 1.50 is well above their 1.4 benchmark for a 'large' effect for within-groups designs; .95 is only just below their benchmark of 1.0 for a 'large' effect for between-groups designs. This suggests that DDL as a whole has a great deal to offer language learners.
- RQ3 is pursued in the sequel to this paper.

References

- Johns, Tim. 1990. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria* 10.. 14-34.
- McEnery, Tony & Andrew Wilson. 1997. Teaching and language corpora. *ReCALL* 9(1). 5-14.
- McKay, Sandra. 1980. Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly* 14(1). 17-26.
- Norris, John F. & Lourdes Ortega. 2000. Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50(3). 417-528.
- Plonsky, Luke & Frederick L. Oswald. 2014. How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning* 64(4). 878-912.
-

All tooled up: Corpus-assisted editing for academic writers

Maggie Charles
Oxford University

This paper reports on a course which aims to teach students how to use corpus tools for editing their texts. Although much has been written about the benefits of data-driven learning, particularly for academic writing (for a review, see Yoon, 2011), less attention has been paid to the potential of individual corpus tools for addressing specific editing concerns. The present course is targeted at doctoral students who have already completed part of their thesis in draft form. After an initial session introducing corpus work, students built two corpora: 1) research articles (RAs) in their own field; 2) their own writing. The freeware *AntFileConverter* (Anthony, 2014) was used for converting batches of pdf files to plain text format to build the RA corpus and *AntConc* (Anthony, 2015) for editing purposes. Class sessions provided demonstrations of how specific tools can be used for editing, followed by individual practice in which students used the tool to edit their own writing.

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, unsurprisingly, the most highly rated tool was Concordance with 95% of responses. This was followed by Clusters (82%), Collocates and Keyword List (both 74%), N-grams (70%), Context Searching (67%), Concordance Plot (63%) and Word List (59%).

While the utility of concordancers and other tools that show collocations has been discussed in the literature (e.g. Flowerdew, 2015), I argue that tools such as Keyword List, Concordance Plot and N-Grams have affordances that are particularly relevant to students who are editing texts. 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 their RA corpus, thereby revealing differences in phraseology. Issues concerning the content of the text can be addressed using both Keyword List and Concordance Plot. 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 course and the affordances of the corpus tools for editing, illustrating the findings with examples of student searches.

References

Anthony, Laurence. 2014. *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/>

- Anthony, Laurence. 2015. AntFileConverter (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net/>
- Flowerdew, Lynne. 2015. Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes* 20. 58–68.
- Yoon, Choongil. 2011. Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes* 10. 130–139.
-

A quantitative corpus-based study on English prepositions: Conceptual contiguity and its pedagogical implications

Alvin Cheng-Hsien Chen

National Changhua University of Education

English prepositions have long been the culprit of learning difficulties for EFL/ESL learners. The present study aims to adopt a quantitative corpus-based method to analyze the conceptual similarity among different prepositions in English and provide a conceptual map where prepositions may be argued to form smaller clusters that are conceptually contiguous. We believe that the emergence of preposition clusters may be utilized to reduce the L2 learners' cognitive loading in prepositions learning, and to shed light on the semantic primitives for space conceptualization in general. In the present study, high-frequency non-projective English prepositions were automatically extracted from the British National Corpus, using regular expressions and the parts-of-speech tags. For each instance of prepositions, its trajector NP (TR) and landmark NP (LM) were automatically identified with the state-of-art shallow parser in Python NLTK module (Bird 2006). Our statistical analyses proceeded as follows. First, we adopted covarying collexeme analysis (Gries & Stefanowitsch 2004) to identify significant trajector-landmark (TR-LM) bigrams for each preposition. Each preposition in turn was quantitatively characterized by a set of collostrengths (Stefanowitsch & Gries 2003), indicating their respective association strength to different types of TR-LM covarying collexeme bigrams. After each preposition was mathematically represented as a multidimensional vector of collostrengths, the conceptual similarity or contiguity between each pair of prepositions was quantitatively measured. Our rationale is: if two prepositions tend to take similar sets of TR-LM bigrams in language use, they are considered conceptually more contiguous. This multivariate representation of prepositions was then submitted to a powerful multivariate statistical method—multidimensional scaling (MDS). The objective of using the MDS was to transform the conceptual distance of the prepositions in a multidimensional space (where each TR-LM bigram constitutes one unique dimension) into one in a visually intuitive low-dimensional space. Our results are summarized in Figure 1, where conceptual contiguity between prepositions is visually represented by Euclidean distance on the graph. The implications of the MDS graph are two-fold. On the one hand, the distribution of the prepositions on the 3-dimensional space will unveil the semantic primitives underlying the conceptualization of these spatial relations. These semantic primitives may provide an alternative for EFL teachers to organize English prepositions on a conceptual basis. On the other, the Euclidean distance among the prepositions would suggest a

number of emerging prepositions clusters in English, suggesting their greater degree of conceptual homogeneity (i.e. clusters may include (1) *across-over* (2) *under-below-against-by* (3) *in* (4) *on*, (5) *at*). The constructs for each dimension in Figure 1 as well as the emerging clusters of prepositions will be discussed and evaluated in comparison with the hypothesis of the *containment-support* continuum in typological studies on space (Bowerman & Choi 2001; 2008; Levinson et al. 2003; Vandeloise 2003).

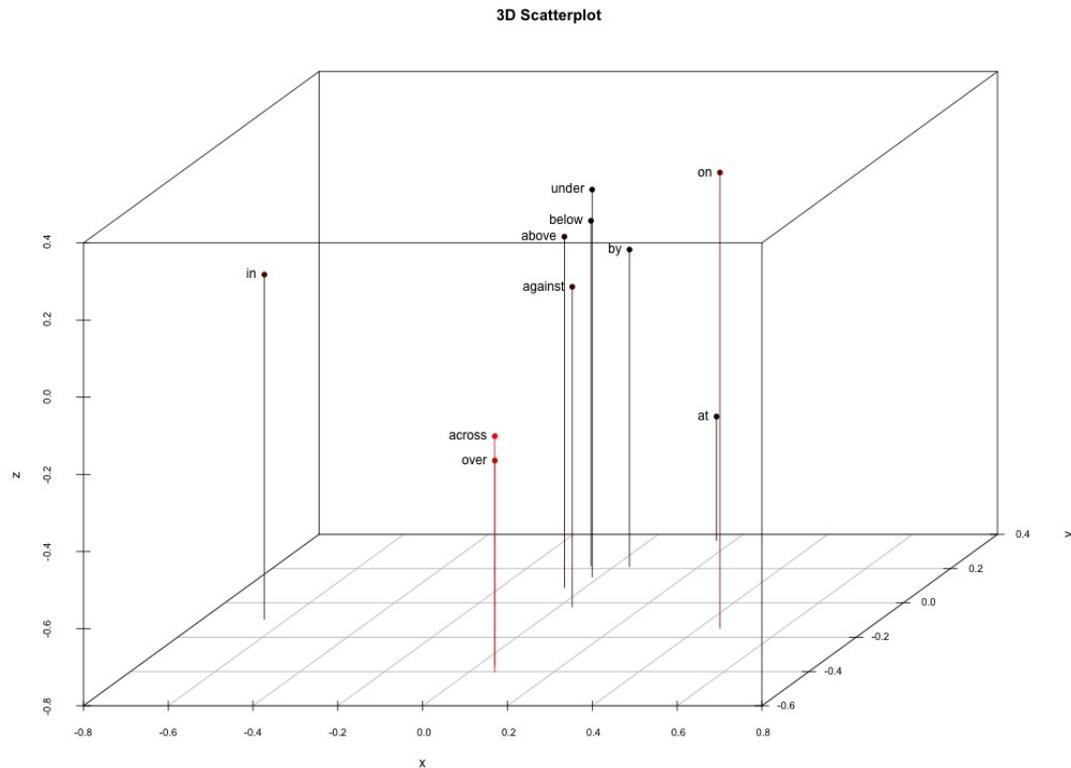


Figure 1: A three-dimensional representation of the conceptual contiguity of spatial particles in English

References

- Bird, Steven. 2006. NLTK: the natural language toolkit. Proceedings of the COLING/ACL on Interactive presentation sessions, 69-72. Sydney, Australia: Associations for Computational Linguistics.
- Bowerman, Melissa & Soonja Choi. 2001. Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In Melissa Bowerman & Stephen C. Levinson (ed.), *Language acquisition and conceptual development*, 475-511. Cambridge, England: Cambridge University Press.
- Feist, Michele I. 2008. Space between languages. *Cognitive Science* 32.1177-99.

- Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Co-varying collexemes in the *into-causative*. In Michel Achard & Suzanne Kemmer (ed.), *Language, culture and mind*, 225-36. Stanford, CA: CSLI Publications.
- Levinson, Stephen C, Sérgio Meira & The Language and Cognition Group. 2003. 'Natural Concepts' in the spatial topological domain – adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language* 79. 485-516.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8. 209-43.
- Vandeloise, Claude. 2003. Containment, support, and linguistic relativity. In Hubert Cuyckens, René Dirven & John R. Taylor (ed.), *Cognitive approaches to lexical semantics*, 393-425. Berlin: Mouton de Gruyter.
-

What are the suitable corpus data to support L1 teaching (not only) in Czech?

Lucie Chlumská, Anna Čermáková
Charles University in Prague

Despite the fact that corpus linguistics tools and methods have proven to be extremely successful in language teaching (L2), there has been, so far, considerably less attention devoted to the potential of applying corpus-based mother tongue (L1) teaching in elementary and/or secondary schools. One of the exceptions, research done by Sealey and Thompson (2004) during their CLLIP project (Corpus-based Learning about Language in the Primary School), suggests that corpora and corpus-based materials can be successfully used already with very small children. It is, however, clear that corpora that are normally available are not readily suitable as pedagogical resource (Braun 2007). One of the solutions suggested by researchers is using for child learners corpora made up of language children are familiar with, that is writing for children (e.g. Thompson & Sealey 2007). This initial question, i.e. what texts should a corpus for children contain to be useful in mother tongue teaching, inspired our larger project with its main objectives of 1) investigating and comparing various text types, focusing mainly on texts written for as well as by children, 2) building a corpus for L1 (Czech) teaching, and as a final step 3) designing a suitable web interface for teachers and pupils/students to work with. This particular study is a follow-up of the pilot research from 2015 (Čermáková & Chlumská 2015) and focuses on the first two steps in the project: analyzing various types of texts and identifying the criteria for their inclusion into the school corpus.

Previous research

The pilot study was based on the analysis of selected linguistic features in a subcorpus of the Czech National Corpus labelled JUN (broadly fiction aimed at children and young readers, 4.76 million tokens). We explored the JUN corpus in terms of the overall frequency characteristics in comparison with three reference corpora: BEL (fiction for adult readers), PUB (newspaper texts), and SKRIPT (children's school

essays). We analyzed the distribution of POS and compared the most frequent vocabulary (the top 1 000 lemmas in each corpus) and specifically focused on lexical verbs, adjectives, nouns, and adverbs in the respective corpora (cf. Sealey & Thompson 2007; Thompson & Sealey 2007). Both JUN and BEL corpora, the two fiction corpora, showed a significant similarity but a more detailed qualitative analysis indicated certain differences as well. The analysis of the SKRIPT corpus, representing student writing, revealed considerable differences between the language of younger pupils (10-14) and students (15-18), pointing out to possibly different needs of these two age groups in terms of corpus design.

Current study

The main objective of this study is to validate our preliminary findings on bigger and more varied data: the JUN subcorpus now includes new texts amounting to 12.6 million tokens in total as well as new text types (not only fiction, but also popular science books and magazines for children and teenagers). In addition to the SKRIPT corpus, representing so far the only source of actual children's language (written schools essays), we also make use of the SCHOLA corpus, containing the transcriptions of school lessons (including the spoken language of pupils/students as well as teachers). The analysis focuses mainly on common lexis, particularly comparing children's passive vocabulary (represented by JUN corpus) and active vocabulary (SKRIPT and SCHOLA corpus), while taking into account different text types and their role in the school corpus design.

References

- Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19(03), 307-328.
- Čermáková, A. & Chlumská, L. (2015). Compiling Corpus for School Children to Support L1 Teaching: Case of Czech. Paper presented at the *Corpus Linguistics 2015* conference, Lancaster, 23th July 2015. Abstract available at <http://ucrel.lancs.ac.uk/cl2015/doc/CL2015-AbstractBook.pdf>.
- Sealey, A. & Thompson, P. (2007). Corpus, Concordance, Classification: Young Learners in the L1 Classroom. *Language Awareness* 16(3), 208-223.
- Sealey, A. & Thompson, P. (2004). 'What do you call the dull words?' Primary school children using corpus-based approaches to learn about language. *English in Education* 38 (1), 80-91.
- Thompson, P. & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics* 12 (1), 1-23.

Corpora

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kovářiková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondříčka, P. – Zasina, A.: *SYN2015: representative corpus of written Czech* (JUN subcorpus). Insti-

tute of the Czech National Corpus, FA CU, Prague. Available at WWW: <http://www.korpus.cz>

Šebesta, K. – Goláňová, H. – Křen, M. – Procházka, P.: SCHOLA2010: corpus of spoken Czech at schools – transcribed recordings of school lessons at Czech elementary and secondary schools. Institute of the Czech National Corpus, FA CU, Prague. Available at WWW: <http://www.korpus.cz>

Šebesta, K. – Goláňová, H. – Jelínek, T. – Jelínková, B. – Křen, M. – Letafková, J. – Procházka, P. – Skoumalová, H.: SKRIPT2012: acquisition corpus of written Czech – transcriptions of school essays written by pupils/students of Czech elementary and secondary schools. Institute of the Czech National Corpus, FA CU, Prague. Available at WWW: <http://www.korpus.cz>

Modifying corpora authenticity to benefit beginner level EFL students: An update on SCoRE

Kiyomi Chujo¹, Kathryn Oghigian², Shiro Akasegawa³

¹Nihon University, ²Waseda University, ³Lago Institute of Language

It could be argued that authenticity is one essential aspect of corpora that benefits researchers, lexicographers and linguists but is not necessarily critical to certain second language learners. In fact, for beginner and intermediate learners, authentic corpora can be overwhelming and can contain complex and even erroneous language (Allan, 2009; Chujo, Oghigian & Akasegawa, 2015). As Braun pointed out, perhaps “it is time for a move from data driven learning (DDL) to needs-driven corpora, activities and methodologies” (2007:316). In re-thinking the purity of corpora, it may be possible to open doors to the benefits of DDL that have thus far been mainly limited to advanced level L2 learners as an intermediate step toward using authentic corpora.

The Sentence Corpus of Remedial English (SCoRE), first introduced in *Multiple Affordances of Language Corpora for Data-driven Learning* (Leńko-Szymańska & Boulton, 2015), currently contains eleven categorized grammar items with 6,000 level-specific, semi-authentic sentences written to satisfy particular pedagogical considerations, i.e., appropriateness and usability, and fair use for copyright issues. In this paper, the authors will present new SCoRE tools and discuss results from a field test in L2 classrooms at a Japanese university.

The newest modifications include changes to the pattern browser to make it more user-friendly and to allow users to locate search results by grammar item, keyword, and/or proficiency level, a simple concordancer (allowing a choice of KWIC or sentence, with minimal functions such as sampling and sorting), a fill-in-the-blank quiz function (for both creating and scoring) aimed at motivating learners by providing instant feedback, and a means to download desired SCoRE data in an EXCEL format to help teachers create DDL worksheets, quizzes, and homework.

Sixty participants in two classes tested the effectiveness of this current version of SCoRE for use in a low proficiency Japanese EFL university course aimed at remediating previously identified grammar issues. Participants analyzed grammatical patterns in SCoRE to form hypotheses, which were confirmed or corrected, and then produced language to consolidate learning. Pre- and post-tests measured learning

effectiveness; the gain between the pre- and post-test was statistically significant and showed improvement in proficiency for the targeted grammar items. Participants also gave feedback on SCoRE on a 42-item questionnaire, indicating that the corpus was useful, they were able to study at their own pace, and the activities were enjoyable. They further reported that the sentences provided clear, observable grammatical patterns and were appropriate in level, length, vocabulary, and structure.

Although this study is only the first of a series, preliminary indications are that it is useful for the target audience (remedial L2 English learners) and we hope it may have a broader use for EFL students of other languages.

References

- Allan, Rachel. 2009. Can a graded reader corpus provide ‘authentic’ input? *ELT Journal* 63(1). 23-32.
- Braun, Sabine. 2007. Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19(3). 307-328.
- Chujo, Kiyomi, Oghigian, Kathryn and Akasegawa, Shiro. 2015. A corpus and grammatical browsing system for remedial EFL learners. In Agnieszka Leńko-Szymańska and Alex Boulton (eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins. 109-128.
- Leńko-Szymańska, Agnieszka and Boulton, Alex. 2015. *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins.
- The Sentence Corpus of Remedial English (SCoRE), version 2. College of Industrial Technology, Nihon University. <http://score.lagoinst.info/> (accessed 1 April 2016).

A meta-analysis of DDL research 2: Variation, good practice and future work

Tom Cobb¹, Alex Boulton²

¹Université du Québec à Montréal, Canada ²Atilf, CNRS & Université de Lorraine, France

Please note that this paper is the second of two submitted conjointly. See Boulton & Cobb on page 15.

So far, our meta-analysis of research on using corpora a learning or reference tool 1991-2014 has determined a mean effect size for within-groups studies (pre/post-test design) of 1.50 and between-groups (control/experimental) of .95, both well “in the neighbourhood” (Plonsky & Oswald, 2014) of strong outcomes in the contexts of applied linguistics and these particular models. The standard deviations, however, are high (SD=.91 within-groups, SD=.99 between) so the sources of variation within this generally successful approach are of interest, and indeed investigating these is one of the benefits of bringing together a large body of work (Lipsey & Wilson, 2001). Once a final set of studies had been determined, they were iteratively coded for moderator variables in the event a strong overall result would indicate the value

of component parcelling, as turned out to be the case. The third research question, then, was as follows:

- RQ3. How can we best account for any variation observed?

The moderator variables that were present in many studies as well as bearing practical/theoretical interest were as follows: publication variables (Do effect sizes vary with time in one direction or another? With length of report?); design variables (Do pre-post or control-experimental designs yield different patterns of effects, in relation to sample size and the type of statistical analysis?); population variables (Do effects vary with sample size? With the nature of the control group where applicable, whether true with zero treatment or merely a comparison group? Whether groups were intact or randomly assigned?); instrumentation variables (Was the required response selected, constrained or free?); geographical variables (In the region and presumed culture where the study took place?); learner variables (Whether the context was EFL or ESL? Whether the motivation was general, specific or academic English? From what first language? For what level, sophistication, purpose for study, in what type of institution?); treatment variables (Classroom or lab? Short or long duration? DDL-delivery by concordancer, CALL program with concordance features, or paper concordances? Using public, local, or parallel corpora?); cognitive operation (To learn or look up?); target language form (vocabulary, lexicogrammar, grammar, or discourse); target language skill (listening, speaking, reading, writing, or translation?). In this wide trawl, some of our variables have yielded fascinating and occasionally counter-intuitive information, while others had insufficient data to draw clear conclusions.

The usual procedure for investigating variance in a meta-analysis is either to perform a regression analysis to see how much of the variance can be attributed to different factors, which assumes continuous moderator variables and has only been done for a couple of meta-analyses in our field, (such as Goldschneider & DeKeyser, 2001) or in the more usual case of categorical variables (like ours) to simply form subgroups and compare average effects. Some highlights from our comparison exercise include that DDL has been most successful in Asia and the Middle East; DDL works well across proficiency levels; DDL works best with language specialists but well enough with social and other scientists; DDL works better ‘hands-on’ than on paper – and at least 20 minutes-worth more to be disclosed in the presentation, which will conclude with suggestions for future research in terms of both fruitful topics and essential practices.

References

- Goldschneider, Jennifer M. & Robert M. Dekeyser. 2001. Explaining the ‘natural order of L2 morpheme acquisition’ in English: A meta-analysis of multiple determinants. *Language Learning* 51. 1-50.
- Lipsey, Mark W. & David Wilson. 2001. *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
- Plonsky, Luke & Frederick L. Oswald. 2014. How big is ‘big’? Interpreting effect sizes in L2 research. *Language Learning* 64(4). 878-912.

Cohesion or coesione? L1 Italian learners' use of linking adjuncts in argumentative essays

Meredith D'Arienzo
Georgia State University

Cohesion is an essential aspect of successful writing and one whose importance has been recognized more fully in recent years. Writing cohesive texts is a challenge for learners of a second language (L2). Learners tend to rely on native language (L1) strategies (Petchprasert, 2013) and often do not produce successful cohesive texts. Past research has shown that cross-linguistic influence (influence on L2 production based on a speaker's L1) affects nearly all areas of L2 competence, including cohesion (Jarvis & Pavlenko, 2008). Previous research on the Italian subset of the International Corpus of Learner English (ICLE; Granger, Dagneaux & Meunier, 2009) has found that native speakers of Italian often produce "foreign sounding and non-native English" at the discourse level due to, among other causes, their use of connectors and discourse markers (Zagrebelsky, 2009, p. 208). Indices of cohesion, along with lexical aspects, are one of the strongest predictors in differentiating the L1 of L2 writers (Crossley & McNamara, 2012). Therefore, errors in cohesion may be unique to L1 groups, and strategies to help learners with these errors could benefit from corpus-based research that identifies errors and variations due to cross-linguistic influence unique to the L1 group of interest.

This study explores the quantitative and qualitative differences in the use of cohesive devices by Italian learners of English as compared to their use by native speakers of English (NSE). Specifically, it compares the use of linking adjuncts (e.g., firstly, in summary, eventually) in academic essays written by Italian learners of English and by NSE. The data come from academic essays that comprise the Italian subset of ICLE and the Louvain Corpus of Native English Essays (LOCNESS; Granger, Sanders & Connor, 2007). The focus of the comparison is on organizational linking adjuncts, as they are a key component of academic texts (Carter & McCarthy, 2006). Additional linking adjuncts ('target linking adjuncts') were selected based on the author's familiarity with Italian and on predictions about which linking adjuncts might be problematic for Italian learners due to cross-linguistic influence. The use of linking adjuncts is analyzed first quantitatively, using log-likelihood values, to determine which linking adjuncts are used significantly more or less by Italian learners. For those linking adjuncts with statistically greater or lesser use, the data are analyzed qualitatively in order to attempt to explain the reason for the divergence in use. Many of the predictions for the target linking adjuncts based on cognates and false cognates between Italian and English are confirmed by the data, and in the instances where the results are unexpected, further qualitative analysis of the linking adjuncts and their context is conducted to explore possible explanations.

Based on the findings of this analysis, pedagogical interventions are suggested to facilitate the learning of linking adjuncts for Italian learners. In particular, teaching materials are proposed to help learners improve the accuracy of their use of linking adjuncts, broaden the range of linking adjuncts in their vocabulary, and understand and avoid mistakes with linking adjuncts typical of Italian learners.

References

- Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English: a comprehensive guide; spoken and written English grammar and usage*. Ernst Klett Sprachen.
- Crossley, Scott A. & Danielle S. McNamara. 2012. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. *Approaching Language Transfer through Text Classification: Explorations in the Detection based Approach. Multilingual Matters*.
- Granger, Sylviane, Estelle Dagneaux & Fanny Meunier (eds.). 2009. *ICLE: International Corpus of Learner English*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.
- Granger, Sylviane, C. Sanders & Ulla Connor. 2007. LOCNESS: Louvain corpus of native English. *Belgium: Universitaires de Louvain*.
- Jarvis, Scott & Aneta Pavlenko. 2008. *Cross-linguistic influence in language and cognition*. New York: Routledge.
- Petchprasert, Anongnad. 2013. A study of cohesive markers used in L1 and L2 essay writing: Translation versus direct composition. *3L: The Southeast Asian Journal of English Language Studies*, 19(1). 19-33.
- Zagrebelsky, Maria T. P. 2009. Computer Learner Corpora: Their Use in Research and in Teaching and Learning. *Using Corpora to Learn about Language and Discourse*, 66, 199.

Dutch and German NN compounds in translation

Hinde De Metsenaere

Ghent University

Dutch and German are two closely related Germanic languages. Nevertheless, the overlap in distinct features, such as an impressive number of nominal compounds with a nominal first constituent (NN) in both lexicons, can be misleading for translator students when translating from one language into the other: Where German uses a NN compound, Dutch may opt for an alternative construction, such as a phrase (Booij/Van Santen 1998:148, Campe 2010:208, Hüning/ Schlücker 2010:791ff, Hüning 2010).

Although it is true that Dutch and German NN compounds may pose challenges to translator students, as has been addressed in various studies involving either language (on Dutch: Ross 2004; on German: Junczys-Dowmunt 2009, Machowski 2010, Balzer Haus 2011, Mohammed 2011), the Dutch preference for alternatives to NN compounds as compared to German has been hardly addressed in the literature. There are studies on German adjective-noun compounds and their corresponding Dutch adjective-noun phrases (Hüning/Schlücker 2010: 801ff, Hüning 2010), but studies on Dutch alternatives for German NN compounds are rare (De Metsenaere et al. 2014). Identifying factors that influence the choice for a NN compound or an alternative construction may, however, be of interest to translator students.

To shed more light on NN compound use in Dutch and German in order to provide translator students with the tools to translate them, the bidirectional PAND corpus was used, which contains twenty German and Dutch novels and their translations, six novels and their translations of which are covered by the present study. From each novel, 250 compound types and their translations were manually extracted and analysed to gain insight in (1) the translation preferences of Dutch and German NN compounds, (2) the constructions other than a NN compound that lead to a NN compound in translation, (3) the translation of recursive NN compounds, and finally (4) meaning relations within NN compounds that may expose regularities amongst the preferences in both languages.

Through quantitative and qualitative corpus analysis preferences could be identified amongst the translations of NN compounds in Dutch and German. Simple nouns, derivatives and adjective constructions occurred most frequently as compound alternatives in both languages, whereas prepositional and appositional constructions on the one hand, and genitive constructions on the other hand turned out to be typical for Dutch and German translations respectively. Recursive compounds were found to be translated more frequently by means of a phrase than non-recursive compounds. Finally, it could be demonstrated that different preferences in the rendering of semantic content led to Dutch phrases where German used a compound.

References

- Balzer Haus, Berit. 2011. Fünf Schwierigkeiten beim Schreiben der Wahrheit: Typologie der Übersetzungsprobleme von Heines Lyrik ins Spanische. *Estudios de Traducción, 1*, 99-115.
- Booij, Geert & Van Santen, Ariane. 1998. *Morfologie. De woordstructuur van het Nederlands*. 2nd edn. Amsterdam: Amsterdam University Press.
- Campe, Petra. 2010. Syntactic variation in German adnominal constructions: An application to the alternatives 'genitive', 'apposition' and 'compound'. *Grammar between norm and variation*. Alexandra Lenz & Albrecht Plewnia (eds.). 193-218. Frankfurt am Main: Peter Lang.
- De Metsenaere, Hinde, Campe, Petra, Vandepitte, Sonia, & Van de Velde, Marc. 2014. Niederländische Übersetzungen deutscher NN- und VN-Komposita. Konvergenzen und Divergenzen in einem literarischen Parallelkorpus. *Germanistische Mitteilungen, 40* (1). 65-78.
- Hüning, Matthias. 2010. Adjective + noun constructions between syntax and word formation in Dutch and German. *Cognitive perspectives on word formation*. Alexander Onysko & Michel Sascha (eds.). 195-218. Berlin, New York: Mouton de Gruyter.
- Hüning, Matthias & Schlücker, Barbara. 2010. Konvergenz und Divergenz in der Wortbildung. Komposition im Niederländischen und im Deutschen. *Kontrastive Germanistische Linguistik*. Antje Dammel, Sebastian Kürschner & Damaris Nübling (eds.). 783-825. Hildesheim, Zürich, New York: Georg Olms Verlag.
- Junczys-Dowmunt, Marcin. (2009). *Niemieckie rzeczowniki złożone i ich polskie*

odpowiedniki. Automatyczna ekstrakcja, analiza i weryfikacja na podstawie korpusów równoległych [German compound nouns and their Polish equivalents. Automatic extraction, analysis and verification based on parallel corpora]. Adam Mickiewicz University, Poznań.

- Machowski, Szymon. 2010. Analiza translatywna niemieckich, polskich i hiszpańskich compositów [Translative analysis of German-Polish-Spanish composita]. *Investigationes Linguisticae*, 19, 133-152.
- Mohamed, Nasser. E. A. 2011. Deutsche Nominalkomposita und ihre Übersetzungsproblemantik ins Arabische. *Lebende Sprachen*, 56 (1). 65-76.
- Ross, Dolores. 2004. Profili morfologici della lingua medica: Contrasti in ambito germanico-romanzo [Morphological profiles of medical language: Conflicts within the Germanic-Romance field]. *Rivista Internazionale di Tecnica della Traduzione*, 8, 107-128.

Using specialized corpora in the ESP classroom to explore corporate pragmatic strategy

Bridgit C. Fastrich, Christoph Wolk
Justus Liebig University Giessen

Multinational corporations (MNCs) are increasingly finding themselves under public scrutiny, making it standard practice for MNCs to produce texts that fall in the realm of corporate social reporting (cf. Fuoli 2012: 56). In addition to standard reporting genres in which communication takes place in a rather top-down, unidirectional manner, firms are taking advantage of digital technologies to communicate with consumers directly, engaging in novel forms of corporate-consumer interaction in which the distinctions between mass and personal communication are blurred (Lüders 2008: 684–685). It is in this context that digital channels of communication like social media pages, message boards and further interactive platforms have gained relevance for LSP and, due to the emergence of English as the business lingua franca (Poppi 2011: 131), ESP pedagogy.

These communicative channels and resulting “plethora of digital genres” (Luzon Marco 2002: 41) present exciting challenges not only for the ESP genre researcher; from a discourse-pragmatic perspective, they are also a potentially valuable source of data that, when made amenable for corpus linguistic study in the form of so-called specialized corpora (cf. Baker 2006: 26–30; Gavioli 2005: 6–8), can be used to reveal patterns in corporate rhetorical strategy. The current paper probes this potential by introducing an example of how data-driven learning (DDL) methods can be used in the ESP classroom to study corporate discourse strategies. While the body of research looking at DDL applications in LSP is steadily increasing (e.g. Charles 2015; Cheng 2010; Gavioli 2005; Rodgers et al. 2011), the teaching of corpus-assisted discourse studies (CADS) is still in its relative infancy for researchers and students alike (Baker 2009: 74), and is thus an interesting prospect in the context of LSP pedagogy.

The specialized corpus compiled for these purposes is based on a multi-national and critically acclaimed campaign called *Our Food. Your Questions*, which was launched by McDonald’s corporation with the explicit purpose of improving public

perceptions of its brand (Laird 2013). This online platform takes a question-and-answer format, in which McDonald's provides often elaborate responses to questions of a neutral (e.g. *What's in your secret sauce?*), critical (e.g. *Are your chicken nuggets really that disgusting pink stuff thats [sic] seen in pics online?*) and flattering (e.g. *I love McDonald's! Why is it so good?*) nature. The resulting corpus, called the *Your Questions Corpus*, was scraped using a Python script and currently includes both an English-language and German-language component, each of which comprises approximately 750,000 tokens of both questions (15% of corpus) and answers (85%).

The current paper will first introduce this monolingual comparable corpus, including both its methodological potential and challenges for use in the LSP classroom. It will then present the findings of a pilot study conducted at a German university in which the *Your Questions Corpus* was used by advanced students of ESP to investigate corporate rhetorical strategy. The paper will report on the procedure undertaken (including the contextual and cultural embedding for the lesson), the ability of students to discover patterns autonomously, and their subjective learning experience with this methodology. The classroom application also includes a productive component, in which students produce their own answers to similar questions. Findings of a feasibility study have already shown that students are capable of identifying and interpreting differences in rhetorical patterns, which sometimes contradict students' cultural and linguistic intuitions.

References

- Baker, Paul. 2006. *Using corpora in discourse analysis*. London: Continuum.
- Baker, Paul. 2009. Issues arising when teaching corpus-assisted (Critical) Discourse Analysis. In Lombardo, Linda (ed.), *Using corpora to learn about language and discourse*, 73–98. Bern: Peter Lang,
- Charles, Maggie. 2015. Same task, different corpus: The role of personal corpora in EAP classes. Leńko-Szymańska, Agnieszka and Alex Boulton (eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* [Studies in Corpus Linguistics 69], 129–154. Amsterdam: John Benjamins.
- Cheng, Winnie. 2010. Hong Kong Engineering Corpus: Empowering professionals-in-training to learn the language of their profession. In Campoy-Cubillo, Mari Carmen, Begona Belles-Fortuno and Maria Lluisa Gea-Valor (eds), *Corpus-based approaches to English language teaching*, 67–78. London: Continuum.
- Fuoli, Matteo. 2012. Assessing social responsibility: A quantitative analysis of Appraisal in BP's and IKEA's social reports. *Discourse & Communication* 6. 55–81.
- Gavioli, Laura. 2005. *Exploring corpora for ESP learning* [Studies in Corpus Linguistics 69]. Amsterdam: John Benjamins.
- Laird, Kristin. Marketer of the Year 2012: McDonald's Canada. 23 January, 2013. *Marketing*. Rogers Media, Inc. <http://www.marketingmag.ca/advertising/marketer-of-the-year-2012-mcdonalds-canada-70067>. (10 December 2015.)
- Lüders, Marika. 2008. Conceptualizing personal media. *New Media & Society*

10(5). 683–702.

- Luzon Marco, Maria Jose. 2002. A genre analysis of corporate home pages. *LSP & Professional Communication* 2(1). 41–56.
- Poppi, Franca. 2011. Companies' websites as vehicles for expressing corporate identity: A case study on the use of English as a lingua franca. In Gotti, Maurizio and Giuliana Garzone (eds.), *Discourse, communication and the enterprise. Genres and trends*, 131–148. Bern: Peter Lang.
- Rodgers, Ornait, Angela Chambers and Florence Le Baron-Earle. 2011. Corpora in the LSP classroom: A learner-centred corpus of French biotechnologists. *International Journal of Corpus Linguistics* 16(3). 391–411.
-

How inter-annotator agreement helps to improve error annotation schemes in learner corpora

Alena Fenogenova, Elizaveta Kuzmenko, Olga Vinogradova
National Research University Higher School of Economics

It is widely acknowledged that learner corpora are mainly useful when error-annotated.² At the same time, annotation performed by humans is subject to influence of various factors [Leech 2005, Glaznieks et al.2014, Bayerl 2008]. The present research describes the situation in which the experiment in evaluating annotation agreement has brought to light problems in annotation practice and thus improved annotation instruction.

Russian Error-Annotated Learner English Corpus (further REALEC) is available at <http://realec.org> with over 1200 pieces of students' writing (about 360 thousand word tokens) and at http://realec.org/hse/#/data_4_staff/IELTS/ with about 2000 essays written by students in the examination comprising almost 434 thousand word tokens. Experts (teachers and students familiar with annotation approaches adopted in REALEC) mark the essays and annotate them according to the error classification scheme established at the start of setting up the corpus [Kuzmenko and Kutuzov 2014].

REALEC error annotation scheme consists of 4 layers: error type, error cause, linguistic 'damage' caused by the error and the impact of the error on general understanding of the text. As the first layer of the annotation scheme has received most attention so far, this paper focuses only on its features and does not touch upon three other layers. The scheme consists of 151 categories organized into a tree-like structure. Annotators are instructed to choose a specific tag for the error they have spotted, however, in some exceptional cases they can apply one of the tags of general categories (Grammar, Vocabulary, etc.).

We have already carried out research on reliability of REALEC annotation scheme [Kutuzov et al. 2015]. The main purpose of the present research is to outline different cases of lack of inter-annotator agreement in order to come closer to uniform approach in annotation in REALEC.

²The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2015-2016, and the authors are member of the team that has won a Research Team Project Competition in 2016 (16-05-0057 at <https://www.hse.ru/en/science/scifund/nug>).

For the annotation experiment we have taken a short text (596 words), which was independently annotated by 12 professors of English. The number of errors found in this text by different annotators varied from 23 to 123; the average number of errors spotted by a person was equal to 69. Overall, 153 different errors have been revealed in the student text, of which some were marked by more, others by fewer annotators. The resulting distribution of types of errors (including even those that were spotted by just one annotator) is demonstrated in Figure 1.

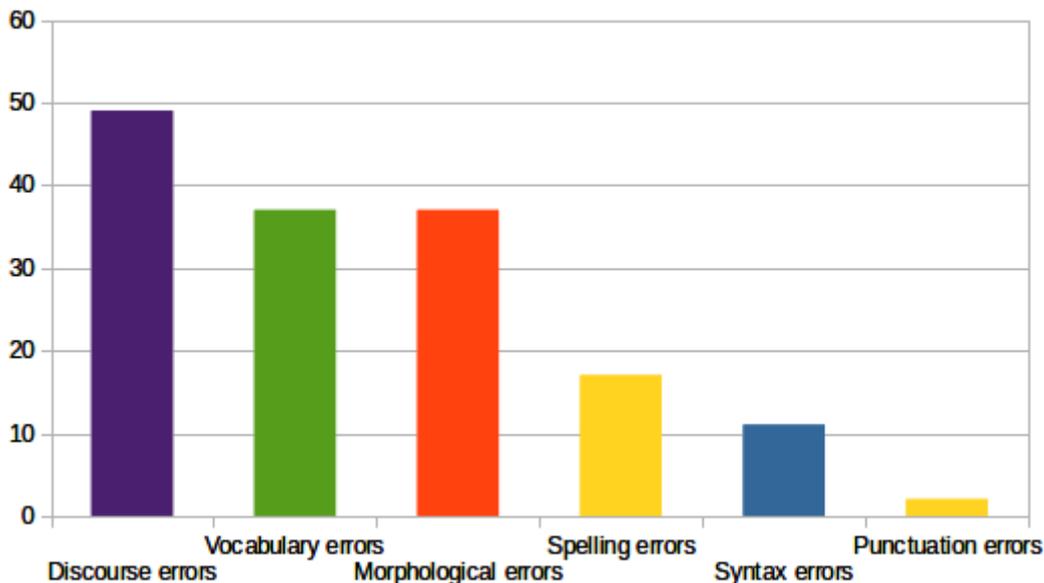


Figure 1: Distribution of errors

We were interested in the level of agreement among annotators, which indicates whether our annotation scheme is consistent, and to what extent the annotators really adhered to the annotation instructions. NLP community has several established means to calculate inter-annotator agreement; the most widely used of them is Krippendorff’s alpha (further KA) [Krippendorff 2012]; see [Passonneau 1997] for explanation on why precision and recall metrics are not feasible for this task.

First we computed KA for the errors as they were assigned, including those that were spotted by just one annotator. It gave us the value of 0.22, which is much lower than in the first experiments. This is due to the fact that contrary to the previous design, we did not predefine error spans. As a result, some annotators tended to hypercorrect, while others did not set themselves the task of thoroughly looking for all the errors. If we eliminate errors spotted by less than 4 annotators, we have KA equal to 0.26, which is closer to the raw counts from the previous experiment. At the next stage we calculated KA taking into account only the upper level of REALEC annotation scheme. We did that by treating all types of grammar errors as one macrocategory ‘Grammar’; the same was true for discourse, vocabulary, etc. In other words, we observed only the upper level of our annotation scheme - macrocategories such as Punctuation, Orthography, Grammar (internally divided into morphology and syntax), Lexis and Discourse. For these nominal categories we compared annotations in a binary way (they either match, or they do not). KA

value was found to be 0.29.

At the next stage in the experiment we turned to the analysis of annotators' disagreement cases. There are 82 wrong tags among the total of 650 suggested by all annotators, and these are cases when an annotator either suggested the wrong correction and as a result the tag is wrong, too, or the annotation suggested does not correspond to the instruction of how to choose a tag adopted in REALEC. However, there also are 32 cases out of 154 different tags in which errors were tagged differently by different annotators, and none of the annotators violated English norms or annotation conventions, and therefore all their decisions can be justified. The following examples can be seen as typical representatives:

(Example 1)

But these methods failed to create a protection from criminality > **protection**
Form of articles

But these methods failed to create a protection from criminality > **prevent**
Choice of lexical item

But these methods failed to create a protection from criminality > **give protection**
Choice of lexical item + Form of articles

But these methods failed to create a protection from criminality > **protect us**
Redundant words (a Vocabulary tag - the combination of CREATE and PROTECTION in native speakers corpora only occurs with the PROTECTION followed by another noun)

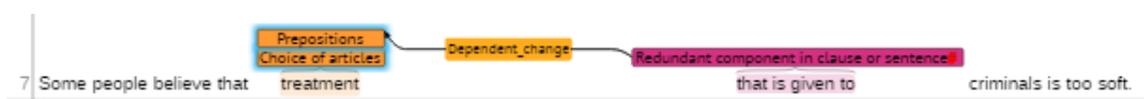
But these methods failed to create a protection from criminality > **protect us**
Redundant component in clause or sentence (a Discourse tag - the result of intuitive approval of the combination of CREATE and PROTECTION as an acceptable lexical expression)

(Example 2)

Some people believe that treatment that is given to criminals is too soft >
criminals are treated too softly **Word choice**

Some people believe that treatment that is given to criminals is too soft >
penalty that is given to criminals is too mild **Choice of lexical item** (both times)

Some people believe that treatment that is given to criminals is too soft > **given**
Redundant component in clause or sentence (a Discourse tag in view of the proximity of another *that*)



(Example 3)

If *person* is quite normal, has no previous conviction and is not dangerous to the society, the court may put **a person** on probation. >**a person** **Form of article** . . . **that person** **Choice of determiner** (Both suggestions are Grammar tags)

If *person* is quite normal, has no previous conviction and is not dangerous to the society, the court may put **a person** on probation. >**a person** **Form of article** . . . **that person** **Lack of referential tool** (The first is a Grammar tag and the second, a Discourse tag)

If *person* is quite normal, has no previous conviction and is not dangerous to the society, the court may put **a person** on probation. >**a person** **Form of article** . . . **him/her** **Lack of referential tool** (The first is a Grammar tag and the second, a Discourse tag)

If *person* is quite normal, has no previous conviction and is not dangerous to the society, the court may put **a person** on probation. >**a person** **Form of article** . . . **the person** **Choice of article** (both corrections deal with articles only)

It is clear that the scope and the level of change suggested by an annotator cannot be formally defined, and besides, it is not often that two persons - native speakers or fluent speakers of a foreign language – will not differ in their intuitive perception of what is acceptable in the language. However, if annotators stick to the decision to restrict corrections to those that they find absolutely necessary to stay within the norm, first, and, second, if for the chosen correction they select tags only for the core change, and not for all the words that have to change as a result of the core change, the variation across annotators is bound to reduce dramatically. Both these requirements accompanied by examples from the corpus are to be included in the REALEC Annotation Manual, and some training based on complicated cases from the experiment described above will be presented to all the annotators.

We have performed analysis of problematic cases of annotators inconsistency to reveal weaknesses and strengths of the annotation scheme. In future we plan to compare our results with the similar experiment carried out over the same text in ICLE, The International Corpus of Learner English, and this will hopefully advance our conclusions and annotation efficiency.

References

- Artstein, Ron, and Massimo Poesio. 2008. "Inter-coder agreement for computational linguistics." *Computational Linguistics* 34.4: 555-596.
- Granger, S. 2003. "The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research." *Tesol Quarterly*, 538-546.
- Bayerl, Petra Saskia. 2007. "Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies" in *Computational Linguistics*, 33(1), pp. 3-8.

- Glaznieks, Aivars, Nicolas, Lionel, Stemle, Egon, Abel, Andrea & Lyding, Verena. 2014 “Establishing a Standardised Procedure for Building Learner Corpora” in *Journal of Applied Language Studies*, vol. 8, 3, 5-20
- Krippendorff, Klaus. 2012. “Content analysis: An introduction to its methodology.” Sage.
- Kutuzov A., Kuzmenko E., Vinogradova O. 2015. “Evaluating inter-rater reliability for hierarchical error annotation in learner corpora”, in *Corpus Linguistics 2015: Abstract Book* Ch. 89. P. 211-213.
- Kuzmenko, E., and A. Kutuzov. 2014. "Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning." *NEALT Proceedings* Series Vol. 22: 87
- Leech, G. 2015. “Adding linguistic annotation”. In *Developing linguistic corpora : a guide to good practice*. Oxbow Books, Oxford, pp. 17-29.
- Passonneau, R. J. “Applying reliability metrics to co-reference annotation.” CoRR URL <http://arxiv.org/abs/cmp-lg/9706011>.

The research grant proposal genre: corpus-based findings and applications for data-driven learning

Lynne Flowerdew

Hong Kong University of Science and Technology (formerly)

The ESP literature on academic writing has witnessed an increasing number of accounts on the issue of ‘writing for publication’, specifically the scientific research article (RA) genre following the trend of the internationalisation of universities (Hyland, 2009). However, Englander (2014) makes the point that the RA is but just one of many different kinds of documents written by scientists. She outlines a chain of interrelated documents accompanying the RA, one of which is the “behind-the-scenes” research grant proposal. As Swales (1990: 178) puts it ‘Published RAs increase the chances of follow-up grants and research grants increase the chances of publishable RAs’. In spite of the importance of the research grant proposal, it is surprising that there are only a few accounts of corpus-based research on this genre in the ESP literature and even fewer accounts of pedagogic applications (see Flowerdew, 2016). My aim in this presentation is two-fold: first, to present a brief survey of the corpus-based research on grant proposal writing and, second, to describe how this research has informed the design and delivery of a corpus-inspired module on grant proposal writing for science and engineering post-graduate students at a tertiary institution in Hong Kong.

Corpus-based studies on research grant proposals are reported in Connor & Mauranen (1999), Connor & Upton (2004), Feng & Shi (2004), Feng (2008), Matzler (2014) and Tardy (2011), which all take a Swalesian (1990, 2004) top-down approach to the analysis supplemented by more qualitative lexico-grammatical investigations. Moreover, Feng & Shi (2008), Tardy (2011) and Matzler (2014) adopt an ethnographic perspective on the data by conducting follow-up interviews with proposal authors to shed light on the findings. Feng’s (2008) study takes a cross-linguistic and cross-cultural perspective on the data from Chinese scholars.

The findings from the aforementioned corpus studies, especially those of Connor & Upton (2004) and Feng & Shi (2004), proved very helpful for alerting students to the prototypical move structure patterning of the research grant proposal. To supplement this top-down approach, students carried out bottom-up lexico-grammatical searches for generic, non-topic specific phrases for particular move structures, e.g. indicating a gap in knowledge. To this end, two freely-available corpora of academic writing were used; the Michigan Corpus of Upper-level student papers, MICUSP (see Römer, 2012) proved an ideal resource as it contains 47 proposals as did the Corpus of Research Articles, CRA (see Lin & Evans, 2012). Examples of search queries student conducted will be presented. Moreover, it was found that these two corpora also contain English as a lingua franca (ELF) type language (Mauranen, 2011), a phenomenon that can be accommodated within Granger's (2015) revised contrastive interlanguage analysis framework. I will conclude by mapping out some future avenues for corpus-based ELF research in ESP written text, while at the same time underscoring the importance of learner corpus research of ESP text (Flowerdew, 2015), and suggest some ways in which ELF and learner language differ and overlap with reference to the domain of scholarly writing.

References

- Connor, Ulla & Anna Mauranen. 1999. Linguistic analyses of grant proposals: European Union grant proposals. *English for Specific Purposes* 18. 47-62.
- Connor, Ulla & Thom Upton. 2004. The genre of grant proposals: A corpus linguistic analysis. In Ulla Connor & Thom Upton (eds.), *Discourse in the Professions. Perspectives from corpus linguistics*, 235-255. Amsterdam: John Benjamins.
- Englander, Karen. 2014. *Writing and Publishing Science Research Papers in English: A global perspective*. New York: Springer.
- Feng, Haiying. 2008. A genre-based study of research grant proposals in China. In Ulla Connor, Ed Nagelhout & William Rozyski (eds.), *Contrastive Rhetoric. Reaching to intercultural rhetoric*, 63-86. Amsterdam: John Benjamins.
- Feng, Haiying & Shi, Liying. 2004. Genre analysis of research grant proposals. *LSP & Professional Communication* 4. 8-32.
- Flowerdew, Lynne. 2015. Learner corpora and language for academic and specific purposes. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 465-484. Cambridge: Cambridge University Press.
- Flowerdew, Lynne. 2016. A genre-inspired and lexico-grammatical approach for helping postgraduate students craft research grant proposals. *English for Specific Purposes* 42. 1-12.
- Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7-24.
- Hyland, Ken. 2009. *Academic Discourse: English in a global context*. London: Continuum.
- Lin, Ling & Stephen Evans. 2012. Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes* 31(3). 150-160.

- Mauranen, Anna. 2011. Learners and users [336?] Who do we want corpus data from? In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (eds.), *A Taste for Corpora. In honour of Sylviane Granger*, 155-171. Amsterdam: John Benjamins.
- Matzler, Pascal Patrick 2014. *A genre analysis of English language research grant proposal abstracts in Chile*. University of Nottingham, UK: MA TESOL dissertation.
- Michigan Corpus of Upper-level Student Papers. 2009. The Regents of the University of Michigan. <http://micusp.elicorpora.info/> (15 November, 2015).
- Römer, Ute 2012. Corpora and teaching academic writing: Exploring the pedagogical potential of MICUSP. In James Thomas & Alex Boulton (eds.), *Input, Process and Product. Developments in Teaching and Language Corpora*, 70-82. Masaryk University Press: Brno, Czech Republic.
- Swales, John. 1990. *Genre Analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, John. 2004. *Research Genres: Explorations and applications*. Cambridge: Cambridge University Press.
- Tardy, Christine. 2011. ESP and multi-method approaches to genre analysis. In Diane Belcher, Ann Margaret Johns & Brian Paltridge (eds.), *New Directions in English for Specific Purposes Research*, 145-173. Ann Arbor, MI: University of Michigan Press.

Hands-on use of corpora by trainee translators

Ana Frankenberg-Garcia
University of Surrey

With the exponential increase in availability of ready-to-use online corpus tools and resources over the past decade or so, the use of corpora is no longer restricted to a small community of researchers working on language description and natural language processing. Anyone with an internet connection is now able to access a good selection of monolingual and multilingual corpora and corpus software, which they can use to look up different types of questions about language, including questions for which there are no clear answers in dictionaries, grammars and other language resources. As widely acknowledged in the literature (see, for example, Bowker and Pearson 2002, Zanettin et al 2003 and Beeby et al. 2009), translators can benefit from this in many ways, from simply looking up parallel concordances to find out how other translators have dealt with similar translation problems before, to building ad hoc specialized language corpora to extract terminology and research phraseology in order to help them translate texts in subject-specific domains. Yet in contrast to the pressure that exists to train translators in the use of computer-assisted translation technologies, there seems to be little or no incentive to teach translators to use corpora. Despite the existence of general modules on Corpus Linguistics at a number of British Universities, there do not seem to be many modules that focus specifically on corpora as tools and resources for translation.

This paper discusses some of the challenges of training translators to use corpora, and then reports on how a group of 13 students studying for an MA in Translation

at the University of Surrey reacted to an eleven-week module on learning to use corpora in everyday translation. The analysis draws on (1) student responses to an anonymous end-of-semester questionnaire and (2) a corpus of graded assignments, where the students were required to write a report on their personal use of corpora (after having been asked to keep a three-month diary with concrete examples of corpus use). The corpus of student assignments was submitted to both a quantitative and a qualitative analysis. The quantitative analysis focuses on verifying the extent to which the students made reference to terms such as *concordance*, *lemma*, *collocation*, *part-of-speech tagging*, *normalized frequency* and so on, and the extent to which the queries described in the reports effectively involved the use of those concepts. The qualitative analysis details a representative selection of examples of how different students used corpora and of their opinions about it. These results were then triangulated with the student responses in the anonymous questionnaire.

References

- Beeby, A., Rodríguez-Inés, P. & Sánchez-Gijón, P. 2009. (eds.) *Corpus use and translating*. Amsterdam and Philadelphia: John Benjamins.
- Bowker, L. & Pearson, J. 2002. *Working with Specialized Language: a practical guide to using corpora*. London: Routledge.
- Zanettin, F., Bernardini, S. & Stewart, D. 2003. (eds.) *Corpora in Translator Education*. Manchester: St. Jerome.

Disagreement in L2 spoken English: From learner corpus research to corpus-based teaching materials

Dana Gablasova, Vaclav Brezina
Lancaster University

Introduction

In order to communicate successfully speakers need to master both linguistic (lexico-grammatical) and social aspects of the interaction. While lexico-grammatical features have received sufficient attention in traditional pedagogy, the more subtle social aspects of communication are often overlooked or not addressed adequately in available teaching materials (Malamed 2010). This study focuses on disagreeing in interactive oral communication by L2 speakers of English. It explores a specific construction for expressing disagreement, the so-called ‘agreement-plus-disagreement’ or ‘yes-but’ construction (Pomerantz 1984; Pekarek Doehler & Pochon-Berger 2011) and the techniques used to soften the impact of the disagreements on the communication. The study first uses a large corpus of spoken L2 English, the Trinity Lancaster Corpus, to investigate the communicative patterns of L2 speakers of three different proficiency levels when expressing disagreement. Next, the study illustrates how findings from a learner corpus can be used in designing activities and materials for teaching disagreement. In particular, the study answers the following two questions:

L2 proficiency	Sub-corpus size (words)	No of speakers
Lower-intermediate	460,012	597
Intermediate	573,443	581
Advanced	308,906	271
Total	1,342,361	1449

Table 1: Overview of the corpus used in the study

RQ1: Is there a difference in the disagreement strategies used by L2 speakers of three different proficiency levels with the ‘yes-but’ construction?

RQ2: How can the corpus findings be used in development of classroom materials to teach ‘disagreement strategies’?

Method

The data for this study were taken from the Trinity Lancaster Corpus (TLC) of L2 production which at present contains over three million running words. The corpus contains transcripts from the Graded Examinations of Spoken English (GESE) conducted by Trinity College London, a major international examination board (Trinity College London 2010). This study draws on a dialogic sub-corpus of the Trinity Lancaster Corpus which contains over 1M words and consists of spoken production from 1449 speakers engaged in two interactive dialogic tasks. The L2 speakers in this sub-corpus represent three proficiency levels: lower-intermediate (B1 level of CEFR), intermediate (B2) and advanced (C1 and C2) level of English proficiency. A more detailed overview of the sub-corpus can be seen in Table 1 below.

Results

The results revealed a considerable differences in how L2 English speakers of three proficiency level expressed the ‘yes-but’ type of disagreement. The results showed a very clear trend with the range and the number of mitigating markers (i.e. markers used to soften the negative impact of disagreement) rising steadily across the three proficiency groups with the advanced speakers using different mitigating techniques (e.g. lexical downtoners and delay/hesitation markers) frequently and combining them for greater effect. These findings contribute to the increasing body of evidence (Pekarek Doehler & Pochon-Berger 2011; Bardovi-Harlig & Salsbury 2004) that shows the connection between proficiency and the use of progressively more complex strategies intended to minimise the disruptive nature of disagreement. These findings, along with examples from the corpus, are then used in designing classroom materials that demonstrate different disagreement mitigating strategies and raise learners’ awareness of potential social impact of their linguistic choices.

References

- Bardovi-Harlig, Kathleen, & Salsbury, Tom. 2004. The organization of turns in the disagreements of L2 learners: A longitudinal perspective. In: Boxer, Diana & Cohen, Andrew (eds): *Studying speaking to inform second language learning*. Clevedon: Multilingual Matters (pp.199-227).
- Malamed, Lewis. 2010. Disagreement: How to disagree agreeably. In: Alicia Martínez Flor; Esther Usó-Juan (eds): *Speech act performance: theoretical, empirical and methodological issues*. Amsterdam: John Benjamins (pp. 199-215).
- Pekarek Doehler, Simona & Pochon-Berger, Evelyne. 2011. Developing “methods” for interaction: a cross-sectional study of disagreement sequences in French L2. In: Kelly Hall, Joan, Hellerman, John & Pekarek Doehler, Simona. (eds): *L2 Interactional Competence and Development* (pp. 206-243).
- Pomerantz, Anita. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shaped. In: Atkinson, J. Maxwell & Heritage, John. *Structures of Social Action*. Cambridge: Cambridge University Press (pp. 57-101).
- Trinity College London. 2010. Graded examinations in spoken English—Syllabus from 1 February 2010.

Anticipatory *it* patterns and rhetorical moves in the critique genre family

James Robert Garner

Georgia State University

Researchers in the fields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP) have become increasingly interested in the lexicogrammatical patterns that characterize certain written genres. Regardless of the approach taken in identifying and classifying these items (e.g. Pattern Grammar, lexical bundles), research has shown that their use varies across both different academic disciplines and different genres (Biber, Conrad, & Cortes, 2004; Cortes, 2004; Durrant, 2015; Hewings & Hewings, 2002; Hyland, 2008a, 2008b). In addition, several studies have identified the ways in which these patterns are used to structure the rhetorical moves of a genre (Ädel, 2014; Cortes, 2013; Römer, 2010). The current study aims to add to this body of literature by examining the use of two anticipatory *it* patterns, *It is (ADV) ADJ to-inf* and *It is (ADV) ADJ that*, in critique essays, a common pedagogical genre at the university level.

This study is based on 221 critique essays (477,127 words) from the British Academic Written English corpus (BAWE; Nesi, 2010). These essays were written by students from all four broad discipline groups in BAWE (Arts & Humanities, Social Sciences, Life Sciences, Physical Sciences). Once all iterations of the *It is (ADV) ADJ to-inf* and *It is (ADV) ADJ that* patterns were extracted, they were subjected to both semantic and rhetorical analyses. For the semantic analysis, each sequence was placed into one of six semantic categories according to a semantic

classification scheme used in previous pattern studies (Francis, Manning, & Hunston, 1998; Groom, 2005). For the rhetorical analysis, each sequence was further classified into categories according to, firstly, the move or step of the genre that the pattern appeared in and, secondly, whether the pattern occurred move initially or move finally. These analyses led to the creation of a three-way classification system (semantic, move, move initial or final) with the aim of revealing how different semantic realizations of anticipatory *it* patterns are used by British university students to structure critique essays in their disciplinary courses.

The findings show that three semantic categories are used most frequently in the student critiques: ***It is DIFFICULTY to-inf*** (e.g. *it is difficult to, it is possible to*), ***It is IMPORTANCE to-inf*** (e.g. *it is necessary to, it is important to*), and ***It is VALIDITY that*** (e.g. *it is clear that, it is likely that*). These patterns were found to vary in frequency across academic discipline groups, specifically between the harder sciences of Life and Physical Sciences and Arts and Humanities and Social Sciences. They were also found to vary in frequency across different rhetorical moves. For example, ***It is IMPORTANCE to-inf*** occurred most often in descriptive moves, while ***It is VALIDITY that*** most often occurred in evaluative moves. All three patterns were also found to occur most often move-initially. These findings are discussed in terms of their implications for the teaching of phraseological patterns in genre-based EAP courses.

References

- Ädel, A. 2014. Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes*. 16. 68-80.
- Biber, D., Conrad, S., & Cortes, V. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*. 25(3). 371-405.
- Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*. 23. 397-423.
- Cortes, V. 2013. *The purpose of this study is to*: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*. 12. 33-43.
- Durrant, P. 2015. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, Advance Access published April 19, 2015. doi: 10.1093/applin/amv011
- Francis, G., Manning, E., & Hunston, S. 1998. *Colling COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.
- Groom, N. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*. 4. 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. 367-383.
- Hyland, K. 2008a. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*. 18(1). 41-62.

- Hyland, K. 2008b. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*. 27. 4-21.
- Nesi, H. 2010. BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, G. Aston & L. Flowerdew (Eds.), *New Trends in Corpora and Language Learning*, 213-228. London: Continuum.
- Römer, U. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Constructions*. 3(1). 95-119.
-

The corpus as a writing tool in first-year composition

James Robert Garner

Georgia State University

One area of language pedagogy that has been heavily impacted by language corpora is English for Academic Purposes (EAP). In addition to research focusing on lexicogrammatical features that characterize academic discourse, research has investigated how direct corpus use by EAP students may enhance their acquisition of academic literacy (C. Yoon, 2011). Corpus instruction in the EAP classroom has generally been implemented in one of two ways. First, corpora have been utilized as research tools, allowing students the opportunity to investigate the rules and regularities that define their chosen discipline or target genre (Charles, 2012; Cresswell, 2007; Lee & Swales, 2006). The second approach focuses on training learners in the use of corpus tools to address problems in their writing (Chang, 2014; Kennedy & Miceli, 2001; O'Sullivan & Chambers, 2006; H. Yoon, 2008). Results from research on both approaches have shown that, given adequate time and training, students can make good use of corpora and apply findings generated from them to their own writing. Furthermore, analysis of learner perceptions has shown that the majority of EAP students find this approach helpful for their writing and not too difficult to use.

However, within the ESL context, most studies of corpus use by EAP students have focused on graduate students, with little to no work done addressing undergraduate students in first-year composition courses. This type of course, which is common in universities across the United States, is one in which corpus consultation could have benefits for students. The use of corpus tools in this type of class could assist students in acquiring a general level of familiarity with the conventions of academic writing at the university level before moving on to their major courses.

The current study attempts to address this gap in the literature by investigating student perceptions of corpus use in the first-year composition classroom. Native-speaking, bilingual, and ESL students in two first-year composition courses at an American university were introduced to corpus tools and trained in their use. Corpus instruction focused on features relevant to the genres the students were completing as well as methods for using corpora to address lexicogrammatical issues in their writing. Student perceptions of corpus use were obtained through the analysis of semester-long student reflective journals, an open-ended questionnaire distributed at the end of the semester, and in-depth interviews conducted following the semester.

Preliminary results show that most students in the study, regardless of language background, found corpus use helpful in writing and revising their essays. Specifically, they claimed that corpus use was most beneficial to their choice of appropriate academic vocabulary. Students also indicated that they intend to use corpus tools in the future and would recommend them for other first-year composition students. This is in spite of challenges they reported facing, such as technical issues and learning the process of corpus searches and concordance analysis. The full results will be discussed in terms of their implications for corpus integration into general EAP courses at the undergraduate level.

References

- Chang, J. 2014. The use of general and specialized corpora as a reference source for academic English writing: A case study. *ReCALL*. 26(2). 243-259.
- Charles, M. 2012. 'Proper Vocabulary and Juicy Collocations': EAP Students Evaluate Do-It-Yourself Corpus Building. *English for Specific Purposes*. 31(1). 93-102.
- Cresswell, A. 2007. Getting to "know" connectors? Evaluating data-driven learning in a writing skills course. In E. Hildalgo, L. Quereda & S. Juan (Eds.), *Corpora in the Foreign Language Classroom*, 267-287. Amsterdam, Netherlands: Rodopi.
- Kennedy, C., & Miceli, T. 2001. An evaluation of intermediate students' approaches to corpus investigation. *Language Learning and Technology*. 5(3). 77-90.
- 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.
- O'Sullivan, I., & Chambers, A. 2006. Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*. 15(1). 49-68.
- Yoon, C. 2011. Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*. 10(3). 130-139.
- Yoon, H. 2008. More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning and Technology*. 12(2). 31-48.

Language for specific purposes corpora and tailor made concordancer: (Semi-) big data corpora and flexible open source software for writing centres

Tobias Gärtner

Leibniz Universität Hannover

Pedagogical implementations of corpus linguistics in Language for Specific Purposes (LSP) contexts require two adjustments: first the texts have to be relevant for the user and the concordancer has to reflect the user's skill. For the engineering students in their writing and language centres the Leibniz University of Hannover, Germany, and the Peter the Great St. Petersburg Polytechnic University, Russia, have compiled a 60 million token corpus containing several hundred PhD and Master

dissertations in the respective fields written in English, Russian and German. The *Deutsch, English and Russkii (DEaR) - Corpus of Civil Engineering* is accompanied by a tailor made concordancing software. The *Hannover Concordancer* (HanConc) provides the latest algorithms and metrics from corpus and computational linguistics, data and text mining programmed with easily editable code (R and Shiny) and is equipped with a user friendly Graphical User Interface (GUI). The queries as well as the results are integrated in a web based interface. Hardly any meta language is required to search for word clusters of up to five items. Advanced queries can be formulated using strings, i.e. words, lemmata and part-of-speech tags. A certain order of these items can be prescribed. HanConc is also capable of scaling the results from, for instance, collocations presented as word clouds to cosine values of associations based on a Latent Semantic Analysis (LSA). So far the compilation of the corpora and the processing of the files has been completed. HanConc and the respective corpora have reached a beta phase and are tested by students and lecturers at both universities. First results indicate that further adjustments are required. While one lecturer needs a sorting algorithm and regular expressions for her course, a second one only uses multi-item Key Words in Context (KWIC). Currently HanConc is altered for their specific needs. Later versions of HanConc will include features such as machine learning algorithms and a dynamic interface to allow for ad hoc adjustments. Thereby users will no longer need a programmer but HanConc adjusts to the way students and lecturers use it. This presentation will report and demonstrate the current state of development. It will introduce the corpora with their respective characteristics and the latest HanConc Beta version.

Speech rate revisited – the effect of task design on speech rate

Tomáš Gráf

Charles University

Speech rate is generally considered one of the most robust components of fluency and a strong predictor of perceived L2 fluency (Kormos and Dénes 2004, Bosker et al. 2012). Consequently it receives regular attention in studies analysing learner language (eg. Götz 2013). One of its most characteristic features is its variability as a result of planning pressure, personal traits, context and genre (Tauroza and Allison 1990), gender (Whiteside 1996), age (Ramig 1983), emotional state and the stress level (Hausner 1987), and last but not least the nature of the task the speaker is performing (Foster & Skehan 1996). It is the aim of this presentation to show to what extent speech rate varies in three different tasks performed by the same speakers, and present a technique which can be reliably used for obtaining speech rate measurements.

Method

The data for the study come from the Czech subcorpus of LINDSEI (Gilquin et al. 2010) which comprises 50 15-minute interviews with advanced speakers of English,

Corpus	Task 1 (tokens)	Task 2 (tokens)	Task 3 (tokens)	Total
LINDSEI_CZ	40,584	42,850	12,535	95,969
LOCNEC	44,320	70,650	7,244	122,214

Table 1: Numbers of tokens in the three tasks.

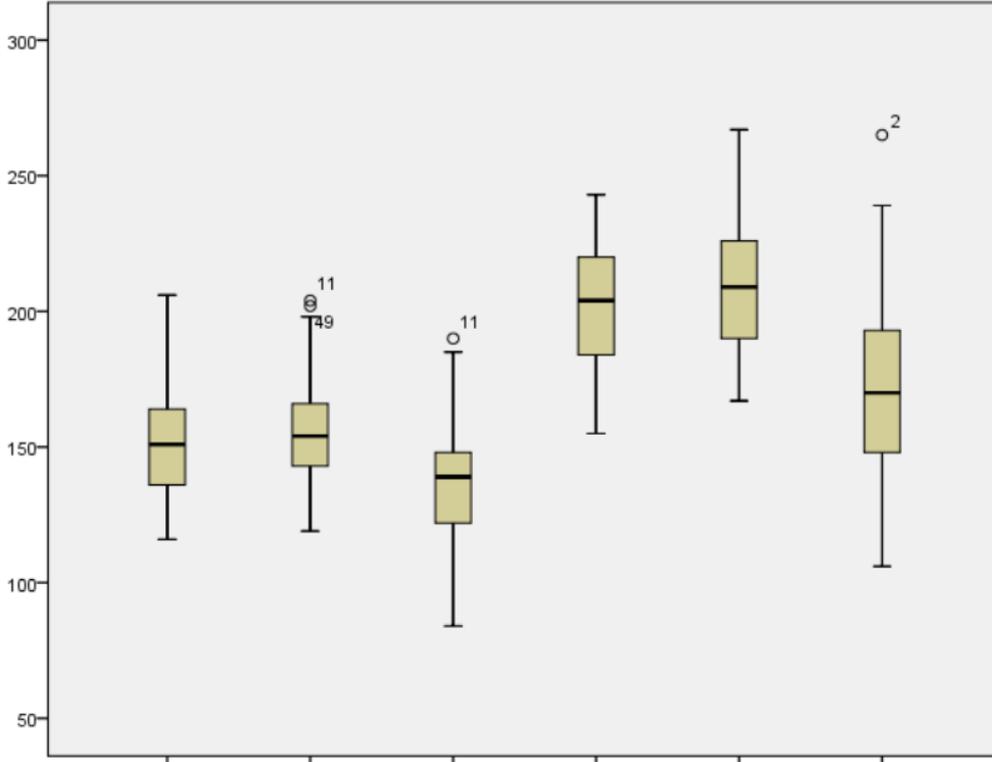


Figure 1: Boxplots showing non-native (L^*) and native (N^*) speech rates for each task separately (*T1—3). The y axis marks speech rate in words per minute (wpm)

and for comparison, the parallel corpus of native speaker conversations LOCNEC. Each interview contained three tasks – a monologue, a dialogue and a picture-based story reconstruction. Speech rate was measured in words per minute (wpm) for the three tasks separately, counting unpruned words, but excluding long periods of silence and non-verbal sounds. Table 1 provides a basic description of the two corpora. The SR measurements were then compared in the two corpora independently using ANOVA and t-tests.

Results and discussion

The learners produced a mean speech rate of 152wpm (SD = 20.97) for Task 1, 157wpm (SD = 19.72) for Task 2, and 138wpm (SD = 22.09) for Task 3. In comparison, the native speakers produced on average 203wpm (SD = 23.51) in Task 1, 210wpm (SD = 24.53) in Task 2 and 174wpm (SD = 34.49) in Task 3. For comparison, see the boxplots in Fig. 1.

T-tests showed significant differences ($p < .05$ for T1, and $p < .0005$) for the

learners between Tasks 1 and 2 ($p < .05$), 1 and 3 ($p < .0005$), and 2 and 3 ($p < .0005$). Significant differences were found for the native speakers between Tasks 1 and 2 ($p < .01$), 1 and 3 ($p < .0005$), and 2 and 3 ($p < .0005$). The results suggest that task design has an effect on speech rate and further indicate that picture description tasks are cognitively considerably difficult, and that monologues require more planning time than informal conversations. It could, however, be argued that the performance in tasks 1 and 2 could also be affected by the nature of the topic under discussion.

Conclusion

The results show that both non-native and native English speakers' speech rates are affected by task design and that especially picture description tasks appear to be most taxing in this respect. This could have implications on assessment, where picture tasks are frequently deployed.

References

- Bosker, H. R., A.-F. Pinget, H. Quene, T. Sanders and N. H. de Jong (2013) "What Makes Speech Sound Fluent? The Contributions of Pauses, Speed and Repairs." *Language Testing* 30 (2): 159–75.
- Foster, P. and Skehan, P. (1996). "The influence of planning and task type on second language performance." *Studies in Second Language Acquisition*, 18, 299–323.
- Gilquin, G., De Cock, S. and Granger, S. 2010. *Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Götz, S. (2013) *Fluency in Native and Nonnative English Speech*. Studies in Corpus Linguistics, volume 53. Amsterdam; Philadelphia: John Benjamins Publishing Company.
- Hausner, M. (1987) "Sprechgeschwindigkeit Als Eine Funktion von Stress: Eine Fallstudie." *Language and Style* 20: 285–311.
- Kormos, J. and M. Dénes (2004) "Exploring Measures and Perceptions of Fluency in the Speech of Second Language Learners." *System* 32 (2): 145–64. doi: 10.1016/j.system.2004.01.001.
- Ramig, Lorraine A. (1983) "Effects of Physiological Aging on Speaking and Reading Rates." *Journal of Communication Disorders* 16 (3): 217–26.
- Whiteside, Sandra P. (1996) "Temporal-Based Acoustic-Phonetic Patterns in Read Speech: Some Evidence for Speaker Sex Differences." *Journal of the International Phonetic Association* 26 (01): 23.
-

Building a large scale corpus of academic written English in a Brazilian university context: An academic learner corpus

Annallena de Souza Guedes, Bárbara Malveira Orfanò, Jessica Ceritello Alves
Federal University of Minas Gerais

In the past few years, we are witnessing the growth of international exchange programmes, in Brazilian universities, in a pace we have never seen before. This process has been contributing immensely to enhance the teaching and learning of English in an academic context. *Corpus Linguistics*, in particular, *Learner Corpus Studies* (Granger, 2002; Granger *et al.*, 2009) and its application to second language acquisition has been playing an important role in this scenario. Granger (1998) has already stated the importance of *Learner Corpora* to a better understanding of learner's interlanguage. Thus, this paper sets out to describe an on-going corpus: *Corpus of English for Academic Purposes (CorIFA)* being developed at the Federal University of Minas Gerais (UFMG)/Brazil bearing in mind its potential to English teaching. The main aim is to present the rationale and the methodology behind building a multidisciplinary academic corpus comprised by undergraduate and postgraduate written productions in an *English for Academic Purposes* course in Brazil. This learner corpus has been being compiled since 2013 and it is formed by five different genres: statement of purpose, argumentative essay, report, literature review and research paper. Each genre requires a specific number of words ranging from 200 to 1,000 words which also depends on students level, for example, pre-intermediate students produce texts belonging to the statement of purpose and argumentative essay genre, intermediate students produce texts belonging to the report genre, while advanced students are required to produce texts belonging to both literature review and research article genre. The compilation process consists of students sending their written production by e-mail and then, teachers organising these writings in files that are converted into *txt*.format in order to be further included in the corpus. Other institutions from different parts of Brazil are collaborating to both increase the size of the corpus and for its representativeness. The description phase is a fundamental part of the whole process. Information such as students specific course, level, type of task and students academic year are crucial for identifying, for example, rhetorical features of specific disciplines and also for creating conditions for longitudinal studies. Building a corpus is far from being an easy task. We will only fully understand and meet our students' needs if we go deep on their interlanguage and compiling a *Learner Corpus* has proved to be an efficient way to pursue this goal. In this paper, we give an in depth description of the process of compiling the *CorIFA* and further investigate linguistic features found in the texts produced by our informants. It is believed that such an analysis can bring interesting insights to the teaching and learning of English in an academic context.

References:

- Granger, Sylviane (Ed.). 1998. *Learner English on Computer*. London & New York: Addison Wesley Longman.
- Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In: Granger, Sylviane; Hung, J; Petch-Tyson, S. (Eds.). *Computer Learner corpora, second*

- language acquisition and foreign language teaching*, 3-33. Amsterdam: John Benjamins Publishing Company.
- Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny & Paquot, Magali. 2009. *The International Corpus of Learner English. Version 2 Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Meunier, Fanny & Sylviane Granger. (Ed.). 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.
-

Extensive reading, false-beginners, and vocabulary development: Investigating the impact of data-driven learning

Gregory Hadley¹, Maggie Charles²

¹Niigata University, ²Oxford University

Data-Driven Learning (DDL), developed in the 1990s by Johns (Johns, 1991), has been shown in numerous studies to be effective among advanced and intermediate learners (Braun, 2007; Charles, 2012; Granger, Hung, & Petch-Tyson, 2002; Sun & Wang, 2003). However, it has had only limited impact among “false beginners”, although small-scale studies have suggested that, with significant levels of scaffolding, DDL has the potential for positively enriching their second language learning experience (Boulton, 2009; Hadley, 2002; St. John, 2001). To date, the linguistic difficulty of currently available corpora has been a major barrier that has precluded these lower-level learners from truly embracing the full potential of this form of second language learning.

We report here on the first iteration of an ongoing research project into the use of DDL in an extensive reading program at a Japanese university. The participants were Japanese, French, Chinese and Korean “false-beginners”, defined here as students of A2 CEFR proficiency level, who occasionally exhibit borderline B1 aspects in certain situations. The corpus was developed by Oxford University Press from their Bookworms Graded Readers, thus ensuring that the data presented to students was at an appropriate linguistic level. An experimental group of 13 students used DDL materials based on the Bookworms corpus, while a control group of 11 students had no DDL input. Both groups read extensively (a minimum of 200,000 words over 10 weeks), completed book reports and vocabulary logs, and participated in similar in-class tasks. The aim of the study was to ascertain whether the use of DDL materials would lead to enhanced vocabulary knowledge and English proficiency in the experimental group. A pre-test/post-test experimental design was employed, using Nation & Beglar's (2007) Vocabulary Levels Test and a C-test (Klein-Braley & Raatz, 1984) constructed from an upper-level Bookworms reader. The results of the pre-test indicated that the two groups were essentially the same statistically and post-test results showed a statistically significant improvement for both groups. However, the control group improved more, as indicated by the C-test results.

We examine some possible reasons that could account for these results, based upon a study of student attitudes and constructs through the use of Personal Construct Repertory Grids (Hadley & Evans, 2001; Jankowicz, 2004; Marsden & Littler, 2000). We conclude by discussing what the further steps we are taking to enhance language learning among false beginners through the use of DDL.

References

- Boulton, Alex. 2009. Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics* 35(1). 81-106.
- Braun, Sabine. 2007. Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19(03). 307-328.
- Charles, Maggie. 2012. "Proper vocabulary and juicy collocations": EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes* 31(2). 93-124.
- Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.). 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam; Philadelphia: John Benjamins.
- Hadley, Gregory. 2002. An introduction to data-driven learning. *RELC Journal* 33(2). 99-124.
- Hadley, Gregory & Chris Evans. 2001. Constructions across a cultural gap. In Julian Edge (ed.), *Action Research*, 129-143. Alexandria, Virginia: TESOL.
- Jankowicz, Devi. 2004. *The easy guide to repertory grids*. Chichester, West Sussex, England: Wiley.
- Johns, Tim. 1991. Should you be persuaded - Two examples of data-driven learning materials. *English Language Research Journal* 4. 1-16.
- Klein-Braley, Christine & Ulrich Raatz. 1984. A survey of research on the C-Test. *Language Testing* 1. 134-146.
- Marsden, David & Dale Littler. 2000. Repertory grid technique - An interpretive research framework. *European Journal of Marketing* 34. 816-834.
- Nation, Paul & David Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7). 9-13.
- St. John, Elke. 2001. A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning and Technology* 5(3). 185-203.
- Sun, Yu-Chih & Li-Yuch Wang. 2003. Concordancers in the EFL classroom: cognitive approaches and collocation difficulty. *Computer Assisted Language Learning* 16(1). 83-94.

Using a corpus-based collocation network explorer to detect/correct L2 learners' collocations

Ping-Yu Huang¹, Nai-Lung Tsao²

¹Ming Chi University of Technology, ²Tamkang University

In this paper we present an innovative approach for representing collocational knowledge for English L2 instruction/learning. Specifically, we demonstrate and evaluate an automated English collocation network explorer, *NetCollo*, which is meant for displaying semantically-related word combinations within networks. Using *NetCollo*, users can: test whether a combination is possible in English (i.e. *say truth*), get suggestions for miscollocations (e.g. *tell truth*), learn other semantically-related correct combinations (e.g. *state fact*), and explore other semantically-related miscollocations which also need to be avoided (e.g. *tell fact*).

The notions of collocation networks or intercollocability were first proposed by Cowie & Howarth (1996). According to them, overlapping collocations taking similar meanings and shared collocates constitute the most difficult collocation type for L2 learning. Learners very often fail to figure out which combinations are possible (e.g. *achieve purpose* and *attain objective*) and which are not (e.g. *attain purpose*) in their target language(s). To our knowledge, Liu, et al. (2009) was the only research attempting to investigate intercollocability, and applied it to miscollocation correction. By identifying overlapping collocations for a miscollocation (e.g. noun collocates of *attain* and verb collocates of *purpose*), Liu, et al. demonstrated that collocation networks were as effective in miscollocation correction as word association or semantic similarity measures. In this study, we utilize a more sophisticated and effective approach to construct collocation networks, and further develop an online tool, *NetCollo*, readily available for English instructors/learners. As a word pair is keyed in, *NetCollo* begins to automatically search English L1 corpora to identify collocates for its elements (e.g. *standard/goal/level/objective* for *attain* and *achieve/state/define/accomplish* for *purpose*). Next, the two word groups are filtered with only the ones which “share” the most collocates with the searched words being left. This filtering process results in a complete and manageable network. The words embedded in such networks, then, are further ranked in order of semantic relevance; that is, the more semantically related to the searched words, the higher ranking (e.g. *achieve/accomplish* for *attain* and *goal* for *purpose*).

To evaluate *NetCollo*'s effectiveness of detecting/correcting miscollocations, we tested it on a total of 168 L2 verb-noun errors reported in Liu (2002). What we intended to know was whether *NetCollo* could detect those errors and, if it could, suggest correct alternative usages. The results collected were rather satisfactory. Among the errors, 153 were found to be miscollocations by *NetCollo*, achieving a high precision ratio of 91.1%. Concerning correction suggestions, the mean reciprocal rank (MRR) that *NetCollo* performed was 0.65, an average comparable to or better than those of previous studies (e.g. 0.66 in Chang, et al., 2008, and 0.518 in Wu, et al., 2010). By considering both intercollocability and semantic relevance, *NetCollo* demonstrates the capability of effectively detecting L2 learners' miscollocations, and providing learners with alternatives which are valid in language and adequate in context.

We conclude our paper by discussing future improvements of *NetCollo*, such as establishing networks based on extremely large textual databases, which contain around 10 billion tokens, to improve candidate collocation extraction.

References

- Chang, Yu-Chia, Jason S. Chang, Hao-Jan Chen & Hsien-Chin Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning* 21, 283-299.
- Cowie, Anthony P. & Peter Howarth. 1996. Phraseological competence and written proficiency. *British Studies in Applied Linguistics* 11, 90-93.
- Liu, Anne Li-Er. (2002). A corpus-based lexical semantic investigation of verb-noun

miscollations in Taiwan learners' English. Taipei: The Tamkang University thesis.

Liu, Anne Li-Er, David Wible & Nai-Lung Tsao. 2009. Automated suggestions for miscollations. *Workshop on Innovative Use of NLP for Building Educational Applications (BEA Workshop)* 4. 47-50.

Wu, Jian-Cheng, Yu-Chia Chang, Teruko Mitamura & Jason S. Chang. 2010. Automatic collocation suggestion in academic writing. *Annual Meeting of the Association for Computational Linguistics (ACL)* 48. 115-119.

Get in line and show me what you've got: Exploring concordance line ranking and selection methods for different language learning goals.

Stephen Jeaco

Xi'an Jiaotong-Liverpool University

Corpora have increased in size over the last few decades and it has been true for a long time that experts making use of concordancers for research or lexicography cannot usually examine in detail every concordance line for a particular query. Approaches to this problem include procedures for sampling concordance lines in a systematic way such as through a retrieval-analysis cycle (Sinclair, 1991), and various means of ranking and selection including use of collocations, the position of repeated elements across sets of concordance lines, the length of sentence and position of the node, and the presence or absence of specific words, symbols or low frequency items (e.g. Collier, 1999; Kilgarriff, Husak, McAdam, Rundell, & Rychlý, 2008). Ranking methods can be a powerful tool in the hands of researchers, and within many software tools it is possible to easily switch between various sorting methods. Expert users are also likely to have some awareness of how well each method might help facilitate their work, while keeping its limitations in mind. However, when concordancers are used by language learners or language teachers who may not have the skills or patience to explore dozens of results with careful manipulation of the ranking method, the advantages and disadvantages of different ranking and selection methods are of even greater importance. Two of the primary aims of using concordancers with language learners are likely to be based on Second Language Acquisition principles: that learners should be exposed to target language in use (Krashen, 1989); and that "intake is what learners consciously notice" (Schmidt, 1990, p. 149). Having the learners themselves make discoveries about patterns in language use is considered important in fields of materials evaluation (Bolitho et al., 2003; Tomlinson, 1994, 2008) and in Data Driven Learning (Bernardini, 2004; Johns, 1991). It is obvious that different ranking methods will make different kinds of patterns more or less obvious. In language learning contexts, it is also important to consider the different uses students and teachers may make of the concordance line results. Frankenberg-Garcia (2012), for example, has considered how the evaluation of examples in dictionaries or concordance lines aiming to aid *comprehension* may differ from the evaluation of examples to aid *production*. As an interactive tool in the classroom, the concordance line results can also be evaluated in terms of their suitability for a specific writing topic.

The purpose of this paper is to present results from language teacher and student evaluations of concordance lines for sets of words and phrases presented using different ranking systems. A concordancer which was designed specifically for English language learning is *The Prime Machine* (Jeaco, 2015), and it has several different concordance line ranking methods including those drawing on procedures outlined by Collier (1999) and Kilgarriff et al. (2008), as well as new approaches based on the node's role in lexical cohesion in its text (see Hoey, 1991) and methods based on combining Collier's system with collocation measures. Three orientations form the basis for the judgements of the output from the software for what would be the first page of concordance line results: general comprehension of the term, productive use of the term for a specific task and productive use of the term more generally.

References

- Bernardini, Silvia 2004. Corpora in the classroom: An overview and some reflections on future developments. In Sinclair, John McHardy (ed.). *How to use corpora in language teaching* (pp. 15-36). Amsterdam: John Benjamins.
- Bolitho, Rod, Carter, Ronald, Hughes, Rebecca, Ivanič, Roz, Masuhara, Hitomi, & Tomlinson, Brian 2003. Ten questions about Language Awareness. *ELT Journal*, 57(3), 251-259.
- Collier, Alex 1999. *The automatic selection of concordance lines*. Liverpool, UK: University of Liverpool Unpublished Ph.D. dissertation.
- Frankenberg-Garcia, Ana 2012. Getting help from corpus examples. Paper presented at the TaLC10 (Teaching and Language Corpora) Conference, Warsaw.
- Hoey, Michael 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.
- Jeaco, Stephen 2015. *The Prime Machine: A user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language*. Liverpool, UK: University of Liverpool Unpublished Ph.D. dissertation. Retrieved from <http://repository.liv.ac.uk/id/eprint/2014579>
- Johns, Tim 1991. Should you be persuaded: Two samples of data-driven learning materials. In Johns, Tim & King, Philip (eds.). *Classroom concordancing* (Vol. 4, pp. 1-13). Birmingham: Centre for English Language Studies, University of Birmingham.
- Kilgarriff, Adam, Husák, Miloš, McAdam, Katy, Rundell, Michael, & Rychlý, Pavel 2008. GDEX: Automatically finding good dictionary examples in a corpus. Paper presented at the Euralex (European Association for Lexicography), Barcelona.
- Krashen, Stephen 1989. We acquire vocabulary and spelling by reading: Additional evidence for the Input Hypothesis. *The Modern Language Journal*, 73(iv), 440-464.
- Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Sinclair, John McHardy 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

- Tomlinson, Brian 1994. Pragmatic awareness activities. *Language Awareness*, 3(3-4), 119-129.
- Tomlinson, Brian 2008. Language acquisition and language learning materials. In Tomlinson, Brian (ed.), *English language learning materials: A critical review* (pp. 3-13). London: Bloomsbury Publishing.
-

An open educational resource: The SOURCe project

Fryni Kakoyianni-Doa, Eleni Tziafa

University of Cyprus

Our aim is to describe the properties and functionalities of the SOURCe project, which includes the search engine for the Searchable Online French-Greek Parallel Corpus for the University of Cyprus (SOURCe) (Kakoyianni-Doa & Tziafa 2013), the Pencil (an alignment tool) (Kakoyianni-Doa et al. 2013), the Synonyms and the Library tools. These are designed as freely available resources for language processing, along with the data to be processed, in usable formats for teachers and learners. Moreover, we will outline its future perspectives and applications, discussing how it can be incorporated into effective learning resources. We will focus on the construction and composition of the SOURCe Project, based on a parallel corpus, the content, annotation, encoding and availability of which are meant to serve the needs of teachers and students of French as a foreign language and also to facilitate future linguistic research. This corpus linguistics approach is undertaken from the perspective of language learning and translation studies. This project is led by Fryni Kakoyianni-Doa and is fully funded by the University of Cyprus.

The core of the project is a collection of parallel corpora, aligned (at sentence level) original and translated texts (Sinclair & Ball 1996), in French and Greek. Research has shown that the use of corpora in the classroom can have remarkable results as regards foreign language learning (Hadley 2002; Landure & Boulton 2010). Moreover, as Kilgariff (2009) suggests, parallel corpora are easier to be disguised as dictionaries and be brought for use in classroom.

In order to support the use of corpus linguistic tools by a diverse range of teachers and learners with no previous experience, we designed a simple interface, through which the user may search existing corpora, upload texts, and see them online. We included different registers (Biber 1993), so that users may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and journalistic language). Commonly used parallel corpora like EUROPARL (Koehn 2005), the JRC Acquis corpus (Steinberger et al. 2006) and other corpora from the Opus open parallel corpus (Tiedemann 2012) were also used, as were literary works available from Project Gutenberg.

To overcome the well known problem of the existing NLP tools and resources not finding their way into the language learning classroom, despite their obvious potential uses in language learning, the main goal of the proposed project is to provide language instructors and learners with ready-made corpora and corpus-based exercises, available for use in a new learning environment. This project is thus intended not only to fill a gap in the literature on corpora used in classroom, but also to make available valuable resources, especially for a less resourced language

such as Greek. Our objective is to develop a knowledge base that will assist teachers to find out about, adapt and apply existing tools.

References

- Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4. 243-257.
- Hadley, Gregory. 2002. Sensing the winds of change: An introduction to data-driven learning. *RELC Journal*, 33(2). 99-124.
- Kilgarriff, Adam. 2009. Corpora in the classroom without scaring the students. *Proceedings of the 18th International Symposium on English Teaching and Learning in the Republic of China*. Taipei: National Taiwan Normal University. <http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc>(14January,2016).
- Kakoyianni-Doa, Fryni & Tziafa, Eleni. 2013. Source: Building a searchable online French Greek parallel corpus for the University of Cyprus. *Revista Nebrija de Lingüística Aplicada* 13 (número especial).
- Kakoyianni-Doa, Fryni, Antaris, Stefanos & Tziafa, Eleni. 2013. A free online parallel corpus construction tool for language teachers and learners. *Procedia - Social and Behavioral Sciences*. Elsevier, 95. 535-541.
- Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand. 79-86.
- Landure, Corinne & Boulton, Alex. 2010. Corpus et autocorrection pour l'apprentissage des langues. *ASp*, 57. 11-30.
- Steinberger, Ralf, Pouliquen, Bruno, Widiger, Anna, Ignat, Camelia, Erjavec, Tomaz, Tufis, Dan & Varga, Daniel. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *5th Language Resources and Evaluation Conference (LREC)*, Genoa, Italy. 2142-2147.
- Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. *8th International Conference on Language Resources and Evaluation (LREC)*. 2214-2218. http://lrec.elra.info/proceedings/lrec2012/pdf/463_Paper.pdf (14 January, 2016).

Word semantics in terms of suffix combinability: L2 acquisition and a specialized electronic corpus

Stela Manova¹, Bartosz Brzoza²

¹University of Vienna, ² Adam Mickiewicz University Poznań

This paper discusses the semantics of complex words of the type BASE-SUFFIX1-SUFFIX2 with data from Polish (e.g., *pis-ar_N-ski_{ADJ}* 'writer's') and suggests strategies for facilitation of L2 vocabulary learning and for creation and automatic annotation of a specialized corpus of word-formation.

The semantics of complex words is modeled through the combinability of the suffixes they contain with the help of the findings of the so-called cognitive approach (Manova 2011) that is usage-based and follows a distributional logic: SUFF1

suffixes followed by the same SUFF2 suffixes derive the same semantics, SUFF1 suffixes with different combinability derive different semantics. With data from a large dictionary (Saloni et al. 2007) and the National Corpus of Polish (Przepiórkowski et al. 2012), it will be demonstrated that SUFF1 and SUFF2 relate in specific ways and all their combinations are either fixed or predictable. Fixed combinations are those in which SUFF1 is always followed by only one SUFF2 of a major lexical category (noun, adjective, verb). In a predictable combination, SUFF1 is followed by more than one SUFF2 of a lexical category but either one of the SUFF2 suffixes dominates over the others, i.e. it derives a great number of types, whereas all other SUFF2 suffixes derive a very limited number of types; or different SUFF2 suffixes of the same lexical category derive different semantics, e.g. an object and an abstract noun. Intriguingly, only a few semantic concepts derive the various SUFF1-SUFF2 combinations in Polish word-formation. For example, for the derivation of complex nouns one has to consider only [+/-person], [+/-object], [+/-place], [+/-abstract]. Moreover, psycholinguistic research on the representation of suffix combinability in the mental lexicon provides evidence that native speakers of Polish know SUFF1-SUFF2 combinations without lexical bases by heart (Manova & Brzoza 2015). We suggest that L2 learners could profit from native speakers' strategies for language processing and will, with a large set of suffixes, demonstrate how the findings of our research can be used for L2 vocabulary learning. We maintain that L2 pedagogical grammars should pay more attention to derivational suffixes and their combinations, especially to the productive ones. Pedagogical grammars should be enriched with generalizations about the semantics of complex words in terms of suffix combinations (Manova & Brzoza, submitted), e.g. that from complex nouns for places only adjectives can be derived (by the attachment of either *-owy* or *-(V)ny*, as in: *kawiar-ni-owy* 'café like', *pracow-ni-any* 'studio-'; *równ-in-owy* 'flatlands-like', *dol-in-ny* 'valley-like', etc.), the only exception being lexicalized action nouns; when used as places, such nouns (e.g., *mieszk-anie* 'living; flat') can serve as bases for derivation of persons (*mieszk-ani-ec* 'resident'). We will also discuss how the observation that the semantics of SUFF1-SUFF2 combinations can be modeled without reference to the semantics of the BASE can be used for creation and automatic annotation of a specialized corpus of word-formation, which will be an invaluable resource for educational and research purposes, especially if one considers that there does not exist an electronic corpus of Polish annotated at the level of morpheme.

References

- Manova, Stela & Mark Aronoff. 2010. Modeling Affix Order. *Morphology* 20: 109–131.
- Manova, Stela. 2011. A cognitive approach to SUFF1-SUFF2 combinations: A tribute to Carl Friedrich Gauss. *Word Structure* 4(2). 272-300.
- Manova, Stela & Bartosz Brzoza. 2015. Suffix combinability and the organization of the mental lexicon. Paper presented at the 10th Mediterranean Morphology Meeting (MMM), Haifa, Israel, 7-10 September 2015.
- Manova, Stela & Bartosz Brzoza. (submitted) Suffix ordering in Polish: Implications for foreign language learning.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski & Barbara Lewandowska-Tomaszczyk. 2012. *Narodowy korpus języka polskiego*. Warszawa: PWN.

Saloni, Zygmunt, Włodzimierz Gruszczyński, Marcin Woliński & Robert Wołosz. 2007. *Słownik gramatyczny języka polskiego*. Warszawa: Wiedza Powszechna.

Teaching language variation in a Spanish L2 class: Analyzing classroom interactions and students' perceptions

Nausica Marcos Miguel

Denison University

Teaching language variation is recommended in second language (L2) courses since this gives learners insights into the sociolinguistic and pragmatic dimensions of an L2 (Geeslin, 2014). This presentation will discuss how teaching of linguistic variation was carried out in a Spanish L2 classroom utilizing different online, Spanish corpora (mainly *Corpus del Español* and *CREA*), how interactions during pair-work were shaped through the use of corpus data and corpora, and what students' perceptions of utilizing these tools were.

The course analyzed was an advanced L2 course on Spanish grammar that was taught during a semester at a North American college. Sixteen English L1 students, whose proficiency ranged from B2 to C1 level and who were majors or minors in Spanish, took this course. Activities using corpora were incorporated throughout the whole term, as homework and as in class activities. The main goal of these activities was for learners to look for different examples of Spanish variation—i.e., lexical and morpho-syntactic variation motivated by textual genres and regions—so that they moved from a pedagogical understanding of language into a descriptive one.

Two weeks of class, i.e., six sessions of fifty minutes, were recorded. The recordings focused on pair-work utilizing both textbook activities and corpus-based activities. To that end, each pair of students received a digital recorder. An analysis based on the task-as-workplan and task-in-process distinction (see Seedhouse, 2004) was carried out. That is, activities were coded as whether the activity planned by the instructor was carried accordingly by the students, whether the activity slightly diverged, and whether the activity strongly diverged. Results suggest that textbook-based activities were most closely carried out whereas in corpus-based activities, there were more divergences in how the activity was understood and completed.

At the end of the term, students completed an anonymous questionnaire about their perceived advantages and disadvantages of utilizing corpora for the teaching of linguistic variation. In general, students acknowledged the importance of corpora for bringing real examples into the classroom. However, whereas some students seem to be willing to move beyond a pedagogical understanding of grammar, for others, this was harder since, according to them, learning about variation could not help them improve their grammar. Students also pointed out difficulties when searching for examples in the online corpora and when understanding the examples because of lack of vocabulary.

Based on these outcomes, some suggestions are given to better incorporate the teaching of language variation utilizing corpora in advanced grammar courses as well as in lower-level courses.

References

- Davies, M. (2002-) *Corpus del Español: 100 million words, 1200s-1900s*. Available online at <http://www.corpusdelespanol.org>.
- Geeslin, K., with Long, A. (2014). *Sociolinguistics and Second Language Acquisition: Learning to Use Language in Context*. New York, NY: Routledge.
- Real Academia Española. Banco de datos (CREA) [online]. *Corpus de Referencia del Español Actual*. Available online at <http://www.rae.es>
- Seedhouse, P. (2004). *The Interactional Architecture of the Language Classroom: A Conversation Analysis Perspective*. Malden, MA: Blackwell Publishing, Inc.

Introducing DDL vocabulary files in an ESP context

Sanja Marinov
University of Split

The aim of this talk is to present the case of introducing the so called DDL vocabulary files in an ESP context, in particular in the context of teaching English for tourism. The idea was triggered both by the general agreement about the usefulness of DDL in language teaching and its absence from regular teaching practice. As Tyne observes, although the researchers “may be quite keen to underline the potential of their work for teaching and learning” there is a “lack of user-friendly applications for general, everyday practice” (Tyne, 2012:114). Boulton and Tyne (2014) note that in France it is almost completely absent from the skill set of language teachers and the situation is the same in Croatia. The suggested DDL vocabulary file is also a response to the general needs and requirements of modern language pedagogy such as demand-high teaching (Scrivener, 2014), enhancing learner autonomy, more attention to vocabulary teaching, flipped classrooms, specific needs of teaching ESP, to name just a few of direct concern to the current case. More specifically, this case addresses three frequently cited methodological difficulties of corpus use in language teaching and learning that need to be overcome, i.e. the time-consuming nature of DDL, the decontextualised nature of corpus samples, and meeting the curricular requirements.

The process of having students work on DDL vocabulary files consisted of several stages. First the file was designed to supplement a particular unit of the English for tourism course. The rationale behind the design will be briefly described since a wide range of approaches to shaping corpora data into appropriate tasks was used exploiting various points on the continuum of DDL exercises (Boulton, 2012). Some of them seriously challenge the original idea of DDL and exploit its “fuzzy boundaries” (Boulton, 2010) trying to “avoid excessive demands, which can lead to frustration on the part of the learners” (Kaltenböck and Mehlmauer-Larcher, 2005). The working material was called Vocabulary file for the ease of reference but the emphasis is on presenting content as well as vocabulary. It is important to note

that the activity was not carried out under laboratory conditions but presented as a viable activity supplementing the current curriculum. Nonetheless, the students were assigned a pre-test and a post-test to check the progress they achieved. A qualitative analysis of the findings will provide the description of the achieved progress. The students also provided feedback about the overall experience in interviews and by filling in a questionnaire. Only the most relevant findings about their attitudes will be mentioned.

References

- Boulton, Alex. 2010. Data-driven learning: the perpetual enigma. In S. Roszkowski & B. Lewandowska-Tomaszczyk (eds.). *Explorations across Language and Corpora*. Frankfurt: Peter Lang.
- Boulton, Alex. 2012. What data for data-driven learning?. *The EUROCALL Review. EUROCALL 2011*. Vol. 20. 23-27.
- Boulton, Alex & Tyne Henry. 2014. *Des documents authentiques aux corpus*. Les éditions Didier, Paris.
- Kaltenböck, Gunther & Mehlmauer-Larcher, Barbara. 2005. Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL17* (1). 65-84.
- Scrivener, J. (2014). Demand-high teaching. *The European Journal of Applied Linguistics and TEFL*, 3(2). 47-58.
- Tyne, H. (2012). Corpus work with ordinary teachers: data-driven learning activities. In A. Boulton & J. Thomas (eds.). *Teaching and language corpora: input, process and product*. Brno: Masaryk University Press. 114-129.

Nominal complexity in learner writing

Jillian Nahm

Yonsei University

While grammatical complexity in L2 writing has been studied for decades, there still remains considerable disagreement as to what measures are admittedly important (Vyatkina, 2015). One of the recent arguments against the traditionally acknowledged measures such as clausal complexity stems from an observation that characteristics of advanced writing are not easily captured by those measures (Biber et al., 2011; Takuchi et al., 2013). This study addresses the issue by examining phrasal complexity as an alternative or complement. Specifically, we focus on the use of nouns and their modifiers in college-level learner writing, assuming that more developed writing entails linguistic ability in enriching and concretizing concepts represented by nominal phrases.

In the present study, nominal complexity is defined in three dimensions: a) length of production, b) amount of modification, and c) amount of concatenation. Length of production can be measured by counting the number of tokens in a noun phrase. For amount of modification, we compute frequencies of various types of noun modifiers, i.e. determiners, quantifiers, relative clauses, attributive adjectives, and prepositional phrases. Finally, frequencies of noun-noun phrases and double

conjunctions are considered to quantify amount of concatenation. In sum, nominal complexity attempts to measure in what extent writers flesh out noun phrases either by adding prepositional or postpositional modifiers to nominal constituents or by joining them together in a chain.

Corpus-based analyses are undertaken to validate a hypothesis that development in noun phrases are central for college-level L2 writers who want to express themselves in more sophisticated ways. We adopt a large collection of EFL learners' writing samples drawn from the International Corpus Network of Asian Learners of English (Ishikawa, 2011) to investigate the following research questions: a) Can nominal complexity indicate L2 learners' proficiency level, in addition to non-nativeness? b) Which dimensions/measures correspond to learners' proficiency levels well? The corpus consists of 3,800 argumentative essays written by English learners from 6 Asian EFL countries in addition to 400 samples provided by native speakers ($N=4,200$).

Using Stanford Parser as a preprocessor for the syntactic analysis over the corpus with 0.95 million words, it was discovered that nominal complexity is a better indicator for learners' writing proficiency than clausal complexity or other classic structural measures heavily relying on T-units. For instance, the mean length of noun phrases was able to differentiate upper and lower levels out of the intermediate learners while the mean lengths of sentences, T-units, and clauses were not (confidence level = .95). Among the measures for the amount of modification and concatenation, two metrics stood out: the occurrences of complex nominals in total and those of attributive adjectives. These empirical results suggest that advanced learners along with native speakers are noticeably better at building up complex nominals or more descriptive noun phrases than lower-level learners, implying that nominal complexity is worth to zooming in to for the development of L2 writing skills in higher education.

References

- Biber, Douglas, Gray, Bethany & Poonpon, Kornwipa. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1). 5–35.
- Ishikawa, Shin'ichirō. 2011. A new horizon in learner corpus studies: The aim of the ICNALE project. In Weir, George R. S., Ishikawa, Shin'ichirō, & Poonpon, Kornwipa (eds.), *Corpora and language technologies in teaching, learning, and research*, 3–11. Glasgow: University of Strathclyde Press.
- Taguchi, Naoko, Crawford, William & Wetzel, Danielle Zawodny. 2013. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly* 47(2). 420–430.
- Vyatkina, Nina (ed.). 2015. Special issue: New developments in the study of L2 writing complexity. *Journal of Second Language Writing* 29. 1–103.
-

Evaluating learner language complexity: Traditional corpus methodology versus online analysers. A case study using the ICCI.

Pascual Pérez-Paredes¹, María Belén Díez-Bedmar²

¹University of Cambridge, ²Universidad de Jaén

The evaluation of complexity is of interest for the analysis of learner language mainly as a dependent variable as a “descriptor of L2 performance and [...] indicator of L2 proficiency” (Bulté & Housen, 2012: 21). However, these authors argue that the study of complexity has been marked by general and vague conceptualizations of the term, which has resulted in a number of diverging interpretations. In this respect, the use of corpora and language-related analytical tools can provide researchers with different measures and metrics that can help us study complexity in ways that favour construct operationalization, research replicability or even language assessment. For example, syntactic complexity features may predict human speaking proficiency scores (Chen & Zechner, 2011).

In this paper we explore the affordances of two different methodologies that may be instrumental in analysing learner language complexity: “traditional” corpus linguistics methodology and the web-based lexical complexity analyser. The former was undertaken by means of WMatrix (Rayson, 2008, 2009). In particular, we have explored the ways in which Wmatrix can offer us lexical complexity metrics that can help us establish different criterial features at the four levels analysed. The latter was explored with the Lexical Complexity Analyzer (LCA) (Ai & Lu, 2010; Lu, 2010; Lu, 2012), a web service that allows users to analyze the lexical complexity of written English language samples using different measures of lexical density, variation and sophistication.

The learner corpus analysed was a subsection of the Spanish subcomponent of the International Corpus of Contrastive Interlanguage (ICCI) (Tono & Díez-Bedmar, 2014), composed of the essays written on the topic ‘Describe your favourite film’ in Grades 7, 8, 11 and 12 (i.e. the first two years of compulsory secondary education and the last two years in non-compulsory secondary education in Spain).

To explore the results of both methodologies, we used an adapted version of the inventory of linguistic complexity measures in task-based studies (Bulté & Housen, 2012) in order to provide a comparison of both approaches for the analysis of learner corpora. The results reveal important differences in learner writing at different levels concerning their most frequent use of, among others, infinitive markers, finality markers, auxiliary verbs and adverbs. The paper discusses the contributions of both methodologies and how these can inform SLA practitioners.

References

- Ai, Haiyang & Xiaofei Lu. 2010. A web-based system for automatic measurement of lexical complexity. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8-12.

- Bulté, Bram & Alex Housen. 2012. De[FB01?]ning and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 32, 21.
- Chen, Miao & Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*(pp. 722-731). Association for Computational Linguistics.
- Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4). 474-496.
- Lu, Xiaofei. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2). 190-208.
- Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13(4) 519-549.
- Rayson, Paul. 2009. Wmatrix: a web-based corpus-processing environment, Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/> (1 February, 2016)
- Tono, Yukio & María Belén Díez-Bedmar. 2014. Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics*, 19(2). 163-177.

Exploring, teaching and learning phraseology with Phrime

Piotr Pezik

Transition Technologies

Phraseology is considered to play a key role in achieving fluency and communicative efficiency in a foreign language (Pawley & Syder 1983), (Paquot & Granger 2012), (Cowie, Mackin, McCaig 1993: X). Automatic extraction of phraseological units from corpora is a widely investigated topic³ and some of the most recent developments in this area are directly applicable in phraseodidactics.

This paper presents Phrime — a language processing solution for exploring and detecting phraseology with experimental modules for language materials development. Phrime uses dependency-based phraseology extraction in order to a) identify potentially prefabricated expressions in large reference corpora, b) organize them into customizable dictionaries c) detect phraseology in user-submitted texts and d) develop and share data-driven language learning and teaching materials.

The method of extracting phraseology from large reference corpora is based on the assumption that phraseological units form connected subgraphs (i.e. subtrees) in the sentence dependency structure (Pezik 2016). It is inspired by the so-called Continuity Constraint (O'Grady 1988) and recent work on dependency syntax (Osborne et al. 2012). By aggregating millions of instances of dependency subtrees, we identify phraseological units of different formal and functional types, ranging from lexical and grammatical collocations and collocational chains, pure and figurative idioms to speech formulas, lexical bundles and sentence stems. Table 1 below shows

³ See Seretan (2011) for a recent review of the work in this area.

#	Chain type	Count	Example
1	noun + preposition + noun	1372	exchange + of + ideas
2	adjective + noun	508	good + idea
3	verb + noun	482	reject + idea
4	noun + verb or adjective	466	idea + go
5	verb + noun + preposition	277	have + idea + of
6	verb + adjective + noun	187	have + clear + idea
7	verb + adjective + noun + preposition	60	have + good + idea + of
8	phrasal verb + noun	42	give + up + idea
9	adverb + adjective + noun	42	pretty + good + idea

Table 1: A summary of automatically acquired 3436 recurrent chains conforming to 9 dependency patterns.

a summary of a Phrime dictionary entry generated for the different collocational catenae in which the see word ‘idea’ was found the COCA corpus.

For example, there are 1372 different recurrent dependency chains in which the noun ‘idea’ has a prepositional object (e.g. ‘exchange of ideas’), 508 potential binary collocations in which it is modified by an adjective (e.g. ‘excellent idea’) and 60 different recurrent chains in which it is a direct object with an adjectival modifier and a prepositional object (e.g. ‘to have a good idea of something’).

Each of these collocational chains has its separate dictionary page with examples of use and information about its structure, distribution and longer recurrent combinations which it tends to be part of. Word and chain entries can be bookmarked, edited and discarded by the users of the dictionary.

This combinatorial dictionary is also used in the phraseology detection module of Phrime to annotate phraseological units in texts submitted to the system. This in turn makes it possible to create gap-fill exercises from working collections of phraseological units validated by the users. Exercises can be used independently or assigned to texts, which in turn can be grouped into courses. Registered users of Phrime can take the role of ‘teachers’ and invite their ‘students’ to join a particular text-based course.

We argue that the current version of Phrime has immediate applications in lexicography and teaching materials development and that with some validation against phraseological dictionaries it can also be used by advanced learners. The software is currently available as a web application at phrime.tt.com.pl. Early access to the service is granted upon request.

References

- Hausmann, Franz Josef. “Was Sind Eigentlich Kollokationen.” *Wortverbindungen-Mehr Oder Weniger Fest*, 2004, 309–34.
- O’Grady, William. “The Syntax of Idioms.” *Natural Language & Linguistic Theory* 16, no. 2 (1998): 279–312.
- Osborne, Timothy, Michael Putnam, and Thomas Groß. “Catenae: Introducing a Novel Unit of Syntactic Analysis: Catenae: Introducing a Novel Unit of

- Syntactic Analysis.” *Syntax* 15, no. 4 (December 2012): 354–96. doi:10.1111/j.1467-9612.2012.00172.x.
- Paquot, Magali, and Sylviane Granger. “Formulaic Language in Learner Corpora.” *Annual Review of Applied Linguistics* 32 (March 2012): 130–49. doi:10.1017/S0267190512000098.
- Pawley, Andrew, and Frances Hodgetts Syder. “Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency.” *Language and Communication* 191 (1983): 225.
- Seretan, Violeta. *Syntax-Based Collocation Extraction*. Vol. 44. Springer, 2011.
-

English intonation of advanced learners: A contrastive interlanguage analysis

Karin Puga

Justus-Liebig-Universität Gießen

Prosodic deviances in nonnative speech can contribute to a perceived foreign accent and/or even impede communication, intelligibility, and comprehensibility (cf. Jilka 2000, 2007; Mennen et al. 2014; Trofimovich & Baker 2007). Also, 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 languages and even advanced learners still deviate from native-like intonation patterns (cf. Bongaerts et al. 1997; Scovel 2000). Despite its importance, researchers have only recently started to take an interest in the description of nonnative intonation patterns (see, e.g. Mennen & de Leeuw 2014; Li & Post 2014; Mennen et al. 2014).

Following this trend, the present study adopts a corpus-based approach contrasting native and interlanguage data. Thus, this study sets out to characterize the intonational features produced by three different L2 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 involving intonational phrasing and pitch height of both native and nonnative speech: How are intonational phrases structured and what is the average timing of intonational and intermediate phrases? Which pitch movement or level is used to mark intonation and intermediate phrase boundaries? How many and what pitch accents are used to highlight or emphasize words?

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) of Czech, German, and Spanish. Through quantitative and qualitative analyses, the frequency and use of pitch and the structure of intonational phrases were annotated using the *Tone and Break Indices* (ToBI) (Silverman et al. 1992) annotation system and 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 the British English variety, and the *Santa Barbara Corpus of Spoken American English* (Du Bois et al. 2000-2005) and

the *New South Voices* (Murrey Atkins Library), representing the American English variety.

Preliminary findings indicate that the German and Spanish groups in their pitch patterns at utterance-final position deviate from each other, as well as the native-speaker control group: While the British native-speakers mainly stick to the usage of falls within major intonation phrases (IP) and levels and falls at minor intermediate (ip) phrases, the learners often misuse these pitch movements at intonation-unit boundaries to different degrees. Both German and Spanish learners of English produce more instances of falling tones within IP boundaries than the NSs of English, whereas the opposite holds true for the usage of falling tones within ip boundaries, where an underuse could be observed. While the German learners' speech was also characterized by a frequent usage of rising tones, the Spanish learners overuse level tones.

In the final part of the paper, I will use regression modelling in order to be able to explain the factors responsible for these deviations (such as L1-transfer, developmental factors, the relationship between the interviewer and the interviewee, genre-dependent differences or other learner variables (age, gender, years of English, stays abroad, etc.)).

References

- Bongaerts, Theo, Chantal van Summeren, Brigitte Planken & Erik Schils 1997. Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition* 19. 447-465.
- De Cock, Sylvie 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*. New Series 2. 225-246.
- Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, & Nii Martey 2000-2005. *Santa Barbara corpus of spoken American English*. Parts 1-4. Philadelphia: Linguistic Data Consortium.
- Gilquin, Gaëtanelle, Sylvie De Cock & Sylviane Granger (eds.) 2010. *LINDSEI: Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Granger, Sylviane 1998. The computer learner corpus: a versatile new source of data for SLA research. In Sylviane Granger (ed.), *Learner English on Computer*. London: Longman. 3-18.
- Jilka, Matthias 2000. *The Contribution of Intonation to the Perception of Foreign Accent*. PhD thesis, Institute of Natural Language Processing. University of Stuttgart.
- Jilka, Matthias 2007. Different manifestations and perceptions of foreign accent in intonation. In Jürgen Trouvain & Ulrike Gut (eds.), *Non-Native Prosody – phonetic description and teaching practice*. Berlin: Mouton de Gruyter. 77-96.
- Li, Aike & Brechtje 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, Ineke & Esther de Leeuw 2014. Beyond Segments: Prosody in SLA. *Studies in Second Language Acquisition* 36. 183-194.
- Mennen, Ineke, Felix Schaeffler & Catherine Dickie 2014. Second language acquisition of pitch range in German learners of English. *Studies in Second Language Acquisition* 36. 303-329.
- Murrey, J. Atkins Library 2016. *New South Voices Collection*. UNC Charlotte.
- Scovel, Thomas 2000. A critical review of the critical period research. *Annual Review of Applied Linguistics* 20. 213-223.
- Silverman, Kim, Mary E. Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert & Julia Hirschberg 1992. ToBI: A standard scheme for labeling prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff. 867-879.
- Trofimovich, Pavel & Wendy 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.

Linguistic accessibility (SDH and AD) in the university programs for teaching AVT: the AVLA project

Juan Pedro Rica Peromingo, Ángela Sáenz Herrero

Complutense University in Madrid

The important change which is taking place with respect to the media and the audiovisual world in Europe needs to benefit all population, in particular those with special needs, such as deaf and hard-of-hearing population and blind and partially-sighted population (Orero, 2007; Díaz Cintas, 2007; Jiménez Hurtado, 2007; Díaz Cintas et al., 2007; Rica, 2016). This recent interest in the field of translation, and particularly in audiovisual translation (AVT), can be observed with the inclusion of the teaching and learning of the different modes of AVT in the degree and post degree courses at Spanish universities (Rica & Sáenz, *in press 2016*), the appearance of a great amount of research projects funded by the Ministry of Education and the universities, PhD theses, etc. which try to expand the interest and practice of AVT linguistic accessibility.

We present a research project led at the Complutense University which is creating a corpus of AVT activities for teaching purposes and tries to analyze the creation and reception of subtitles for deaf (SDH) and audio description for the blind (AD) population: the AVLA Project (*Audiovisual Learning Archive*), which consists of a corpus of audiovisual materials carried out by university students (in different Spanish universities) on different AVT modes (subtitling for hearing and SDH, AD and dubbing). In particular, in this study we present the webpage which contains the materials created by the students (<https://avlearningarchive.wordpress.com/>) with respect to those modes referring to linguistic accessibility. A group of deaf and blind population has been in charge of testing the students SDH and AD corpus of audiovisual materials. Some questionnaires have been used in order to evaluate the students' production and preliminary results will be presented with respect to the most important aspects, difficulties and deficiencies in the SDH and AD included

in the corpus, together with some changes and improvements in the quality of the SDH and AD analyzed will be suggested. In the end, a demand for the use of corpus linguistics (Rica, 2014) for teaching and learning these specific AVT modes at a university level will be suggested.

References

- Díaz Cintas, Jorge. 2007. Traducción audiovisual y accesibilidad. In C. Jiménez Hurtado (ed.), *Traducción y accesibilidad. Subtitulación para sordos y audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual*. Frankfurt am Main: Peter Lang, 9-3.
- Díaz Cintas, Jorge, Pilar Orero & Aline Remael (eds.). 2007. *Media for All: Subtitling for the Deaf, Audio description and Sign Language*. Amsterdam: Rodopi.
- Jiménez Hurtado, Carmen (ed.). 2007. *Traducción y accesibilidad. Subtitulación para sordos y audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual*. Series: Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation - Volume 39. Frankfurt am Main: Peter Lang.
- Orero, Pilar 2007. La accesibilidad en los medios: una aproximación multidisciplinar. *TRANS. Revista de Traductología* 11: 11-14.
- Rica, Juan P. 2014. La traducción de marcadores discursivos (DM) inglés-español en los subtítulos de películas: un estudio de corpus. *JoSTrans – Journal of Specialised Translation*, 21: 177-199.
- Rica, Juan P. 2016. *Aspectos lingüísticos y técnicos de la traducción audiovisual (TAV)*. Series: Linguistic Insights – Studies in Language and Communication. Frankfurt am Main: Peter Lang.
- Rica, Juan P. & Ángela Sáenz. *in press 2016*. Audiovisual Translation (AVT) Modes: Dubbing, Audio description (AD) and Subtitling for Hearing and Deaf and Hard-of-Hearing (SDH) as an L2 Learning Pedagogical Tool. In C. Herrero & I. Vanderschelden (eds.), *Using Film and Media in the Language Classroom: Reflections on Research-Led Teaching*. Manchester Metropolitan University Publications.

L2 learners' developing knowledge of English verb constructions: Usage-based views and implications for teaching

Ute Römer¹, Cynthia M. Berger¹ & Nick C. Ellis²

¹Georgia State University, ²University of Michigan

This paper adopts a usage-based perspective on language acquisition to investigate how knowledge of verb-argument constructions (VACs) develops in second language learners across proficiency levels. We will first present findings from an analysis of L1 German and L1 Spanish learner use of English VACs, such as the 'V about n' (e.g., *let's talk about the weather*) or the 'V with n' construction (e.g., *he always agreed with her*). We will then discuss what our findings mean for language

teaching. Our analysis is based on a corpus of learner writing at different levels of proficiency, described in further detail below. We were interested in determining (1) how VACs develop in second language (L2) writing as proficiency increases and (2) how the use and emergence of VACs is affected by the learner's first language.

The paper builds on previous work on learner knowledge of VACs carried out in a usage-based linguistics tradition (Römer, et al., 2014a and 2014b; Gries, & Wulff, 2005). This work has shown that advanced learners of English have constructional knowledge, that learners' VAC knowledge differs in systematic ways from that of native speakers, and that learners' verb-VAC associations differ across L1 groups. What previous studies have not been able to address, mostly due to the unavailability of pertinent data at lower proficiency levels, is how this constructional knowledge unfolds over time (though see Li, Eskildsen, & Cadierno, 2014). Likewise, only few studies have systematically contrasted learners from different L1 backgrounds to investigate the role of transfer from the first language. The present paper seeks to take steps to closing both of these gaps.

To gather information on learner VAC use at different proficiency levels, we use subsets of the Education First-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou, & Korhonen, 2013), consisting of writing samples by learners of a range of L1s who were placed into 16 different proficiency levels. For our study, we retrieved sets of texts written by German and Spanish learners at Common European Framework of Reference (CEFR) levels A1 through C2. The EFCAMDAT subsets we compiled—over 28,000 texts and 2.8 million words from L1 German learners, and over 40,000 texts and 3.2 million words from L1 Spanish learners—constitute a pseudo-longitudinal learner corpus that complements existing corpus resources. From these EFCAMDAT subsets, we exhaustively retrieved instances of 19 different VACs.

The EFCAMDAT datasets enabled us to describe learners' dynamically evolving abilities and trace the emergence of constructions over time. We will report trends in frequency developments of VACs, type/token ratios, dominant verb-VAC associations, correlations between verbs produced by learners at different levels, and correlations between verbs produced by learners of different first languages (German vs. Spanish). We will also discuss how insights from our study and usage-based L2 research in general can inform pedagogical practice and have a positive impact on second language teaching. Pedagogical implications and recommendations will consider form-focused instruction, the use of input floods, and awareness-raising activities, especially for learners of first languages that are typologically different from the target language.

References

- Römer, U., M. B. O'Donnell & N. C. Ellis. (2014a). Second language learner knowledge of verb-argument constructions: Effects of language transfer and typology. *The Modern Language Journal*, 98(4), 952-975.
- Römer, U., A. Roberson, M. B. O'Donnell & N. C. Ellis. (2014b). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, 38, 59-79.

- Geertzen, J., T. Alexopoulou & A. Korhonen. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). Proceedings of the 31st Second Language Research Forum (SLRF). Carnegie Mellon University: Cascadilla Press.
- Gries, S. T., & S. Wulff. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, 3, 182-200.
- Li, P., S. W. Eskildsen, & T. Cadierno. (2014). Tracing an L2 learner's motion constructions over time: A usage-based classroom investigation. *The Modern Language Journal*, 98, 612-628.
-

The use of smallwords in the speech of German learners of English: a corpus-based study of the factors of instruction and natural exposure

Anna Rosen

University of Freiburg

Spontaneous, unplanned speaking is considered an essential and yet often the most challenging skill in acquiring a foreign language – and one which is difficult to teach (e.g. Brown & Yule 1983, Bailey & Savage 1994). It has been suggested that learners' use of smallwords, such as, for instance, discourse and stance markers, would be one way of bridging their planning time while speaking and of improving the naturalness and fluency of their speech (e.g. Götz 2013).

This paper builds on earlier research which has found that i) greater fluency in the speech of English learners correlates with the use of a greater number and variety of so-called smallwords like, for example, *well, right, and stuff, I mean* (Hasselgren 2002) and that ii) contact with native speakers and exposure to natural input during stays abroad seem to help learners to acquire a more native-like usage of discourse markers (Müller 2005, Götz 2013). Bringing both findings together, the present paper investigates the consequences of foreign language learning in an instructional setting only versus instruction plus additional natural exposure for the use of fluency-enhancing smallwords. Based on data from the German component of the LINDSEI database (Gilquin et al. 2010), it compares the occurrence of a list of smallwords, established as fluency-enhancing in Hasselgren (2002), in the speech of German undergraduates of English who spent some time abroad in an English-speaking country and those who did not. The paper thus complements the studies of discourse markers in German learner English by Müller (2005) and Götz (2013) and offers a fine-grained analysis of individual differences in advanced learners' speech in the realm of communication strategies in spontaneous speech.

In line with previous research, the results show that there is indeed an overall tendency for fluency-enhancing smallwords to be used more often by German students who spent some time in an English-speaking country, although corpus evidence also suggests that their usage still differs quantitatively and qualitatively from the one of native speakers of English of the same age group. The results are taken as an indication that the kind of language training university undergraduates have received so far at secondary school and university in Germany is not sufficient or

explicit enough for them to be able to use such features of spoken conversation in a helpful and appropriate way. This also confirms earlier work on the topic, which pleads for more consideration of spoken grammar features and fluency-enhancing strategies in (German) secondary school teaching and strongly suggests that teacher trainees should spend some time in an English-speaking environment as their speech will provide an important model for generations of German pupils to come (Rühlemann 2008, Mukherjee 2009, Goh & Burns 2012, Rogge 2012, Diao-Klaeger & Thörle 2013, Götz 2013).

References

- Brown, Gillian & George Yule. 1983. *Teaching the spoken language. An approach based on the analysis of conversational English*. Cambridge: Cambridge University Press.
- Bailey, Kathleen & Lance Savage. 1994. *New ways in teaching speaking*. Alexandria, VA: TESOL.
- Diao-Klaeger, Sabine & Britta Thörle. 2013. Diskursmarker in L2. In Christoph Bürgel & Dirk Siepmann (eds.), *Sprachwissenschaft – Fremdsprachendidaktik: Neue Impulse*, 145-160. Baltmannsweiler: Schneider Hohengehren.
- Gilquin, Gaëtanelle, Sylvie De Cock & Sylviane Granger (eds.). 2010. *Louvain International Database of Spoken English Interlanguage* (CD-ROM + Handbook). Louvain-la-Neuve: Presses universitaires de Louvain.
- Götz, Sandra. 2013. *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.
- Goh, Anne & Christine Burns. 2012. *Teaching speaking: A holistic approach*. Cambridge: Cambridge University Press.
- Hasselgren, Angela. 2002. Learner corpora and language testing: Smallwords as markers of learner fluency. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, 143-173. Amsterdam: John Benjamins.
- Müller, Simone. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.
- Mukherjee, Joybrato. 2009. The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In Karin Aijmer (ed.), *Corpora and language teaching*, 203-230. Amsterdam: John Benjamins.
- Rogge, Michael. 2012. Sagen können, was man zu sagen hat: Mündliche Kompetenz mit Sprechaufgaben fördern. *Der fremdsprachliche Unterricht English* 116. 2-6.
- Rühlemann, Christoph. 2008. A register approach to teaching conversation: Farewell to Standard English? *Applied Linguistics* 29(4). 672-693.
-

The effects of corpus and dictionary use: Error correction in L2 writing

Yoshiho Satake
Surugadai University

Introduction

While the strength of corpus and dictionary use in language learning has been stated (ex. Satake 2015, Tono, Satake & Miura 2014), it remains unclear about how corpora and dictionaries have different effects on improving L2 writing because there are few longitudinal comparative studies. More empirical research needs to be conducted in this area.

Research question

This study investigated the effects of corpus and dictionary use on learners' correcting their errors in L2 writing. Students consulted a corpus and dictionaries to correct their errors based on teacher and peer feedback. To investigate how corpus and dictionary use influenced correcting learners' errors, the author made corpora of students' writing and compared the effects of corpus and dictionary use. The research question is as follows: How do a corpus and dictionaries have different effects on error correction in L2 writing?

Methods

Thirty Japanese undergraduate students at a university in Tokyo participated in the study for nine weeks. They were intermediate English learners and reached level B1 in the Common European Framework of Reference for Languages (CEFR).

The following procedure was taken.

(1) The timed essay task was given (twenty-five minutes). The students did not consult a corpus or dictionaries.

(2) After the task, the revision session was held, in which the students were given highlighted feedback for problematic points by the author (Week 1, 3, 5, 7, 9) and by their classmates (Week 2, 4, 6, 8).

(3) The students revised the errors on which they were given feedback, consulting Corpus of Contemporary American English (COCA) and dictionaries (fifteen minutes). If time remained, they were encouraged to correct other errors.

Results and discussion

The results indicated the following:

(1) Both COCA and dictionaries helped the students correct more than seventy percent of errors.

(2) COCA helped the students revise more omission and addition errors of articles and prepositions.

(3) Dictionaries helped the students revise more lexical errors of nouns and verbs.

(4) While dictionaries helped the students correct many verb errors based on teacher feedback, dictionaries less helped them correct verb errors based on peer feedback.

Both COCA and dictionaries were effective as they contributed to correcting the majority of errors. The results suggest that concordance lines that provided more context information than dictionaries were more effective for revising omission and addition errors, while dictionaries that provided semantic information were more effective for revising lexical errors. Considering that dictionaries helped the students correct more verb errors based on teacher feedback, it is suggested that they were given different kinds of feedback by the author and their classmates and the different effects of teacher and peer feedback should be studied further. The findings suggest that balanced use of a corpus and dictionaries is needed to improve accuracy in learners' L2 writing, as a corpus and dictionaries contribute to correcting different kinds of errors.

References

- Satake, Yoshiho. 2015. Comparison of dictionary use and corpus use: different effects on learning L2 phrases. *Proceedings of ASIALEX (The Asian Association for Lexicography) 2015 Hong Kong*. 222-229.
- Tono, Yukio, Satake, Yoshiho & Miura, Aika. 2014. The effects of using corpora on revision tasks in L2 writing. *ReCALL* 26(2). 147-162.

Which corpus management systems for non-specialists in computer science? The case of French trainee teachers of German.

Eva Schaeffer-Lacroix
University Paris-Sorbonne

This paper is based on the idea that foreign language teachers would accept to invest their energy and time in getting acquainted with corpus tools if they understand that these tools are suited to address barriers of learning (Breyer 2009). In addition, the corpus management system they are confronted with must have a certain number of characteristics which make them accessible even to non-specialists in computer technology (Mukherjee, 2004; Römer 2009). These statements underlay a research project which will be presented in the first part of this paper. This project was offered to future teachers of German, enrolled in the first year of a master's program at the teacher training department of the university Paris-Sorbonne. During a class observation period preceding the project, the trainee teachers were given the opportunity to detect learning needs. They identified following pitfalls: sometimes, the resources used in the classroom did not seem convincing, individual learner differences were not always taken into account, linguistic concepts were absent or unclear, and it happened that the learning activities were not meaningful enough. As shown by Breyer (2009), the use of corpus tools can contribute to resolve such issues. During the project, three groups, each composed of three participants, were invited to design learning scenarios containing at least one corpus-based activity.

To do this, they had to get familiar with a corpus management system, in this case with *Sketch Engine* (Kilgarriff, Rychly, & Pomikalek nd), used by Thomas (2015) in a similar context. Following research data have been collected: audio recordings of lesson interactions, screen casts filmed during corpus exploration sessions, and the filmed final interviews made with each of the small groups by an external researcher. We analysed the students' actions and their reactions to *Sketch Engine*. A certain contrast between convinced and sceptic users could be identified. Our results suggest that the reluctance of some of the participants to use *Sketch Engine* is not as much motivated by technical difficulties than by reasons of pedagogical and/or financial nature. We found out that from a technical point of view, *Sketch Engine* is well suited not only for advanced corpus explorers, but also for beginners. However, beyond the 30 days free trial, the users are charged for keeping their *Sketch Engine* account; a free version of *Sketch Engine* called *No Sketch Engine* exists (Natural Language Server, nd), but its features are limited. This is why we consider introducing next year's cohort to *TXM* (Heiden, Magué, Pincemin 2010; Heiden 2010), an open-source textual analysis software. In the second part of this paper, we compare the main features of *Sketch Engine* and *TXM*, with the following question in mind: to what extent do the interface and the functions of both systems meet the technical, practical and heuristic needs of novice users aiming to design corpus-based activities for German teaching purposes?

References

- Breyer, Y. (2009). "Learning and teaching with corpora: reflections by student teachers." *Computer Assisted Language Learning* 22 (2), pp. 153-172. doi: 10.1080/09588220902778328.
- Heiden, S., Magué, J-P., Pincemin, B. (2010)." TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement". In I. C. Sergio Bolasco (eds.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010* (Vol. 2, p. 1021-1032). *Edizioni Universitarie di Lettere Economia Diritto*, Roma, Italy. <https://halshs.archives-ouvertes.fr/halshs-00549779/fr/>
- Heiden, S. (2010). "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme". In K. I. Ryo Otoguro (eds.), *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24* (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan. Online.
- Kilgarriff, A., Rychly, P. & Pomikalek J. (nd). *Sketch Engine* [Corpus management system]. <http://www.sketchengine.co.uk/>
- Mukherjee, J. (2004): "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany", *Applied Corpus Linguistics: A Multidimensional Perspective*, ed. Ulla Connor & Thomas Upton. Amsterdam: Rodopi. 239-250 (pdf-version)
- Natural Language Server (nd). *No Sketch Engine*. <http://nl.ijs.si/noske/index-en.html>
- Römer, Ute. 2009. Corpus research and practice: What help do teachers need and

what can we offer? In: Aijmer, Karin (ed.). *Corpora and Language Teaching*. Amsterdam: John Benjamins. 83-98. pdf version
Thomas, J. (2015). *Discovering English with Sketch Engine*. Brno: Versatile.

The influence of task-based factors on L2 assignment writing – Corpus-based evidence from ESP learner language

Christine S. Sing

Vienna University of Economics and Business

Investigating L2 writing is a continuing concern within learner corpus research. Recent developments in the field have generated an increased interest in English for specific purposes (ESP) writing. A much debated question is how discipline-specific text-types may be characterised in order to be effectively utilised in writing instruction. One such identifying characteristic is the frequency of occurrence of technical vocabulary. The latter is notoriously difficult to identify (Chung & Nation 2004), given that it tends to have limited range and its frequency varies greatly even within specialist texts (Nation & Hwang 1995: 36f.). However, even if Paquot's (2010: 13) systematic approach involving three frequency-based criteria - keyness, consistency and dispersion - is applied to identifying technical terms in a specialised corpus, two important groups of vocabulary items remain unaccounted for: A group of so-called topical vocabulary; a group of genre-specific words that reflect the nature of the language represented in the corpus (Coxhead & Nation 2001: 264f.).

In order to address this limitation, the present study pursues a twin objective: 1) To investigate to what extent these vocabulary items are indicative of language prompts that have permeated into the students' writing. 2) To compare the usage patterns identified in the research corpus with those in reference corpora, such as the corpus of British Academic Written English (BAWE) and a corpus consisting of the sources used by the student writers. It will be argued that task-based factors, such as the wording of the task and the sources used, account for the variation in the corpus data.

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 is grounded in a mixed methodology, combining bottom-up, inductive, corpus analyses with top-down analyses focussing on larger portions of discourse (Flowerdew 2005).

One interesting finding is that topical vocabulary accounts for more than one quarter of the running words in the corpus, clustering in specific sections of the papers. 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 are in fact 'language re-uses' (Flowerdew & Li 2007). Writers re-use

language for different reasons: They may have pooled commonly used patterns for future (re-)use; they may imitate recently encountered linguistic options; they may have been otherwise primed (Hoey 2005).

These results, while preliminary, suggest that, in ESP writing, a more careful distinction between technical and topical vocabulary is in order. 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 strong reliance on expert uses modelled on the source literature provides strong evidence in favour of learning by imitation (Limburg 2014). The findings of this present study have thus important implications for L2 writing pedagogy and serve to inform teaching materials and curriculum design.

References

- Boulton, Alex, Shirley Carter-Thomas & Elisabeth Rowley-Jolivet (eds.). 2012. *Corpus-informed research and learning in ESP: Issues and applications*. Amsterdam: Benjamins.
- Chung, Teresa M. & Paul Nation. 2004. Identifying technical vocabulary. *System* 32(2). 251–263.
- Coxhead, Avril & Paul Nation. 2001. The specialised vocabulary of English for academic purposes. In John Flowerdew & Matthew Peacock (eds.), *Research perspectives on English for academic purposes*, 252–267. Cambridge: Cambridge University Press.
- Flowerdew, John & Yongyan Li. 2007. Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics* 28. 440–465.
- Flowerdew, Lynne. 2004. The argument for using English specialized corpora to understand academic and professional language. In Ulla Connor & Thomas A. Upton (eds.), *Discourse in the professions. Perspectives from corpus linguistics*, 11–37. Amsterdam: Benjamins.
- Flowerdew, Lynne. 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.
- Gavioli, Laura. 2005. *Exploring corpora for ESP learning*. Amsterdam: Benjamins.
- Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.
- Limburg, Anika. 2014. Imitationslernen in den Wirtschaftswissenschaften: Unterschiede zwischen Examensarbeiten und Forschungsartikeln. *Zeitschrift Schreiben. Schreiben in Schule, Hochschule und Beruf* (<http://www.zeitschrift-schreiben.eu>, retrieved 7 January 2016).
- Nation, Paul & Kyongho Hwang. 1995. Where would general service vocabulary stop and special purposes vocabulary begin? *System* 23(1). 35–41.
- Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.
- Römer, Ute. 2008. Corpora and language teaching. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics, Computational linguistics. Corpus Linguistics. An International Handbook. Volume 1*, 112–131. Berlin: De Gruyter.

- Szmrecsanyi, Benedikt. 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, Christine M. 2012. Writing and Language for Specific Purposes. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, 6266–6274. Oxford: Wiley-Blackwell.
- Tognini-Bonelli, Elena. 2008. Corpora and LSP: Issues and Implications. In Carol T. Torsello, Katherine Ackerley & Erik Castello (eds.), *Corpora for university language teachers*, 31–47. Bern: Peter Lang.
- Tribble, Christopher. 2001. Small Corpora and teaching writing. In Mohsen Ghadessy, Alex Henry & Robert L. Roseberry (eds.), *Small corpus studies and ELT. Theory and practice*, 381–409. Amsterdam: Benjamins.
- Tribble, Christopher. 2002. Corpora and corpus analysis: new windows on academic writing. In John Flowerdew (ed.), *Academic discourse*, 131–149. Harlow: Longman.
-

Corpus construction and specialist vocabulary learning by Chinese and Finnish EAP students

Simon Smith¹, Nicole Keng²

¹Coventry University ²University of Vaasa

It has been shown that language learners can benefit from a discovery-based learning process whereby they construct as well as consult their own specialist corpora and vocabulary portfolios, both for the purposes of translator training (Castagnoli 2006) and for academic English learning (Charles 2012, Smith 2011). It is perhaps in the acquisition of specialist vocabulary and authentic collocational information that the approach offers the most promise.

In this study, there are two cohorts of academic English learners, majoring in Finance or Business Studies. Cohort 1 study at a university in Finland, and the majority of participants are Finnish. Cohort 2, at a UK university, consists mainly of Chinese direct-entry students. Chinese L1 groups have been characterized as unlikely to benefit from autonomous learning approaches, according to Cheng and Dörnyei (2007), while studies such as Groom and Maunonen-Eskelinen (2006) report success with Finnish students using such approaches, in particular involving the creation of student portfolios. A comparative study is, therefore, well motivated.

The research questions are:

1. What are the perceptions of students regarding a corpus-based approach to specialist vocabulary study?
2. Are there generalizable differences between the reported perceptions of the Finnish and Chinese cohorts?
3. Are there significant differences in the progress made by the Chinese and Finnish cohorts ?

Students created mini-corpora based on study materials created by their Business and Finance tutors, uploaded on the VLE. The corpora had specific themes: examples were *Transfer Pricing* and *International Accounting Standards*. Materials included lecture PowerPoint slides, seminar notes, and past exam papers. Students then explored the corpora using Sketch Engine, using concordancing and keyword lists. They were asked to verify the keyword lists to see which keywords genuinely represented the domain. After that, they extended the corpora (using the WebBoot-Cat tool) to create a larger web corpus on the same theme.

The web corpus made available relevant and authentic texts which students could download and read. They also continued to investigate corpus keywords, and used the resource to create and maintain a vocabulary portfolio in spreadsheet format. Here, they would record domain keywords and collocations, along with contextual examples from their corpus. There was also space for them to record other information about the word, for example a business dictionary definition. The tasks could be completed by students at home, but in practice most of the work actually took place in class, in computer lab sessions.

An online questionnaire on the utility of the approach for vocabulary study was used at the end of the intervention, to answer RQs 1 and 2. Pre- and post-tests addressed RQ3. Keywords identified from the authors' own corpus exploration of the Business/Finance tutor materials were tested by a combination of gap-fill and definition questions. These were compiled from the materials themselves where available, but for some definitions recourse was made to a business dictionary.

Student perceptions from both cohorts were varied but mostly positive. The Chinese cohort, unused to this style of learning, were pleasantly surprised by the opportunity to study autonomously when adequate support was given.

References

- Castagnoli, S. (2006). Using the Web as a source of LSP corpora in the terminology classroom. In M. Baroni & S. Bernardini (Eds.), *Wacky! Working papers on the Web as corpus* (pp. 159–172). Bologna: Gedit.
- Charles, M. (2012). Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31, 93-102.
- Cheng, H., & Dörnyei, Z. (2007). The use of motivational strategies in language instruction: The case of EFL teaching in Taiwan. *Innovation in Language Learning and Teaching*, 1(1), 153–174.
- Groom, B., & Maunonen-Eskelinen, I. (2006). The use of portfolios to develop reflective practice in teacher training: A comparative and collaborative approach between two teacher training providers in the UK and Finland. *Teaching in Higher Education*, 11(3), 291-300.
- Smith, S. 2011. Learner Construction of Corpora for General English in Taiwan. *Computer Assisted Language Learning*, 24(4): 291-316. <http://curve.coventry.ac.uk/open/file/96362d30-9307-404e-81e8-16091fb2c118/1/smithcomb.pdf>

Investigating Linguistic Characteristics of L2 Writing Development through the Examination for the Certificate of Proficiency in English: A Multi-Dimensional Analysis

Shelley Staples¹, Xun Yan²

¹Purdue University, ²University of Illinois, Urbana-Champaign

Linguistic complexity has been recognized as a core component of second language (L2) writing development (Biber, Gray & Staples, 2014; Lu, 2011; Taguchi, Crawford, & Wetzel, 2014). Previous research suggests that distinctions across levels of writing proficiency are not necessarily associated with individual linguistic features, but rather with the co-occurrence of multiple features (Biber, Gray & Staples, 2014; Friginal, Li & Weigle, 2014; Jarvis, Grant, Bikowski & Ferris, 2003). In addition, although numerous studies of language development from a corpus-based and usage-based perspective have emphasized the intersection of lexis and grammar, most studies in assessment contexts continue to investigate these as separate areas of development (Römer, 2016). Finally, it is not clear that ‘complexity’ as it is currently operationalized captures key linguistic resources used by writers at different levels of development.

The goal of this study is to further the exploration of linguistic aspects of L2 writing development by using Multi-Dimensional (MD) analysis. Specifically, we investigated a corpus of 595 essays from the Examination for the Certificate of Proficiency in English (ECPE), a large-scale test used to certify advanced level of English proficiency for academic and professional purposes. In order to examine co-occurrence patterns of linguistic features, we conducted MD analysis, a corpus-based analytic framework developed by Biber (1988) for exploring linguistic variation in spoken and written English texts (see also Conrad & Biber, 2001). MD analysis accounts for the co-variation among a wide array of linguistic features and reduces these features to a smaller number of functionally interpretable linguistic dimensions. The advantage of MD analysis is that it represents writing development through a few holistic linguistic dimensions while accounting for all the individual features that contribute to the linguistic dimensions in the analysis of writing performance.

The 41 linguistic features investigated in this study include both individual lexical features (e.g., word frequency and range and type-token ratio), phraseological features (lexical bundles), grammatical complexity features (e.g., phrasal and clausal features) stance features (such as modals and lexical bundles functioning to convey stance), and semantic categories of particular grammatical features (e.g., mental verbs) which have been found to characterize written discourse (e.g., Biber, 1988; Biber, 2006; Biber, Gray & Poonpon, 2011) and writing development (Biber, Gray, & Staples, 2014; Chen & Baker, 2014; Grant & Ginther, 2000; Lu, 2011; Staples, Egbert, LaFlair, & Biber, 2013; Taguchi, Crawford, & Wetzel, 2014). Following Conrad and Biber (2001), we used factor analysis to identify co-occurring lexico-grammatical features and interpreted the resulting factors as dimensions. We calculated dimension scores for each essay in our corpus and examined relationships between the dimension scores and scores on the ECPE.

Results of the MD analysis indicate five underlying factors, representing five functional dimensions of ECPE essays: oral vs. literate discourse, topic-related content, prompt-dependence vs. lexical variety, overt suggestions, and stance vs. referential discourse. Together, these five dimensions accounted for 35% of the holistic score variance. Dimension scores on three of the five dimensions demonstrated linear, significant correlations with the holistic ECPE writing scores awarded by raters: oral vs. literate discourse was positively correlated with writing score while prompt-dependence vs. lexical variety and stance vs. referential expressions were negatively correlated. These dimensions also showed significant differences across score levels.

The findings of this study add to the previous research on L2 writing development by showing the interaction of lexis and grammar in L2 writers' development of register awareness of written discourse, lexical variety at both the individual word and phraseological level, and the importance of linguistic features not usually accounted for in models of complexity such as stance.

References

- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: CUP.
- Biber, Douglas. 2006. *University language*. Philadelphia: John Benjamins.
- Biber, Douglas, Bethany Gray, & Kornwipa Poonpon. 2011. Should we use the characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45. 5–35.
- Biber, Douglas, Bethany Gray, & Shelley Staples. 2014. Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*.
- Chen, Yu-Hua and Paul Baker. 2014. Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2, and C1. *Applied Linguistics*.
- Conrad, Susan, & Douglas Biber. 2001. Multi-dimensional methodology and the dimensions of register variation in English. In Susan Conrad & Douglas Biber (eds.), *Variation in English: Multidimensional studies* (pp. 13-42). New York: Routledge.
- Friginal, Eric, Man Li, & Sara Weigle. 2014. Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays, *Journal of Second Language Writing* 23. 1–16.
- Jarvis, Scott, Leslie Grant, Dawn Bikowski, & Dana Ferris. 2003. Exploring multiple profiles of highly rated learner composition, *Journal of Second Language Writing* 12. 377–403.
- Lu, Xiaofei. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writer's language development, *TESOL Quarterly* 45. 36–61.
- Römer, Ute. 2016. Language assessment and the inseparability of lexis and grammar. In *Connecting Corpus Linguistics and Language Assessments ILTA colloquium*, American Association of Applied Linguistics (AAAL), Orlando, FL.
- Staples, Shelley, Jesse Egbert, Allison McLair, and Douglas Biber. 2013. Formulaic

sequences and academic writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, 12(3), 214-225.

Taguchi, Naoko, William Crawford, & Danielle Zawodny Wetzel. 2013. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program, *TESOL Quarterly* 47. 420–30.

Developing a first year composition L2 writing corpus and repository

Aleksandra Swatek, Hadi Banat, Shelley Staples
Purdue University

A number of student academic writing corpora (e.g., ICLE, MICUSP, BAWE) have been developed in the past few decades, showing the interest in and importance of representing this domain of language use. These corpora have been used for countless research studies, as illustrated by the extensive bibliography on the CECL and LCA websites.

Our project, the Purdue Second Language Writing corpus (PSLW), builds on this base but aims to represent the writing produced by first year international students in the U.S. in composition courses. Such courses are provided at virtually every university in the U.S., but to date no large-scale projects have been completed. Our corpus currently includes 4012 texts (3,472,260 words) representing 5 different genres (literacy narrative, proposal, annotated bibliography, interview report and argumentative essay), and we are currently processing a comparable amount of texts to be available by Summer 2016. The corpus contains three drafts of each assignment. The samples are annotated with writers' TOEFL scores, nationality, and gender, among other characteristics.

Importantly, the corpus is part of a larger interdisciplinary project that represents a collaboration among students and faculty from both applied/corpus linguistics and composition studies, called CROW (Corpus and Repository of Writing). Two main features of this larger project include the development of an online interface where scholars can eventually submit their own texts, and the inclusion of pedagogical artifacts that accompany the production of the texts, including syllabi, assignment sheets, pre-writing readings, and schema building activities. Providing these additional materials sheds light on how the texts in the corpus are developed and shaped by these instructor-designed texts. We believe that such efforts are an important way to advance corpus linguistic and language teaching research.

Our presentation will focus on two strands: the methodology for developing this new kind of corpus project, and research that has been conducted using our corpus. In terms of methodology, we will briefly cover our corpus compilation process, but focus more on the interdisciplinary practices used to guide the development of the online platform and integration of corpus texts and artifacts. We will provide a discussion of several best practices from usability design: 1) the development of persona scenarios (e.g., novice international graduate student instructor); 2) environmental scans of corpus and repository websites (e.g., MICUSP, COCA and Pedagogy Toolkit).

A number of research projects have been conducted using the PSLW corpus. We will report on the findings of one of these studies, which investigated the use of reporting verbs in students' literature reviews. Using a framework drawing on the work of Francis, Hunston, and Manning (1996), Charles (2006), and Friginal (2013), the study showed that although L2 writers in the corpus used many verbs in the semantic categories of *argue* and *show*, mostly for textual attribution, they also employed more *think* verbs than advanced L1 student writers, particularly for making general statements or to express their own opinions. After discussing our research findings, we will end the presentation by offering implications of our project for corpus development and research in general.

References

- Charles, Maggie. 2006. Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes* 25(3). 310–331. doi:10.1016/j.esp.2005.05.003. <http://www.sciencedirect.com/science/article/pii/S0889490605000529> (17 March, 2016).
- Francis, Gill, Susan Hunston & Elizabeth Manning (eds.). 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. Amsterdam: John Benjamins Publishing Company.
- Friginal, Eric. 2013. Developing research report writing skills using corpora. *English for Specific Purposes* 32(4). 208–220. doi:10.1016/j.esp.2013.06.001. <http://www.sciencedirect.com/science/article/pii/S0889490613000392> (15 January, 2016).

How learners make sense of corpus data

Kuei-Ju Tsai

National University of Kaohsiung

Since the coinage of the term “data-driven learning” (DDL) in the 1990s, much has been done to test the efficacy of learners' direct manipulation of corpus resources to assist learning. Notwithstanding the widespread interests in the learning product of corpus consultation, a dearth of studies has investigated the processes underlying such seemingly straightforward endeavors. The present study employed a mentalistic measure (i.e. concurrent think-aloud) in conjunction with a behavioristic one (i.e. corpus queries) to uncover what lay beneath the observable learner corpus consultation.

Corpus is notable for its capacity to make the otherwise elusive linguistic patterns salient, particularly habitual co-occurrences of words, such as collocations. It is thus sensible to exploit corpus resources for collocation learning/teaching. This paper reports on a study in which 17 respondents induced collocations from web-based corpus data, and verbalized their induction processes simultaneously (i.e. think-aloud) as they undertook the collocation task. The information provided by corpora is not as straightforward and organized as other reference tools, so the user may need to manipulate corpus data physically or cognitively to obtain the information needed.

Participants' concurrent verbal reports were audio-recorded, and their queries on web-based reference tools were manually documented. These two parallel datasets were then aligned to allow the researcher to infer the thinking processes going on as the respondents consulted corpora for target collocations.

Findings suggest that learners did not simply mirror what was observed in corpora, they consciously or unconsciously activated innate capacities to process the language input generated by concordancers. The learners were fairly resourceful and innovative in coordinating 1) physical mediating tools (e.g. web-based concordancers and online resources); 2) symbolic mediating tools (e.g. real language data in corpora); and 3) their own mental capacities, including cognitive and metacognitive strategies. Cognitive strategies employed to manipulate language input were *deduction*, *induction*, *substitution* and *translation*. In the course of induction learners may also group or differentiate language data to infer collocations from a wide range of knowledge sources. Metacognitive functioning was also found to regulate the undertaking of the DDL task through *planning*, *monitoring* and *evaluation*. The study provided thick description of how respondents availed themselves of the strategies above to make sense of corpus data and eventually induce the target collocations. Previous studies on corpus consultation processes are diagnostic in nature, prescribing how learners should proceed with corpus consultation from expert users' perspective. In contrast, this study is exploratory in nature, describing how learners actually fare with corpus data. Corpus consultation attempts were by no means as straightforward as prescribed by previous studies, they were highly discursive, entailing constant changes in strategies (cognitive or metacognitive) and reference resources in the light of instantaneous findings. Moreover, various external factors (e.g. task demands, reference tools, search results) and internal ones (e.g. individual differences, prior knowledge) come into play as the learner consults corpus resources. The study concluded that corpus consultation endeavors are idiosyncratic, and cannot be easily reduced to any linear, generic model.

Developing spelling with data-driven learning in L1 French

Henry Tyne

Université de Perpignan Via Domitia

Mastery of spelling in L1 French is notoriously complex and difficult (perhaps surprising so for non-natives who may wonder what all the fuss is about) and numerous publications (ranging from research papers and newspaper articles to blog pieces and forum posts) bear witness to this situation. One of the chief difficulties is that of homophony (often grammatical): there are many sets of relatively high frequency forms sharing the same sounds (give or take regional differences in pronunciation) despite different spellings and different meanings. These are often given as pairs (although there may be more than two forms potentially), as in the following: *mes* [my prl] and *mais* [but], *à* [prep] and *a* [conj vb to have]. Interestingly, while the notoriety of this problem is well flagged in the literature (and in general), solutions for teaching/learning are typically lacking (Brissaud 2011). Some specialists even go so far as to argue for spelling reform. However, for the time being, the problem

remains and large numbers of French children, even at university level, have difficulty with spelling (Brissaud 2011). This study concentrates on the homophonic morphographic set /sE/ (with the following spellings: *c'est* [it is], *s'est* [conj vb to be], *ces* [these], *ses* [his/hers prl], *sais/t* [conj vb to know]). Working with primary school children (n=161) in southern France, the use of data-driven learning for promoting awareness of morphographic differences was investigated over a short period. Children (10-11 years old) in the experimental group were invited to work autonomously in “hands-off” (Boulton 2012) mode with concordances and corpus-derived exercises; the control group used conventional teaching/learning methods. The results show that whereas initially there is no difference between the two groups (i.e. following pre-test, fewer errors are made in general at the time of learning by both experimental and control groups), there is subsequently less attrition for the data-driven learning group, with some spellings progressing even at the post-test stage. In conclusion, it is suggested that data-driven learning may be used for developing the mastery of spelling in L1 French as learners are exposed to enriched input and engage in awareness raising activities. Also, the question of how a predominantly L2-based practice can be applied to L1 pedagogy is addressed with particular attention being paid to the methodological implications.

References

- Brissaud, Catherine (2011): “Didactique de l’orthographe: avancées ou piétinements?” *Pratiques* 149-150: 207-226.
- Boulton, Alex (2012): “Hands-on / hands-off: varying approaches in data-driven learning”, in J. Thomas & A. Boulton (eds.), *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno, Masaryk University Press: 152-168.

Instructed second language acquisition and longitudinal learner corpus research: The case of lexical and syntactic complexity

Nina Vyatkina¹, Hagen Hirschmann², Felix Golcher²

¹University of Kansas, ²Humboldt-Universität zu Berlin

Our study explores the development of syntactic and lexical complexity in a written longitudinal learner corpus that comprises data collected from US university students. We address existing research gaps by focusing on 12 *ab initio* learners of a Second Language (L2) other than English (German) and utilizing dense data collection waves (17 over 4 semesters of study). The data comes from KANDEL, a freely and publicly available learner corpus (www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/research/kandel). The corpus has been annotated on multiple linguistic layers, which allows us to employ learner corpus research methods that have rarely been used in instructed L2 acquisition research.

We define complexity as “[t]he extent to which the language produced in performing a task is elaborate and varied” (Ellis 2003, p. 340). Previous research has

shown that generally, L2 complexity increases with growing proficiency but it also does not grow linearly but is subject to periodic waxing and waning as the result of the interaction among different aspects of complexity as well as different contextual variables (Bulté & Housen 2014; Mazgutova & Kormos 2015; Spoelman & Verspoor 2010; Verspoor et al. 2008).

Our study contributes to this research by focusing on fine-grained complexity measures in beginner learner language. We use multilevel modeling methods that are considered imminently appropriate for longitudinal studies but are only beginning to gain traction in L2 research (Cunnings 2012; Gries & Deshors 2015). We operationally define syntactic complexity as the system of syntactic modifiers, or optional elements extending the basic sentence frame (Graesser et al. 2004). Using parts-of-speech as proxies for syntactic modifiers (e.g., attributively used adjectives as prenominal modifiers), we recently showed that learners modified their writing from the very onset of language study but the composition of the modification system changed continuously and was characterized by a decrease in cognitively and grammatically easier (uninflected and lexical) categories and an increase in cognitively and grammatically more difficult (inflected and clausal) categories (Vyatkina et al. 2015). In this presentation, we relate previously found syntactic growth trends to the development of lexical complexity, operationally defined as lexical variety (Mazgutova & Kormos 2015). The results demonstrate distinct relationships between syntactic and lexical complexity measures in our data. For example, frequencies of attributively used adjectives increase over time as a group trend and in the writing of all individuals, and this goes hand-in-hand with the growth in frequencies of new adjectives. In contrast, the use of predicatively used adjectives as a syntactic category declines over time but the frequency of new types used by the students in this category grows. Finally, for some modifiers that do not show any discernable increase or decline (e.g., prepositional phrases), a lexical growth is still found.

Our study thus contributes to developmental profiling of L2 writing and pinpoints complexity measures appropriate for the study of L2 production of beginners. Our results are consistent with the dynamic usage-based L2 acquisition approaches (e.g., Verspoor et al. 2012) showing complex dynamic relationships between different variables as well as high inter- and intra-individual variability along with uniform group trends. We conclude with teaching implications of our research.

References

- Bulté, Bram & Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26. 42–65.
- Cunnings, Ian. 2012. An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28(3). 369–382.
- Ellis, Rod. 2003. *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavioral*

- Research Methods Instruments and Computers* 36(2). 193–202.[2028?]
- Gries, Stefan Th. & Sandra C. Deshors. 2015. EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1). 130-159.
- Mazgutova, Diana & Judit Kormos. 2015. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing* 29. 3–15.
- Spoelman, Marianne & Marjolijn Verspoor. 2010. Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics* 31(4). 532–553.[2028?]
- Verspoor, Marjolijn, Wander Lowie & Marijn van Dijk. 2008. Variability in second language development from a dynamic systems perspective. *Modern Language Journal* 92(2). 214–231.[2028?]
- Verspoor, Marjolijn, Schmid, Monica S. & Xiaoyan Xu. 2012. A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing* 21(3). 239–263.
- Vyatkina, Nina, Hagen Hirschmann & Felix Golcher. 2015. Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing* 29. 28-50.
-

The EAP teaching implications of a corpus investigation into the semantic patterning of grammatical keywords in undergraduate History and PIR essays

Karin Whiteside

University of Warwick & University of Reading

This paper reports the findings and discusses pedagogical implications of a PhD research project which adopted a corpus-driven approach to analysis of phraseologies in successful undergraduate writing across two degree programmes, History and PIR (Politics and International Relations), at two UK HE institutions.

Although neglected until relatively recently, there is now a growing body of research into the disciplinary and genre features of undergraduate academic writing (e.g. Bruce, 2010; Nesi & Gardner 2012; Hardy & Roemer 2013; Durrant 2014, 2015). The current study contributes to this knowledge by taking a bottom-up approach to the examination of lexico-grammatical features of student writing which may be linked to discipline, and by contrastively analysing student writing in subjects which sit quite closely on the disciplinary spectrum. History and PIR are relatively closely-related ‘soft’ disciplines (Becher & Trowler 2001) sharing the ‘the Essay genre family’ (Nesi & Gardner 2012) as the main genre in which students write.

The project broadly follows the approach taken in Groom’s (2007) study of disciplinary difference in professional academic writing, that of choosing salient grammatical words as the focus for close analysis. Four discipline-specific specialized corpora were created consisting of recent successful third-year writing from the two target disciplines in two institutions (the cross-institutional dimension of the study was intended to measure the extent to which factors other than discipline may explain differences found in student-level writing).

Using Wmatrix, keyness analyses of the four corpora were undertaken using as a reference corpus the 'BNC Sampler Written'. Key 'grammatical' words (Scott & Tribble 2006; Groom 2007) were manually identified and those that were key across all four disciplinary/institutional corpora (most saliently, *of*, *and*, *that*, *as* and *this*) were subject to further analysis.

A qualitative, bottom-up approach was taken to analysis of these salient grammatical words. Groom's (2007 p102) notions of 'semantic motif' and 'semantic sequence' were adopted as a starting point for categorization since the primary goal of analysis of the phraseological patterning of each key word was semantic rather than grammatical. Groupings and resulting categories for each key word were generated from close analysis of random samples of concordances (for each target key word a 300-line random sample was taken from a corpus created by pooling all four sub-corpora, and, once firm categories were established, an analysis of a 300-line random sample from each disciplinary and institutional sub-corpus was undertaken).

The paper will report and attempt to account for findings of continuities across the two disciplines as well as findings of differences between the two disciplines. The implications for teaching the essay genre within EGAP and ESAP contexts will be discussed.

References

- Becher, Tony. & Trowler, Paul R. 2001. *Academic Tribes and Territories*, 2nd edn. Buckingham: Open University Press
- Bruce, Ian. 2010. Textual and discursive resources used in the essay genre in sociology and English. *Journal of English for Academic Purposes* 9. 153-166
- Durrant, Phillip. 2014. Discipline and Level Specificity in University Students' Written Vocabulary. *Applied Linguistics* 3(35). 328-356
- Durrant, Phillip. 2015. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics* (in press)
- Hardy, Jack. & Römer, Ute. 2013. Revealing disciplinary variation in student writing: a multi-dimensional analysis of the Michigan Corpus of Upper-Level Student papers (MICUSP). *Corpora* 8(2). 183-207
- Groom, Nicholas. 2007. *Phraseology and epistemology in humanities writing: a corpus-driven study*. Birmingham: University of Birmingham PhD thesis.
- Groom, Nicholas. 2010. Closed-class keywords and corpus-driven discourse analysis. In M. Bondi and M. Scott (eds.), *Keyness in Texts*. Amsterdam: John-Benjamins. 59-78
- Nesi, Hilary. & Gardner, Sheena. 2012. *Genres across the Disciplines: Student writing in higher education*. Cambridge: Cambridge University Press
- Scott, Mike. & Tribble, Christopher. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John-Benjamins
-

Direct quotation in second language writing: A corpus-based study of intertextuality in academic learner English

Leonie Wiemeyer
University of Bremen

Intertextuality, i.e. the explicit reference to prior literature in the field, is deemed a central feature of academic writing (Hyland 2002). However, there is a notable lack of corpus-linguistic research into its manifestations (Keck 2015). This contribution aims to address this by exploring the use of direct quotation in L2 academic writing. It outlines a larger learner corpus-based project on intertextuality in L2 writing and presents a study investigating the form, function, textual integration, and attribution of direct quotes in research papers written by advanced German learners of English. Direct quotes are passages from a source text which writers integrate into their writing without changing the original wording or proposition. They are fully acknowledged according to academic citing conventions (Petrić 2012) and optionally framed by reporting structures. Quotation serves a variety of (often discipline-specific) functions (Davis 2013). Perhaps most importantly, it allows writers to “partly or wholly detach themselves from what they depict” (Clark & Gerrig 1990). While various studies have addressed direct quotation in published academic writing (e.g. Hyland 2004), such studies of L2 writing are scarce, even though academic writing competence is often judged based on the ability to cite felicitously (Shaw & Pecorari 2013). Petrić (2012) found that L2 writers of high-rated Master’s theses incorporated mostly source text fragments and quoted significantly more than their less successful peers, who tended to quote entire clauses and longer passages. Verheijen’s (2015) study revealed that due to a lack of familiarity with citing conventions, L2 writers often struggled with integrating direct quotes, made lexico-semantic and grammatical errors, and overused certain reporting structures. Though these studies offer valuable insights, little is known about how quotation differs from other intertextual strategies, how L2 writers attribute direct quotes, and whether they encode evaluation in reporting structures. This corpus-based study aims to clarify the use of direct quotes by German learners of English. Specifically, it will address the questions of 1) how these writers embed quotations into their writing, i.e. which lexical and grammatical alterations they make, 2) which reporting structures they use and whether they encode evaluation, and 3) how they attribute and reference quoted source text material. For this purpose, direct quotes will be identified in research papers from the Corpus of Academic Learner English (Callies & Zaytseva 2013) using a method proposed by Hyland (2002). Drawing on existing taxonomies (Borg 2000; Clark & Gerrig 1990; Petrić 2012; Verheijen 2015), they will be classified based on length, integration into co-text, and function. Reporting structures and strategies of embedding will also be analysed. Further analyses focus on which parts of the original text are quoted, referencing and attribution to the original author, and evaluation of quotations. Finally, direct quotation will be compared to other intertextual strategies, e.g. paraphrasing. As it has often been asserted that the appropriation of sources is challenging, especially in the L2 (Davis 2013; Hirvela & Du 2013; Pecorari 2003; Thompson, Morton & Storch 2013), this paper will also consider how corpus-based investigations of intertextuality can

inform EAP writing practice.

References

- 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.
- Clark, Herbert H. & Gerrig, Richard J. 1990. "Quotations as demonstrations". *Language* 66 (4), 764–805.
- Davis, Mary. 2013. "The development of source use by international postgraduate students." *Journal of English for Academic Purposes* 12 (2), 125–135.
- Hirvela, Alan & Du, Qian. 2013. "'Why am I paraphrasing?': Undergraduate ESL writers' engagement with source-based academic writing and reading". *Journal of English for Academic Purposes* 12, 87–98.
- Hyland, Ken. 2002. "Activity and evaluation. Reporting practices in academic writing." In: Flowerdew, John (ed.), *Academic Discourse*. Harlow: Longman. 115–130.
- Hyland, Ken. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor: University of Michigan Press.
- Keck, Casey. 2015. "Investigating textual borrowing in academic discourse: A need for a corpus-based approach". In: Cortes, Viviana/Csomay, Eniko (eds.), *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*. Philadelphia: John Benjamins. 177–196.
- Pecorari, Diane. 2003. "Good and original: Plagiarism and patchwriting in academic second-language writing". *Journal of Second Language Writing* 12, 317–345.
- Petrić, Bojana. 2012. "Legitimate textual borrowing: Direct quotation in L2 student writing". *Journal of Second Language Writing* 21, 102–117.
- Shaw, Philip & Pecorari, Diane. 2013. "Source use in academic writing: An introduction to the special issue". *Journal of English for Academic Purposes* 12 (2), A1–A3.
- Thompson, Celia; Morton, Janne & Storch, Neomy. 2013. "Where from, who, why and how? A study of the use of sources by first year L2 university students". *Journal of English for Academic Purposes* 12 (2), 99–109.
- Verheijen, Lieke. 2015. "The language of quoting in academic writing". *Dutch Journal of Applied Linguistics* 4 (1), 101–121.
-

Cognitive changes after inductive data-driven language learning

Kiyomi Yatabe, Yasuyo Minagawa
Keio University

Purpose

The present study examines the cognitive changes associated with foreign language learning in young adults in order to investigate how a particular approach such as the inductive data-driven approach and the student-centered teaching approach may affect students' language proficiency and general executive function.

Method

Seventeen university undergraduates from native Japanese speakers' families without the experience of living abroad and without the experience of learning German for more than six months took part in our experiment. One student cancelled the participation after the beginning of the experiment. Based on the prior scores of a cognitive (30 odd items in Raven's Standard Progressive Matrices (Raven, 1941)) and German test (20 questions from German Diploma in Japan (Gesellschaft zur Förderung der Germanistik in Japan, 2014;2015)), the participants were divided into two groups with roughly the same distribution of non-verbal IQ and command of German, one group who were to be taught German using the inductive data-driven teaching approach and another who were to be taught the language using the deductive teaching approach. Both groups were taught the same content with the same materials. The only difference between the two was in the type of instruction: participants assigned to the first group, the group for the inductive teaching approach, were first exposed to examples from a corpus created in-house for learners of German, and then were encouraged to discover grammatical rules by themselves, while participants assigned to the second group, the group for the deductive teaching approach, were taught grammatical rules explicitly from the start and were exposed to examples for practice only later. Before and after the three-month-long extensive training including nine class meetings and homework to be completed using iPod touches (Apple Inc.), behavioral sessions were conducted in order to test not only the participants' German grammatical judgement ability but also their ability to perform a type of reasoning which is known to require a cognitive ability with general executive function.

Results and Discussion

The participants' accuracy in identifying the grammaticality of German sentences and phrases after the training was significantly better compared to that assessed before the training ($t(16) = 5.03, p < .0005$). There was no significant difference between the two approaches in their effectiveness in helping the participants to acquire German morph-syntactic knowledge; there was no significant difference in the participants' accuracy either before ($t(15) = -0.56, p > 0.58$) or after the training

($t(11.49) = -0.48, p > 0.63$) between the two groups, and there was no significant difference between the two ($t(15) = 0.11, p > 0.91$) in the amount of improvement, either. Thus, the participants in the two groups can be said to have learned the German language's rules equally well.

However there was a significant effect of teaching approach on the participants' ability to perform relational reasoning. There was no significant difference between the two groups in the participants' accuracy in the complex reasoning test either before ($U = 27.00, p > 0.30$) or after the training ($U = 25.50, p > 0.17$). However, the amount of improvement (i.e. the post-training score minus the pre-training score) of the participants in the inductive teaching approach group was significantly larger than that of the participants in the deductive teaching approach group ($U = 15.5, p < .015$ (one-tailed)).

It is argued that this result supports the hypothesis that the inductive data-driven teaching approach contributes to the improvement in the general executive function, if not in the language proficiency suggested in the previous research (Ellis, 2006; Erlam, 2003).

References

- Ellis, Rod. 2006. Current Issues in the Teaching of Grammar: An SLA Perspective. *TESOL Quarterly* 40(1). 83–107. doi:10.2307/40264512.
- Erlam, Rosemary. 2003. The Effects of Deductive and Inductive Instruction on the Acquisition of Direct Object Pronouns in French as a Second Language. *The Modern Language Journal* 87(2). 242-260.
- Gesellschaft der Förderung der Germanistik in Japan. 2014. *Dokken kako mondaishu 2014 nenban 5kyu 4kyu 3kyu / Doitsu gogaku bungaku shinkokai hen*. Tokyo: Ikubundo Verlag.
- Gesellschaft der Förderung der Germanistik in Japan. 2015. *Dokken kako mondaishu 2015 nenban 5kyu 4kyu 3kyu / Doitsu gogaku bungaku shinkokai hen*. Tokyo: Ikubundo Verlag.
- Raven, John. C. 1941. Standardisation of Progressive Matrices, 1938. *British Journal of Medical Psychology* 19(1). 137–150.
-

Posters

KoLaS – Commented Learner Corpus of Academic German

Melanie Andresen, Dagmar Knorr
University of Hamburg

In this poster presentation, we introduce the newly released corpus KoLaS⁴ ('Kommentiertes Lernendenkorpus akademisches Schreiben', Commented Learner Corpus of Academic German). This resource was developed and collected at the *writing centre multilingualism* ('Schreibwerkstatt Mehrsprachigkeit', Knorr & Neumann 2014) at the University of Hamburg.

The *writing centre multilingualism* offers peer tutor feedback on academic texts. Students from all disciplines can get advice on all kinds of writing tasks they have to complete as part of their studies at the University of Hamburg. In return, the writing centre is allowed to carry out analyses of the texts and make them available to the academic public. The first version of the resulting corpus KoLaS has been released in December 2015. It comprises 454 texts/text versions written by 49 students between September 2011 and December 2013.

Several features make KoLaS a valuable resource:

1. The corpus comprises authentic texts written and submitted by students as assessed work.
2. Metadata on the students' languages and educational background are available. The students represent a great variety of cultural and language backgrounds (26 languages in total).
3. Many of the texts have been commented on by our peer tutors. This makes the analysis of commenting behaviour etc. possible which is highly relevant for writing centre research (e. g. von Gunten 2015).
4. Many students come to the writing centre several times and therefore the corpus includes many texts in different versions. This can be used to investigate text development and revision strategies.

The corpus has been used for research (e. g. Andresen 2016) and didactic purposes in our peer tutor training. With this poster, we wish to encourage other researchers to use the corpus for their own research or training purposes. So far, the texts

⁴ www.uhh.de/uk-kolas

are available as a collection of doc files only. In the future, we intend to make the data more easily accessible by converting it into software-independent formats and enabling online queries. Additionally, the corpus will be extended by adding further texts collected in 2014 and 2015.

References

- Andresen, Melanie. 2016. Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Sprechhandlungsverben in Wissenschafts- und Pressesprache. *Zeitschrift für angewandte Linguistik* 64(1). 47–66. doi: 10.1515/zfal-2016-0001.
- Gunten, Anne von. 2015. Wie und wozu angehende Lehrpersonen Texte kommentieren. Forschungsdesign und Analyse-Instrumente eines Projektes. *Zeitschrift Schreiben*. <http://www.zeitschrift-schreiben.eu>.
- Knorr, Dagmar & Ursula Neumann. 2014. Die Schreibwerkstatt Mehrsprachigkeit – (Lehramts-)Studierende mit Migrationshintergrund der Universität Hamburg schreiben. In Dagmar Knorr & Ursula Neumann (eds.), *Mehrsprachige Lehramtsstudierende schreiben: Schreibwerkstätten an deutschen Hochschulen*, 119–144. (FörMig-Edition 10). Münster: Waxmann.

Corpora, collocations, classroom kids: Teaching collocations to young learners at primary school

Martina Bredenbröcker
University of Paderborn

This poster presentation focusses on collocations in the primary EFL classroom. The analysis is based on the 2012 version of the Oxford Children’s Corpus (OCC), which has been compiled by Oxford University Press for use when writing children’s dictionaries. This version comprises 126 million tokens (Wild et al. 2012), including material written for 5- to 14-year old children (e.g. fiction, curriculum material, websites, magazines) and over 86 million tokens of writing by children themselves (mainly short stories collected in a BBC competition held since 2012. Using Sketch Engine (cf. Kilgarriff et al. 2004), a subcorpus (OCC-SUB) was compiled, which contains ca. 26 million tokens. It is specifically designed with this study in mind and so it contains only writing done by primary school children (ages 7 to 11).

Many previous studies (cf. e.g. Meunier/Granger 2008, Aijmer 2009, Gavioli/Aston 2001, Reppen 2010, O’Keeffe et al. 2007) have argued for a transfer of corpus-linguistic results to the EFL classroom, but this has not yet been seriously attempted for pupils at primary level.

Drawing from lemmatized frequency lists derived from the OCC-SUB, verbs and nouns are selected for a study of their collocational behaviour. The aim is to develop a basic vocabulary set to be taught in the first two years of EFL; however, the set is not based on single words, but rather on collocations, which are identified by applying different statistical measures (the MI and *t* test). The resulting set of collocations will be used as the starting point of an investigation of language-pedagogical

consequences. By applying reliable vocabulary selection criteria such as learnability, availability, familiarity, and coverage (cf. Nation 2001), the collection can be evaluated with regard to the specific needs of young learners whose L1 acquisition is not yet completed.

As one example: the verb MAKE collocates frequently and statistically significantly with a variety of object nouns in the subcorpus, e.g. with WAY, FRIEND, PLAN, WISH, DECISION, FUN, SENSE, MISTAKE, and CAKE. However, not all of these collocations are suitable for primary pupils, who first and foremost require **basic** vocabulary. Thus, teaching should focus in the early years on more concrete uses (e.g. MAKE [a] FRIEND/WAY/WISH/MISTAKE/CAKE), and less on abstract and/or delexical uses (e.g. MAKE [a] DECISION/SENSE). The latter collocations are of course also considered to be of great importance because of their frequency in native speech and their error-proneness in non-native usage (cf. Altenberg/Granger 2001); however, their teaching should be postponed until secondary school, when students are cognitively better equipped to understand such concepts in a foreign language.

References:

- Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22(2). 173-194.
- Aijmer, Karin. 2009. *Corpora and language teaching*. Amsterdam: Benjamins.
- Gavioli, Laura & Guy Aston. 2001. Enriching reality: language corpora in language pedagogy. *ELT journal* 55(3). 238-246.
- Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In Williams, Geoffrey & Sandra Vessier (eds.). *Proceedings of the eleventh Euralex Congress*. Lorient: UBS.
- Meunier, Fanny & Sylviane Granger. 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: Benjamins.
- Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom. Language use and language teaching*. Cambridge: Cambridge University Press.
- Reppen, Randy. 2010. *Using corpora in the language classroom*. Cambridge: Cambridge University Press.
- Wild, Kate, Adam Kilgarriff & David Tugwell. 2013. The Oxford Children's Corpus: Using a children's corpus in lexicography. *International Journal of Lexicography* 26(2). 190-218. Oxford: Oxford University Press.
-

Building the Trinity Lancaster Corpus of L2 spoken production

Vaclav Brezina¹, Dana Gablasova¹, Tony McEnery¹, Elaine Boyd²

¹Lancaster University ²Trinity College London

The Trinity Lancaster Corpus of L2 spoken production (Gablasova et al. 2015) which is currently being developed at Lancaster University in collaboration with the Trinity College London represents the largest corpus of its kind. It is based on examinations of spoken English conducted by the Trinity College London, a major international examination board, and contains interactions between exam candidates (L2 speakers of English) and examiners (L1 speakers of English).

The corpus allows studying individual differences between speakers and the effect of speaker background (L1, gender, age, education etc.) on their language production. At present, the corpus contains approximately 3 million running words coming from over 1700 L2 users from nine different countries (Spain, Italy, China, India, Sri Lanka, Mexico, Argentina, Brazil and Russia). The age of L2 speakers ranges from 9 to 72 years, with a balanced sample of each of the following age groups: young (under 20 years of age), middle-aged (20-30 years old) and mature (31 and older) speakers. The corpus also provides a balanced sample of L2 speech in terms of L2 proficiency, covering B1 – C2 levels of the Common European Framework. The corpus allows us to study each L2 speaker in two to four different speaking tasks. The tasks were designed to elicit language in monologic (formal presentation) as well as in highly interactive contexts (e.g. discussion, conversation). Since the exam allows the candidates to bring in their own topics for the presentation and some aspects of this topic are also discussed in the discussion, the corpus contains spoken L2 production on a great variety of topics.

The corpus, which will be made publicly available in 2017, can serve a large range of research purposes including the study of lexico-grammatical development, pragmatics of L2 speech and communicative strategies. The corpus consists of two datasets: 2012 and 2014 dataset. The former is completely transcribed and post-processed, the latter represents work in progress. The aim of the poster is:

- To introduce the project & describe the corpus structure
- To describe method of data collection, the nature of the data and speaker metadata
- To demonstrate possible directions of research
- To discuss methodological decisions connected with sampling, transcription and post-processing

References

Gablasova, Dana, Brezina, Vaclav, McEnery Tony & Boyd Elaine. 2015. Epistemic stance in spoken L2 English: The effect of task type and speaker style. *Applied Linguistics* (Advance Access)

Guangwai Lancaster Chinese Learner Corpus: Filling the gap in learner corpus research

Vaclav Brezina¹, Hai Xu²

¹Lancaster University, ²Guangdong University of Foreign Studies

So far, the research in the area of learner language has focused predominantly on English as the target language (Granger, 2012; Pravec, 2002). However, with the rise of China as a global power, Chinese as a major world language has become an increasingly popular foreign language. As a result, the number of L2 Chinese learners is growing rapidly (e.g. Dube, 2009). Yet, the specific challenges that learning Chinese presents still remain to be systematically investigated. In this context, a learner corpus can provide the much needed systematic and robust evidence about the process of L2 Chinese acquisition and help both researchers and L2 practitioners.

This poster presents Guangwai Lancaster Chinese Learner Corpus (GLC), a corpus of learner Mandarin Chinese, which consists of more than one million words (exact token count: 1,214,122). The corpus is a result of the cooperation between the Guangdong University of Foreign Studies and Lancaster University. GLC has both a spoken (588,398 tokens) and a written (625,724 tokens) part and covers a variety of task types: The written corpus data range from essays under test conditions to free compositions; the spoken corpus data consist of structured oral tests and free conversations. In terms of the interaction, the corpus includes both dialogues and monologic tasks. The corpus covers three broadly defined proficiency levels: beginner, intermediate and advanced and includes speakers from 72 different countries (L1 and cultural backgrounds).

The poster focuses on the following areas:

- Corpus design and methodology
- The nature of spoken and written L2 Chinese
- Error coding (different levels: grammatical, lexical, collocational etc.)
- Availability and release of the data

References

- Dube, C. 2009. Chinese language study is rising fast. *USC US-China Institute*.
- Granger, Sylviane 2012. Learner corpora. *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell.
- Pravec, N. A. 2002. Survey of learner corpora. *ICAME journal*, 26(1), 8-14.
-

Introducing data-driven learning to PhD students for research writing purposes

John Flowerdew, Meilin Chen
City University of Hong Kong

As a result of globalisation, universities around the world face the growing challenges of remaining competitive internationally in research. PhD students worldwide are now under pressure to publish internationally in order to enter the academic community (Li, 2002). As users of English as an additional language, novice researchers in places like Hong Kong often face linguistic difficulties or other disadvantages in getting their research published (Belcher, 2007; Curry, & Lillis, 2004; Flowerdew, 1999, 2000, 2001). However, support for PhD students in research writing in Hong Kong, mainland China and elsewhere is still inadequate (Li & Flowerdew, 2009; Kwan, 2010).

The past two decades have witnessed rapid development of machine-readable corpus applications (referred to as data-driven learning (DDL) (Johns, 1990) in EAP writing pedagogy (e.g. Charles, 2010, 2012, 2014; Cresswell, 2007; Lee & Swales, 2006; Thurnbull & Burston, 1998), due to the advances in computer technology. The effectiveness of the DDL approach nevertheless still remains a mystery to many language teachers and learners. As Ädel (2010: 41) observed, “corpus-based courses on academic writing are still at an exploratory stage”.

This presentation further explores the value of DDL in the teaching and learning of L2 academic writing, specifically, in this case, research writing. The presentation describes how half-day workshops were given to over 100 PhD students (divided into three groups) from more than 10 different disciplines at a Hong Kong university to help them improve research writing by using both online and self-built corpora. The first part of the seminar focused on introducing participants to free online corpora and training them to use these resources to solve lexico-grammatical problems they encounter during research writing. After the introduction to basic concepts about corpora, students were guided to carry out simple and slightly complex searches in online corpora, sort and break down concordances and interpret search results. In the second part, students were introduced to a freely accessible concordancing tool, *AntConc* (Anthony, 2014), using discipline-specific corpora of research articles built by one of the authors, and taught to build personal corpora that suited their specific needs. Activities were also designed to familiarize the participants with common functions of the concordancer and how it can be used to explore language patterns that serve different rhetorical functions. Through hands-on activities and discussion, students were introduced to key issues in corpus construction and started creating a corpus of their own using high-quality research articles in their own research domains.

Pre- and post-workshop questionnaires were distributed in order to collect the participants' background information and their evaluation of the approach. Feedback from the surveys shows generally positive results. The great majority of the participants were not familiar with DDL prior to the workshops, and they found the workshop of high quality, informative, and useful for improving research writing. Most participants expressed their willingness to continue to use corpora in research

writing after the workshop and would recommend the workshop to a friend. Many critical suggestions collected from the open-ended survey questions addressed the intensity of the workshop; while participants complained about the length of time involved (three and a half hours), they wanted more examples or exercises that could further show them how to improve research writing through the use of corpora. While the findings suggest that intensive introductory workshops can be an effective way of teaching PhD candidates to learn to write for publication purposes independently using the DDL approach, there is a trade-off between highly concentrated work and more time spent on more exemplification and practice. Interviews conducted two months after the workshops will provide further data on this issue and others. The introduction of this approach among PhD candidates, nevertheless, is a highly valuable and feasible alternative or additional solution for institutions in which writing support is inadequate for PhD students.

References

- Ädel, Annelie. 2010. Using corpora to teach academic writing: Challenges for the direct approach. In Campoy, Mari Carmen, Begona Belles-Fortuno, Maria Lluisa Gea-Valor (eds.), *Corpus-based approaches to English language teaching*, 39-55. London: Continuum.
- Anthony, Laurence. 2014. *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>.
- Belcher, Diane. 2007. Seeking acceptance in an English-only research world. *Journal of Second Language Writing* 16. 1-22.
- Charles, Maggie. 2010. Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In Frankenberg-Garcia, Ana, Lynne Flowerdew & Guy Aston (eds.), *New trends in corpora and language learning*, 81-104. London/New York: Continuum International Publishing Group.
- Charles, Maggie. 2012. 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus building. *English for Specific Purposes* 31(2). 93-102.
- Charles, Maggie. 2014. Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes* 35. 30-40.
- Curry, Mary Jane & Theresa Lillis. 2004. Multilingual scholars and the imperative to publish in English: Negotiating interests, demands, and rewards. *TESOL Quarterly* 38. 663-688.
- Flowerdew, John. 1999. Problems in writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing* 8. 243-264.
- Flowerdew, John. 2000. Discourse community, legitimate peripheral participation, and the nonnative-English-speaking scholars. *TESOL Quarterly* 34. 127-150.
- Flowerdew, John. 2001. Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly* 35. 121-150.
- Kwan, Becky. 2010. An investigation of instruction in research publishing offered in doctoral programs: The Hong Kong case. *Higher Education* 59(1). 55-68.

- Lee, David & John Swales. 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.
- Li, Yong Yan. 2002. Writing for international publication: The perception of Chinese doctoral researchers. *Asian Journal of English Language Teaching*, 12. 179–193.
- Li, Yong Yan & John Flowerdew. 2007. Shaping Chinese novice scientists' manuscripts for publication. *Journal of Second Language Writing* 16. 100–117.
-

Developing didactic principles for annotation based learning

Julia Hübner, Marc Schutzeichel

University of Münster

Learners often conceive grammar as an irrelevant and even boring subject. As a result they significantly lack the key competence to describe and analyze grammatical structures (Nänny 2014: 183; Habermann 2013:35).

When conveying grammatical knowledge, corpus based methods such as manual annotation by the learners still provides a new approach to access grammatical categories. The main purpose of annotation is the assignment of features and attributes to individual segments of an utterance. Acquiring knowledge about grammar is a very similar process. Therefore annotation tasks provide an appropriate means for reifying and visualizing grammatical entities and concepts. However the required didactics for this purpose are still to be defined.

Our poster aims at contributing a model (Annotation Abstraction Model; AAM) for the different abstraction levels of the annotation process of grammatical items. Our AAM combines a triad consisting of the three entities learner, teacher and theory which are linked by the annotation as a guiding principle. Teaching grammar is used as an example to demonstrate opportunities for the didactization of corpus linguistics. Learners access more sophisticated concepts and entities of grammar such as morphological primitives with increasing qualification. Our model visualizes a corresponding relation between the linguistic level of abstraction (grammatical models; theoretical framework) and the technical means (paper based annotation; computer tool). Therefore we present concrete methods for adapting the technical level used in the annotation exercises to the level of abstraction reached by the learner.

The reduction with regard to technical means and contents allows for the inductive development of language awareness “Sprachbewusstheit” (cf. Bredel 2013), thus enabling the learner to discover language as describable subject matter rather than simply as a means of communication. Moreover the model contributes to a better comprehension of didactic necessities when working with corpora in the context of teaching.

References

- Bredel, Ursula. 2013. *Sprachbetrachtung und Grammatikunterricht*. Paderborn: Schöningh.

- Habermann, Mechthild. 2013. *Von der Schule zur Universität. Zum Funktionswechsel von Grammatik im BA-Studium*. In: Köpcke, Klaus-Michael & Ziegler, Arne (eds.): *Schulgrammatik und Sprachunterricht im Wandel*. Berlin: de Gruyter.
- Nänny, Stephan. 2014. *Grammatisieren - sprachliche Strukturen erforschen*. In: Bredel, Ursula & Schmellentin, Claudia (eds.). *Welche Grammatik braucht der Grammatikunterricht?* Baltmannsweiler: Schneider Verlag Hohengehren. 183-193.

Exploring the semantic prosody of attributive verbs in academic texts

Hsueh-chao Marcella Hu & Shu-jung Wu

The Overseas Chinese University, Taiwan

Semantic prosody, arising from the phraseological tradition that focuses on the typical behaviour of individual lexical items, has attracted considerable research attention in the past two decades since Sinclair (1991) claimed that “many uses of words and phrases show a tendency to occur in a certain semantic environment” (p. 112). Borrowing the concept from Firth’s discussion on phonological items, Louw (1993) first introduced the term ‘semantic prosody’ by claiming that the habitual collocates of a lexical item is established through the semantic consistency of its subjects. Semantic prosody has thus been closely related to collocation learning in language acquisition research. In the context of collocation learning, near-synonyms particularly pose a difficulty for most foreign language learners due to their similar denotational meanings but un-interchangeable semantic prosody (Xiao & McEnery, 2006).

In the context of academic writing, some attributive verbs may have similar meanings but differ in the degree or density of that meaning (Hunston, 1995, 2002, 2007). The present corpus-based study aims to compare and contrast the function and usage of some near-synonymous attributive verb pairs (i.e., *suggest* vs. *indicate*, *confirm* vs. *support*, *argue* vs. *discuss*) by examining the academic texts in the field of applied linguistics. The attributive verbs were chosen based on a list from the writing center at the University of Texas (<http://uwc.utexas.edu/wp-content/handouts/Verbs-of-Attribution.pdf>). They were also checked against in the Merriam-Webster Online Dictionary of Thesaurus (<http://www.merriam-webster.com/thesaurus/>) and Collins Online Dictionary (<http://www.collinsdictionary.com/english-thesaurus>) to ensure they share similar meanings. The self-created corpus consists of 507,515 words derived from 86 pieces of peer-reviewed academic papers in well-renowned journals (i.e., *Language Teaching Research*, *Language Learning*). The words being compared occurred no less than 10 times in the corpus and appeared in 5 or more pieces of texts. At least 5 examples of a specific verb within a co-text of no less than 20 words were used to analyze the data before a conclusion was drawn. As evidenced in the concordance lines, the preliminary analyses indicate the following findings for the three pairs:

Suggest vs. Indicate

The subject position of SUGGEST and INDICATE is usually some research findings (e.g., *results* and *data*) or studies (e.g., *research* and *analysis*). When used to report results from previous studies, SUGGEST tends to take the author of the attributed study as subject whereas INDICATE almost always take the piece of work as the subject. Furthermore, the subtle difference construes more transitivity in favor of SUGGEST, but not for INDICATE.

Support vs. Confirm

Both CONFIRM and SUPPORT can be used to give testimony to the truth of something, research findings/hypotheses in particular. However, several differences were found as to how they are used in the PA Corpus. Statistical terms and were found to appear exclusively in the co-text of CONFIRM. Also, while CONFIRM allows overt negatively evaluated complements, this is not the case for SUPPORT.

Argue vs. Discuss

ARGUE often implies conviction with evidence or reasons offered nearby while DISCUSS is a neutral word. Finally, the analysis generally confirmed Hunston (1995) but discovered an exception where a statement attributed by ARGUE is countered by another statement attributed by ARGUE serving as the final move in a conflict.

The purpose of the study was not only to examine the density and degree of the synonymous attributive verbs, hoping to offer insights into expert performance and contribute to classroom practices. Some implications will also be discussed.

References

- Hunston, Susan. 1995. A corpus study of some English verbs of attribution. *Functions of Language* 2. 133-158.
- Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2). 249-268.
- Louw, Bill. 1993. "Irony in the Text or Insincerity in the Writer? In M. Baker (ed.), *The Diagnostic Potential of Semantic Prosodies*." *Text and Technology: In Honour of John Sinclair*, 157-176. Amsterdam: John Benjamins.
- Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Xiao, Richard, & McEnery, Tony. 1996. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1). 103-129.

The DGS-L2C: Creation of a German Sign Language learner corpus at the University of Cologne

Thomas Kaul, Alejandro Oviedo, Reiner Griebel, Leonid Klinner & Monika Krumpfen
University of Cologne

Our corpus will comprise annotated videos from learners of German Sign Language (henceforth DGS) and will provide data for applied researching as well as a tool to improve the university teaching of DGS as L2.

Sign language corpora

Sign languages of the Deaf and spoken languages, despite having different transmission channels, share similar structures as well as equivalent acquisition and learning processes (Meier et al. 2009).

Sign language corpus linguistics is a relatively recent development (Felon et al. 2015). Automatic processing of signed data was only made available by software produced within the last 10 years, permitting the synchronization of videos with written annotations (Crasborn & Slotjes 2008). Currently, some research groups worldwide are working on sign language corpora (Felon et al. 2015). Among them are at least three running projects dedicated to learner data (Schönström & Mesch 2014).

DGS

DGS is the officially recognized natural language of approximately 80,000 deaf people in Germany (Hase 1996). DGS has received linguistic descriptions since 1989 (Eichmann et al. 2012) and is taught in numerous institutions across the country.

The DGS-L2C

The University of Cologne offers A1 to C1 (Council of Europe, 2001) DGS-levels distributed among 64 courses, which are attended by around 1,000 students every year. After completion of the A1 and A2 levels, students perform a series of tests in contexts of reception, production and interaction, a portion of which have been recorded since 2013. This material (about 22 hours in 230 individual files comprised of individual productions –induced by a drawing– as well as dialogues between deaf teachers and informants) constitutes the primary DGS-L2C data. Metadata related to the videos includes age, gender and hearing status of the informants as well as the course visited, semester of data collections and scores received.

Videos are attached to basic annotation files in the software ELAN (Crasborn & Slotjes 2008). Basic annotations are synchronized to the videos and consist of parent tiers containing a German translation as well as one ID-Glossing.

An ID-Glossing is a surrogate of the citation form of a sign and assumes the written form of the German word(s) historically related to the basic meaning of that sign (cfr. Johnston 2010). In the DGS-L2C, ID-Glossings are contrasted with the WebDGS, a vocabulary of around 8,000 entries developed in our University.

Differences observed between signs in the DGS-L2C and the corresponding models of WebDGS are labelled for future research as a deviation of the native-signer norm.

A series of child tiers derivates from the ID-Glossing to describe the activity of manual and nonmanual articulators (i.e. left hand, eye gaze, nose, mouth patterning, etc.). For ease of annotation, every child tier is attached to a controlled vocabulary (i.e. „eye gaze“: up, down, up-left, etc.).

Currently development stand

We are now creating the individual files in ELAN. After completing this phase (Februar 2016) we will begin with the annotations of every single file. This second phase should be completed in two years (Februar 2018).

References

- Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Press Syndicate of the University of Cambridge.
- Crasborn, Onno & Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In Onno Crasborn, Thomas Hanke, Eleni Efthimiou, Inge Zwitterlood & Ernst Thoutenhoofd (eds.), *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*, 39-43. Paris: ELDA.
- Eichmann, Hanna, Martje Hansen & Jens Hefmann (eds.). 2012. *Handbuch Deutsche Gebärdensprache. Sprachwissenschaftliche und anwendungsbezogene Perspektiven*. Seedorf: Signum.
- Felon, Jordan, Adam Schembri, Trevor Johnston & Kearsy Cormier. 2015. Documentary and corpus approaches to sign language research. In Eleni Orfanidou, Bencie Woll & Gary Morgan (eds.), *Research methods in sign language studies. A practical guide*, 156-172. West Sussex: Wiley/Blackwell.
- Hase, Ulrich. 1996. Zur aktuellen Situation gehörloser Bürgerinnen und Bürger in Deutschland. In Peter Donath, Ulrich Hase, Siegmund Prillwitz & Karin Wempe (eds.), *Eine Minderheit verschafft sich gehör. Textdokumentation zur Anerkennung der Gebärdensprache Gehörloser*, 79-88. Hamburg: Signum.
- Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1). 106-131.
- Meier, Richard, Kearsy Cormier & David Quinto-Pozos (eds.). 2009. *Modality and structure in signed and spoken languages*. Cambridge: Cambridge University Press.
- Schönström, Krister & Johanna Mesch. 2014. „Use of nonmanuals by adult L2 signers in Swedish Sign Language – Annotating the nonmanuals. In: Onno Crasborn, Eleni Efthimiou, Stravroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette H. Kristoffersen & Johanna Mesch (eds.), *Beyond the manual channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages*, 153-156. Reykjavik: ELRA.

Towards a *Discipline-Specific Vocabulary Core* for postgraduate Law students: reporting on a pilot study

Jenny Kemp

University of Leicester

Studies have demonstrated that university students require discipline-specific vocabulary, including specialised meanings of more widely used lexical items (see e.g. Hyland & Tse, 2007). Indeed Durrant (2014, p. 11), in his investigation of BAWE (British Academic Written English) subcorpora, found that subject-specific lexis provided a mean coverage of text of 18%. If learners do not know this vocabulary, it is likely to significantly hinder their reading comprehension as well as their writing. It is therefore important that teachers of English for Specific Academic Purposes are aware of the specialised lexis their students need.

The aim of this doctoral research project is to identify the discipline-specific vocabulary required by postgraduate Law students for reading. To this end I am in the process of building a law corpus. Existing law corpora tend to focus on prescriptive texts such as law reports (e.g. Williams, 2007; Marín & Rea, 2012); yet these only form a part of students' reading needs. The corpus in this study will be much more representative of their needs as it is based on the texts specified in module reading lists and includes a variety of genre types.

As we now have a greater understanding of the phraseological tendency of language, it is not enough to focus only on single words: there is a need for a list which includes “an appropriately disciplinary-sensitive repertoire of bundles” (Hyland, 2008, p. 8) in the form of collocations and multiword units (MWUs), as well as single words. Yet to date research has tended to focus on either single words or MWUs. The Discipline-Specific Vocabulary Core (DSVC) will be an innovative list as it will contain both. However, as a result it will be challenging in terms of the methodology used.

This poster presentation reports on the methodology and findings of the pilot study involving texts in the domain of World Trade Organisation Law. A DSVC list was extracted from the corpus by combining computational methodology with expert judgments (see Simpson-Vlach & Ellis, 2010, and Ackermann & Chen, 2013). The poster will present and critically evaluate the methods used and highlight the challenges faced in the attempt to produce a balanced and representative DSVC.

References

- Ackermann, K. & Chen, Y.-H., 2013. Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12. 235-247.
- Durrant, P., 2014. Discipline and level specificity in university students' written vocabulary. *Applied Linguistics* 35(3). 328-356.
- Hyland, K., 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes* 27. 4-21.

- Hyland, K. & Tse, P., 2007. Is there an "academic vocabulary"? *TESOL Quarterly* 41(2). 235-253.
- Marín, M. J. & Rea, C., 2012. A corpus-based proposal for the grading of vocabulary teaching materials for the Legal English class. *Journal of Second Language Teaching and Research* 3(1). 24-36.
- Simpson-Vlach, R. & Ellis, N., 2010. An Academic Formulas List: New methods in phraseology research. *Applied Linguistics* 31(4). 487-512.
- Williams, C., 2007. *Tradition and Change in Legal English: Verbal constructions in prescriptive texts*. 2nd edn. Bern: Peter Lang.

Corpora for engineers writing in a foreign language: Methods and applications for Language for Specific Purposes corpora

Marina Kogan¹, Tobias Gärtner², Sigrun Schroth-Wiechert²

¹Peter the Great St. Petersburg Polytechnic University. ²Leibniz Universität Hannover

One area of languageteaching is of a special interest for language teachers and researchers: Language (in particular, English) for Academic Purposes. Data Driven Learning (DDL) is considered to be a very prospective method in scientific and academic writing areas. A number of publications have been devoted to studies in this area. They suggestthat students using corpus data perform better than students taught with vocabulary listsand dictionaries, although results were not conclusive in all. Studies, which quantify the success rate of error correction/self-correction, show more positive outcomes, suggesting that students are able to apply corpus data to solving language problems (e.g., [Gaskell & Cobb, 2004]). The results of these studies are promising. Using a corpus as a reference source for writing academic English can be ineffective and demotivating if the corpus does not contain examples of language use in students' specific technological/scientific areas [Chang, 2014].

The overwhelming majority of studies on using DDL in language teaching in general and in teaching Language for Specific Purposes (LSP) in particular consider issues related to English language acquisition. Boulton's comprehensive survey (2010) of 93 papers English was not a target language in 11, with only 3 studies devoted to the acquisition of German and none to Russian. So, it can be concluded that DDL is a much less popular method in teaching other languages than English. Needless to say that all language learners need as much assistance when they struggle to produce a serious outcome in a target language (e.g. to write Master thesis in Engineering) as learners of English do.

With the Deutsch, English and Russki – Corpus of Civil Engineering we have developed a new resource for engineering students learning either of the three languages. The corpus and newly programmed concordancer (HanConc) are intended to step in where user specific solutions are needed. They help to overcome the knowledge gap between language students and teachers. Our poster will present the first results of the first trial phase.

References

- Boulton, Alex. 2010. Learning outcomes from corpus consultation // In Jaén, Maria Moreno, Valverde, Fernando Serrano, & María Calzada Pérez (eds),

Exploring New Paths in Language Pedagogy: Lexis and corpus-based language teaching. London: Equinox. 129–144.

Chang, Ji-Yeon. 2014. The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL* 26(2). 243–259.

Gaskell, Delian & Cobb, Thomas. 2004. Can learners use concordance feedback for writing errors? *System*. 32 (3). 301–319.

Building a corpus of written academic texts in Portuguese

Tanara Zingano Kuhn^{1,2}, José Pedro Ferreira²

¹University of Lisbon ²University of Coimbra

This contribution reports on the process of compilation of a corpus specially designed for an online, corpus-driven dictionary of academic Portuguese for university students⁵. Users of this dictionary will be speakers of Portuguese as a mother tongue and as an additional language who attend a variety of courses (within different areas of knowledge) at different levels of study (undergraduation and graduation) and use Brazilian and European varieties of Portuguese. This synchronic corpus aims to reflect academic language use by expert writers in Brazil and Portugal. Thus, it is composed of written texts extracted from peer-reviewed, open source online journals of distinct disciplines. Based on the principles of Corpus Linguistics (McEnery *et al.*, 2006; Sinclair, 2003, 2005), the process of compilation involved a series of actions. First, texts were extracted from the Internet and converted into .txt format. Next, corpus clean-up involved biographical information, footnotes, bibliography and abstracts removal. In addition, verification of texts' language resulted in confirmation of language variety (Brazilian Portuguese or European Portuguese) and exclusion of texts written in other languages. Then, texts were distributed in four broad areas according to its discipline of origin. This classification into broad areas followed the disciplinary groups division used in the British Academic Written English (BAWE) corpus, namely: Arts and Humanities; Life Sciences; Physical Sciences; and Social Sciences. After that, metadata were gathered and headers with information on text genre (article, interview, report, and review), language variety, disciplinary area, author(s), year and journal of publication were given. Next step was automatic corpus annotation, consisting of tokenisation, part-of-speech (POS) tagging and lemmatisation. Finally, detection of original documents' character encoding and conversion to UTF-8 (see Kilgariff *et al.*, 2012) completed the process of corpus building.

References

Heuboeck, Alois, Jasper Holmes & Hilary Nesi. 2010. *The BAWE corpus manual v. III*. Retrieved from [http://www.coventry.ac.uk/Global/05%20Research%20section%20assets/Research/British%20Academic%20Written%20English%20Corpus%20\(BAWE\)/Microsoft%20Word%20BAWEmanual%20v3%20%20BAWEmanual%20v3.pdf](http://www.coventry.ac.uk/Global/05%20Research%20section%20assets/Research/British%20Academic%20Written%20English%20Corpus%20(BAWE)/Microsoft%20Word%20BAWEmanual%20v3%20%20BAWEmanual%20v3.pdf) (29 November, 2015.)

⁵The first author is a Capes scholarship holder, process number 0973/13-0.

- Kilgarriff, Adam, Jan Pomikalek, Miloš Jakubiček & Pete Whitelock. 2012. Setting up for corpus lexicography. In Fjeld, Ruth Vatvedt & Julie Matilde Torjusen (eds.), *EURALEX* (European Association of Lexicography), 778–785. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
- McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: an Advanced resource book*. Abingdon: Routledge.
- Sinclair, John. 2003. Corpus processing. In Sterkenburg, Piet van (ed.), *A practical guide to lexicography*, 179-193. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Sinclair, John. 2005. Corpus and text - basic principles. In Wynne, Martin (ed.), *Developing linguistic corpora: a Guide to good practice*, 1-16. Oxford: Oxbow Books. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm> (10/01/2016)
-

CollGram profiles and n-gram frequencies as gauges of phraseological competence in EFL learners at different proficiency levels

Agnieszka Leńko-Szymańska

University of Warsaw

The study of phraseology has recently become one of the most active fields of second language acquisition. However, the literature does not as yet offer a thorough account of the process of development of phraseological competence in EFL learners. Most of the studies concentrate on advanced students (e.g. De Cock 2000; Nesselhauf 2003; Chen & Baker 2010). There are almost no accounts of formulaic language emerging at early stages of learning (see Vidakovic and Barker 2010; and Lenko-Szymanska 2012 for counterexamples).

The methodology applied by corpus-based studies of L2 phraseology consists in comparing frequencies of n-grams retrieved from learner data at two or more proficiency levels. These studies do not analyse the use of phraseology in individual learner texts, but examine them holistically in a corpus or subcorpora. Such approach ignores individual variation in the use of multi-word units. In addition, it does not consider the strength of co-occurrence of the retrieved items.

A new method of studying phraseology was recently proposed by Bestgen & Granger (2014). CollGram is “a technique that assigns to each pair of contiguous words (bigrams) in a learner text two association scores (mutual information and t-score) computed on the basis of a large reference corpus.” It produces three measures, which together form a CollGram profile, and which, according to the authors, “quantify the collocation strength of each text” (p. 31). The profile includes: the mean MI value; the mean t-score; and the proportion of idiosyncratic bigrams that are absent from the reference corpus.

So far CollGram has been applied to trace the development of phraseological competence in intermediate and advanced learners (Granger & Bestgen 2014; Bestgen & Granger 2014). The purpose of this study is to analyse learner data from a wider range of proficiency levels. In addition, the study aims to compare the two

measures – CollGram profiles and n-gram frequencies – as instruments for describing and comparing phraseological competence of learners at these levels.

The data analysed in the study were drawn from the ICCI and ICLE corpora. The International Corpus of Crosslinguistic Interlanguage is a collection of essays written by primary and secondary school students (Tono et al. 2012). 90 essays were selected from the Polish, Spanish and Austrian sections of the corpus representing a range of grades. To complement the ICCI data with text written by more advanced learners, the International Corpus of Learner English was used (Granger et al. 2002). 30 essays were drawn from the Polish, Spanish and German sections of this collection.

The 120 essays analysed in this study were rated on the CEFR scale (Council of Europe 2001) by three raters. Next, a CollGram profile was computed for each essay using COCA as a reference corpus. Finally, each essay was searched for the occurrence of 1825 3-grams appearing at least 10 times in COCA. The CollGram profiles and 3-gram frequencies were correlated using Pearson product-moment coefficients. Additionally, changes in the values of the two measures for six CEFR proficiency levels were analysed using ANOVA.

The presentations will discuss the results of the analyses which demonstrate that both measures are valuable instruments in tracing the development of phraseological competence in EFL learners. It will also present the pattern of growth in the use of native phraseology by learners at different proficiency levels.

References

- Bestgen, Y., & Granger, S. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26: 28–41. <http://doi.org/10.1016/j.jslw.2014.09.004>
- Chen, Y. H. & Baker, P. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14 (2): 30–49.
- Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- De Cock, S. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (eds). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi. 51–68.
- Granger, S., & Bestgen, Y. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52(3): 229–252. <http://doi.org/10.1515/iral-2014-0011>
- Granger, S., Dagneaux, E. Maunier, F. & Paquot, M. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.
- Lenko-Szymanska, A. 2012. The Emergence of Formulaic Language in EFL Writing – A Cross-linguistic and Cross-sectional Perspective. Paper presented at 10th International Conference on Teaching and Language Corpora TaLC10, 12-14 July 2012, Warsaw.
- Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24: 223–242.

- Tono, Y., Kawaguchi, Y. & Minegishi, M. (eds). 2012. *Developmental and Cross-linguistic Perspectives in Learner Corpus Research*. Amsterdam and Philadelphia, PA: John Benjamins.
- Vidakovic, I & Barker, F. 2010. Use of words and multi-word units in Skills for Life Writing examinations. *Cambridge ESOL: Research Notes*, 41: 7–14.
-

Developing a culturally annotated corpus for foreign language teaching

Çalışkan Nihal

Yıldırım Beyazıt University

In this poster presentation, we want to report on the procedures that we followed to develop a culturally annotated small corpus of Turkish data. The role of culture in foreign language teaching has been recognized and widely discussed among scholars during the past half century (Brooks 1968, Valdes 1986, Byram 1989, Byram and Esarte-Sarries 1991, Byram et al. 1994, Kramsch 1991, Hinkel 1999, Byram and Grundy 2003). One of the very practical concepts to understand the cultural elements of language teaching is the distinction made as Big “C” culture and little “c” culture (Brooks 1968, Seelye 1993). While big “C” culture elements are explicit constituents of culture, such as famous writers, musicians, historical figures, places to see etc., the little “c” culture elements are more implicit, like how people react in specific situations, daily routines, the food types they consume, the tiny things they give importance to, etc. It is a hot topic how to integrate especially these little “c” culture in foreign language teaching (Hendon 1980, Galloway 1985, Lafayette 1988, Seelye 1994, Herron et al. 1999, Stephens 2001, Allen 2004, Akpınar and Öztürk 2009, Bueno 2009, Lee 2009). We believe that a culturally annotated corpus can help learners become more acquainted with the cultural elements of foreign language in question. However, this type of annotation isn’t an easy task. It is very difficult to assign C-tags (culture tags) to linguistic items, since the probable tokenization process is quite complicated due to the pitfalls of defining boundaries. Bearing in the mind this complexity, we decided to compile a culturally sensitive corpus. This type of corpus can be created collecting texts directly addressing the cultural issues. Our source is Ekşi Sözlük (Dictionary of Sourstimes), a popular collaborative hypertext dictionary in Turkish (Gürel and Yakın 2007). There is no limit on the number and the nature of entries made by individual authors. The entries are generally in the form of a dictionary entry referring to the heading. This dictionary gives a great amount of data about Turkish culture. Some of the headings are listed to give an idea: The Turkish type academics, raising a child in a Turkish style, Turkish serials, the barking women in Turkish films, the things that Turkish people can never learn, the hospitality of Turkish people, etc. We compiled a 10.000+- words small size corpus extracting data from the headings that contain the phrase “Turkish people, Turkish life, and Turkish culture”. The annotation scheme developed in this study is hierarchically structured and comprised of two nodes, the big “C” and little “c” culture elements of Turkish. Each dictionary entry is given a C-tag. The annotation process is still in operation. We aim to annotate the whole corpus until the end of April 2016. We will then implement it in the classroom, and it will be possible

to share some findings about how learners react this kind of data. We believe that this poster presentation will suggest some technics to compile a culturally annotated corpus and using it in foreign language classrooms.

References

- Akpınar, K. D., & Öztürk, M. S. (2009). Teaching culture and language through folktales. *Ekev Academic Review*, 13(40), 69-84.
- Allen, L. Q. (2004). Implementing a culture portfolio project within a constructivist paradigm. *Foreign Language Annals*, 37(2), 232-239.
- Brooks, N. (1968). Teaching Culture in the Foreign Language Classroom. *Foreign Language Annals*, 1:3, 204-228.
- Bueno, K. A. (2009). Got film? Is it a readily accessible window to the target language and culture for your students? *Foreign Language Annals*, 42(2), 318- 339.
- Byram, M. (1989). *Cultural studies in foreign language education*. Clevedon, England: Multilingual Matters.
- Byram, M., & Esarte-Sarries, V. (1991). *Investigating cultural studies in foreign language teaching: A book for teachers*. Clevedon, England: Multilingual Matters.
- Byram, M., & Grundy, P. (2003). *Context and culture in language teaching and learning*. Tonawanda, NY: Multilingual Matters.
- Byram, M., Morgan, C., & Colleagues. (1994). *Teaching and learning: Language and culture*. Bristol, PA: Multilingual Matters.
- Galloway, V. B. (1985). *A design for the improvement of the teaching of culture in foreign language classrooms*. ACTFL project proposal.
- Gürel, E. & Yakın, M. (2007). "Ekşi Sözlük: Postmodern Elektronik Kültür". *Selçuk İletişim-Selçuk Üniversitesi İletişim Fakültesi Dergisi*. Cilt: 4. Sayı: 4: 203-219.
- Hendon, U. (1980). Introducing culture in the high school foreign language class. *Foreign Language Annals*, 13(3), 191-199.
- Herron, C., Cole, S., Corrie, C., & Dubreil, S. (1999). The effectiveness of a video-based curriculum in teaching culture. *The Modern Language Journal*, 83(4), 518-533.
- Hinkel, E. (1999). *Culture in second language teaching and learning*. Cambridge, U.K: Cambridge University Press.
- Kramsch, C. (1991). Culture in language learning: A view from the United States. In K. De Bot, R. B. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 217-240). Philadelphia, PA: John Benjamins.
- Lafayette, R. C. (1988). Integrating the teaching of culture into the foreign language classroom. In A. J. Singerman (Ed.), *Toward a new integration of language and culture* (pp. 47-62). Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages.
- Lee, L. (2009). Promoting intercultural exchanges with blogs and podcasting: A study of Spanish-American telecollaboration. *Computer Assisted Language*

- Learning*, 22(5), 425-443.
- Seelye, H. N. (1993). *Teaching culture: Strategies for intercultural communication* (3rd ed.). Lincolnwood, IL: National Textbook Company.
- Stephens, J. L. (2001). Teaching culture and improving language skills through a cinematic lens: A course on Spanish film in the undergraduate Spanish curriculum. *ADFL Bulletin*, 33, 22-25.
- Valdes, J. M. (1986). Culture in literature. In J. M. Valdes (Ed.), *Culture bound* (pp. 137-147). New York, NY: Cambridge University Press.
-

Translation and Do-it-Yourself Corpora used as pedagogical tools in EAP activities

Paula Tavares Pinto, Anna Luisa Lopes Alves
Universidade Estadual Paulista “Júlio de Mesquita Filho”

Over the past two decades, the use of corpora for teaching and translation pedagogy has been consolidated as a productive research area. In classifying the pedagogical corpus applications (i.e. the use of corpus tools and methods), a useful distinction can be made between direct and indirect applications. This means that, ‘indirectly’, corpora can help with decisions about what/how/when to teach a particular topic, but they can also be accessed ‘directly’ by learners and teachers in the language or translation classroom, so as to assist students during their learning processes (Leech, 1997). In this context, the continuous release of scientific publications in English has demanded special attention by researchers in general who, when writing their own articles or requiring their translations to specialized professionals, make the study of academic genre and scientific language a vast field to be described. Nowadays, the theoretical and methodological approach of Corpus Linguistics has favored the investigation of large amounts of texts with the aid of computational tools developed for the purpose of Linguistics research. Based on this approach, we compiled a corpus of original scientific papers and their translations in the field of Medicine. This corpus, called *MedCorp*, was the basis for the analysis of technical translation style and the compilation of three digital glossaries. The emphasis was given to the analysis of language and terminology used by professional translators (Paiva et al. 2008; Paiva, 2009). One of the results showed that the use of passive voice, reported by Johns (1991) as troublesome in the interpretation of abstracts written in English by Brazilian researchers, was not as problematic in the abstracts translated by professionals. One of the hypotheses was due to non-professional translations or to the unawareness, by Brazilian researchers, of the structural and lexical constraints related to the patterns used by the international scientific community. Based on this hypothesis, we proposed a study aimed at analyzing the scientific language produced by Brazilian researchers who use English as “additional language” (Swales and Feak, 2009) or, rather, the language for international publication. The main objective was to compare the language produced by these researchers to the language produced by researchers from international institutions who have published in high-impact journals worldwide. In order to compile the corpus of texts written by Brazilian researchers, we offered an English for Academic Purposes Course for three years in a row, in order to have the abstracts in the areas of Humanities,

Biological Studies and Hard Core Sciences produced by the students. Besides the compilation of this “learner corpus” (Granger, 2002), students compiled their own personal Do-it-Yourself (DIY) corpora so that, during classes, they discussed the specificities and lexical patterns of each different area before producing their own abstracts. As a result, students have become aware of rhetorical functions as well as the terminology and lexical patterns used in each distinct area of research. As examples, we will present considerations about the use of rhetorical moves, lexical patterns regarding vocabulary of “urgency” and “importance” and the passive voice in the three areas. The results of this study have been used as guidance in teaching EAP classes.

References

- Granger, Sylviane. 2002. A bird’s-eye view of learner corpus research. *Computer learner corpora, second language acquisition and foreign language teaching*. 3-33.
- Johns, Tim. 1992. It is presented initially: linear dislocation & interlanguage strategies in Brazilian academic abstracts in English and Portuguese. *Ilha do Desterro*. J. L. Meurer (ed). 1992. Florianópolis: Universidade de Santa Catarina, n. 27, 9-32.
- LEECH, Geoffrey. 1997. *Introducing corpus annotation*.
- Paiva, Paula T. P. Camargo, Diva C. Xatara, Cláudia M. 2008. A reflection on the elaboration of a preliminary bilingual lexicon of cardiology based on the terms used in a parallel and in two comparable corpora. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, v. 24, n. 1. 1-22.
- Paiva, Paula T. P. 2009. *Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em corpus*. PhD Dissertation. Universidade Estadual Paulista “Júlio de Mesquita Filho”.
- Swales, John M. Feak, Christine B. 2009. *Abstracts and the Writing of Abstracts*. Michigan: University of Michigan Press.

Just another brick – Reflections from PAC-ToE and ICE-IPAC in the teaching/learning of oral English for non-native learners.

Anne Przewozny-Desriaux¹, Nadine Herry-Bénit²

¹University of Toulouse, ²University of Paris VIII

Although there has been increasing interest in learner corpora (Durand, Gut and Kristoffersen 2014, Ballier and Martin 2015), most research on non-native speech is not directly concerned with pedagogical issues except such works as Detey *et al.* in French (2009, 2010, PFC Corpus) or Trouvain and Gut (2007, LeaP corpus). It has been acknowledged that a corpus-linguistic approach can supplement current research methods in second language learning (Biber, Conrad, and Reppen 1998, Botley *et al.* 1996, Kettemann and Marko 2002, Granger, Hung and Petch-Tyson 2002, Sinclair 2004, Granger 2004, Trouvain and Gut 2007). Non-native speech

corpora have been in use in the classroom for a few years, while the recognition of their pedagogical value is growing (Ghadessy, Henry and Roseberry 2001, Kettmann and Marko 2002, Granger, Hung and Petch-Tyson 2002, Sinclair 2004; Detey et al. 2010).

One of the long-term goals of the PAC programme (with recordings from 31 English-speaking locations, Durand and Przewozny 2012, 2015) is to bring together researchers and teachers with a shared interest in the description and teaching of the phonology of a second language. In this paper, the PAC-ToE (PAC-Teaching of English) and the brand new ICE-IPAC corpora (Interphonology of Contemporary English in PAC) enable us to deepen the analysis of the segmental and suprasegmental systems of native and non-native speakers. From a technical point of view, following Ballier and Martin's typology of spoken (learner) corpora (2013, 2015), the PAC corpus falls into the category of a multilayer annotated phonetic corpus with concerns in diatopic, diastratic and diaphasic variation. The ICE-IPAC corpus is a phonological learner corpus (Granger 2002) of English across varieties, launched in 2013. It offers to look at sociophonological variation as observed in learners' speech situations.

We present the first outcomes of PAC-ToE and ICE-IPAC data, as both ranges of data are closely interconnected when dealing with the teaching of oral English to French learners in an academic context. After having explained the respective goals of PAC-ToE and ICE-IPAC, our paper focuses on a number of web-based facilities and pedagogical resources for teachers, researchers, learners and advanced students of English (from phonology to oral syntax) in an academic context (from the practice of comprehension and oral skills in General English studies and English for Special Purpose, to secondary education competitive examinations and in-service training for teachers of English, Przewozny 2015). We focus on the issues of foreign speech assessment ("always prone to subjectivity", Carranza et al. 2014) and recurrent pronunciation errors among learners. Doing so we evaluate how the PAC-ToE and ICE-IPAC corpora may add a brick to improve the teaching and learning of oral English on the basis of a scientific description of authentic contemporary English.

References

- Andreassen, H. N., Herry-Bénil, N., Kamiyama, T. & Lacoste, V. 2015. The ICE-IPAC project: Testing the protocol on Norwegian and French learners of English. (ICPhs), 18th International Congress of Phonetic Sciences, Satellite meeting: Workshop on Phonetic Learner Corpora, 10-14 August 2015 SECC Glasgow Scotland UK.
- Ballier, N. & Martin, P. 2015. Speech annotation of learner corpora. In Granger S., F. Meunier & G. Gilquin (eds.), *Handbook of learner corpus research*, Cambridge: CUP, 107-134.
- Ballier, N. & Martin, P. 2013. Developing corpus interoperability for phonetic investigation of learner corpora, in Díaz-Negrillo, A., Ballier, N. and Thompson, P. (eds.), *Automatic treatment and analysis of learner corpora*. Amsterdam: Benjamins, Studies in corpus linguistics series, 33-64.
- Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus linguistics: investigating language*

- structure and use*. Cambridge: Cambridge University Press.
- Bonfiglioli, A., Trifu, I & Herry-Bénil, N. 2015. ICE-IPAC: An acoustic description of French learners' speech productions in English: Towards a classification of deviant forms. International conference: Language in contrast diachronic, variationist, and cross-linguistic studies of English, 4-5 Dec. 2015, University of Paris VIII and Paris X.
- Botley, S., Glass J., Mc Enery T., & Wilson A. (eds.) 1996. *Teaching and Language Corpora 1996*. Lancaster: UCREL technical papers volume 9.
- Brulard, I., Carr, P. & Durand, J. (eds.) 2015. *La Prononciation de l'anglais contemporain dans le monde. Variation et structure*. Toulouse: Presses Universitaires du Midi, coll. Amphi 7.
- Carranza, M., Cucchiari, C., Llisterri, J., Machuca, M. J., & Ríos, A. 2014. A corpus-based study of Spanish L2 mispronunciations by Japanese speakers. Edulearn14, 6th International Conference on Education and New Learning Technologies. July 7th-9th, 2014, Barcelona, Valencia: IATED Academy, 3696-3705.
- Detey, S., Lyche, C., Tchobanov, A., Durand, J. & Laks, B. 2009. Ressources phonologiques au service de la didactique de l'oral: le projet PFC-EF. Mélanges CRAPEL, 31, 223-236.
- Detey, S., Durand, J., Laks B. & C. Lyche (eds.) 2010. *Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignement*. Paris: Ophrys. www.projet-pfc.net (15 March, 2016).
- Durand, J, Carr, P., Pukli, M. (eds.) 2004. *English pronunciation: accents and variation*, Vol. 36 of La tribune internationale des langues vivantes. www.projet-pac.net (15 March, 2016).
- Durand, J., Gut, U. & Kristoffersen, G. (eds.) 2014. *The Oxford Handbook of Corpus Phonology*, Oxford: Oxford University Press.
- Durand, J. & Przewozny, A. 2015. La variation et le programme PAC. In I. Brulard, P. Carr, J. Durand, *La prononciation de l'anglais : variation et structure*, Toulouse: Presses Universitaires du Midi, 55-91.
- Durand, J. & Przewozny, A. 2012. La phonologie de l'anglais contemporain : Usages, variétés et structure, *Revue française de linguistique appliquée*, vol. XVII-1, Oral, norme et variation(s), 25-37.
- Ghadessy, M. G. Henry, A. & Roseberry, R. 2001. *Small corpus studies and ELT*. Amsterdam: John Benjamins.
- Granger, S., Hung, J. & Petch-Tyson, S. (eds.) 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins.
- Granger, S. 2004. Computer learner corpus research: current status and future prospects. In: Ulla Connor and Thomas Upton (eds.), *Applied corpus linguistics. A multidimensional perspective*, Amsterdam: Rodopi, 123-145.
- Herry-Bénil, N. 2010. *Evaluations objectives et subjectives de la prosodie de l'anglais parlé par des Français*. Paris: Editions Publibook.
- Herry-Bénil, N. 2011. *Didactique de la phonétique anglaise*. Rennes, Presses Universitaires de Rennes.
- Kettemann, B. & Marko, G. (eds.) 2002. *Teaching and learning by doing corpus*

- analysis*. Amsterdam: Rodopi.
- Lacoste, V., Herry-Bénil, N & Kamiyama, T. 2015. ICE-IPAC: Interphonology in English-learning contexts: new perspectives from EFL and ESL research. 11th International PAC conference, University of Toulouse Jean Jaurès, 9-13 April 2015.
- Przewozny, A. 2015. Research environments for pedagogical tools: learning and teaching with the PAC programme. Oral presentation, RANACLES 23rd Congress, University of Toulouse Jean Jaurès, 26 Nov. 2015.
- Sinclair, J. M. (ed.) 2004. *How to use corpora in language teaching*. Studies in Corpus Linguistics, 12, Philadelphia: John Benjamins.
- Trouvain, J. & Gut, U. (eds.) 2007. *Non-native prosody: phonetic description and teaching practice*. Berlin: Mouton de Gruyter.
-

Novice academic writing in Norwegian and English: a new corpus project

Sylvi Rørvik, Marte Monsen

Hedmark University of Applied Sciences

In this poster presentation we give an overview of plans for a corpus compilation project comprising student academic writing in L1 Norwegian and L2 English by students attending a master's program at a Norwegian University College. The corpus compilation was begun in 2015 and will continue until 2019. During this period, we will collect student texts in L1 Norwegian and L2 English. The texts will be directly comparable, since they will be written by students attending the same or similar courses in a master's program which allows for specialization in either English, Norwegian, or music, with the latter two groups producing texts in Norwegian. The rationale behind the project is a desire for evidence-based conclusions about the acquisition of academic language skills, in order to provide better courses in academic writing for the student group in question. In addition, the insights gained from the study of the material collected for this project will add to those from the relatively few existing studies of student academic writing in Norway that include a contrastive perspective (see e.g. Fossan 2011). A contrastive perspective including texts in the students' L1 is especially valuable when investigating learner language, since it allows the researcher to control for the potential influence of transfer. Existing research on published academic writing in Norwegian and English (see e.g. Fløttum et al 2006) can then serve as a yardstick against which to measure the students' texts, by representing the standards to which they ultimately aspire.

The poster will also present results from a pilot study of one type of metadiscourse in English and Norwegian novice academic writing, taking as its starting point the concept of "metatext" which can be defined as "the writer's explicit commentary on her own ongoing discourse" (Ädel 2006: 183-184).

References

- Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

- Fløttum, Kjersti, Trine Dahl & Torodd Kinn. 2006. *Academic voices: Across languages and disciplines*. Amsterdam/Philadelphia: John Benjamins.
- Fossum, Heidi. 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*. Oslo: University of Oslo MA thesis.
-

Investigating the use of cohesive devices by advanced learners of German through contrastive interlanguage analysis

Carola Strobl
Ghent University

Cohesion plays an important role in academic writing. Yet, it is a stumbling stone for advanced foreign language (FL) writers, as it requires attention to a global textual level (Kellogg, 2008; Kirkland & Saunders, 1991; Strobl, forthcoming). Despite its impact on academic writing quality, the adequate use of cohesive devices is an under-researched topic in advanced language instruction (Lee, 2002). To close this gap, first, the characteristics of learner language with regard to cohesion building need to be identified. To our knowledge, this has not been done in a comprehensive way for German FL, which is the scope of the presented project. To this aim, we depart from a Contrastive Interlanguage Analysis (CIA) approach. Corpus-based CIA allows to identify characteristics of non-native speakers' (NNS) writing in comparison with those of native speakers (NS) (Granger, 2015). The corpus under analysis is *Falko* (Reznicek, Lüdeling, & Schwantuschke, 2012) which comprises texts of two academic genres, summary and essay, written by NS and NNS. It therefore forms a good starting point for a CIA of cohesion in advanced learners' writing in German. *Falko* contains annotations on the levels of part-of-speech, lemmatisation, syntactic function, sentence structure, and target hypotheses in case of non target-like learner output. Up to date, it does not contain annotations specifically related to cohesion. The presented project, which is in its initial stage, follows a two-step approach to carry out a CIA based on the *Falko* data: First, the corpus in its current annotation status will be explored by ways of a pilot study, using the corpus search tool ANNIS (Zeldes, Lüdeling, Ritz, & Chiarcos, 2009). ANNIS allows for highly flexible queries combining information on different annotation levels. Following a top-down approach based on previous scholarship, some hypotheses with regard to characteristics of cohesion in learner language will be tested that do not require a specific annotation for cohesion, such as particular aspects of connector usage (Belz, 2005; Benazzo & Andorno, 2010; Bolton, Nelson, & Hung, 2003; Dimroth, Andorno, Benazzo, & Verhagen, 2010; Granger & Tyson, 1996) and syntactic coordination patterns (Vyatkina, 2013). For a second, more comprehensive, bottom-up analysis, a new annotation layer will be added to the *Falko* corpus in which lexical cohesion and co-reference are tagged semi-automatically. To this aim, we will follow the procedures and guidelines established by Lapshinova-Koltunski and Kunz (2014) for the annotation and analysis of cohesion in a multilingual corpus for German and English (<http://www.gecco.uni-saarland.de/>). The results of this in-depth analysis

will provide insights into the characteristics of advanced learner language concerning cohesion building in academic writing, and will therefore serve as a valuable information source for advanced writing instruction in German FL.

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.
- Benazzo, S., & Andorno, C. (2010). Discourse cohesion and Topic discontinuity in native and learner production: Changing topic entities on maintained predicates. *EUROSLA Yearbook*, 10(1), 92-118.
- Bolton, K., Nelson, G., & Hung, J. (2003). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics*, 7(2), 165-182. doi:10.1075/ijcl.7.2.02bo1
- Dimroth, C., Andorno, C., Benazzo, S., & Verhagen, J. (2010). Given claims about new topics. How Romance and Germanic speakers link changed and maintained information in narrative discourse. *Journal of Pragmatics*, 42(12), 3328-3344. doi: 10.1016/j.pragma.2010.05.009
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24. doi: 10.1075/ijlcr.1.1.01gra
- 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
- Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1(1), 1-26.
- Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing Student Performance in Summary Writing: Managing Cognitive Load. *TESOL Quarterly*, 25(1), 105-121.
- 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
- 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.
- Strobl, C. (forthcoming). Learning to write syntheses through collaboration and strategy instruction. *Journal of Second Language Writing*.
- Vyatkina, N. (2013). Specific Syntactic Complexity: Developmental Profiling of

Individuals Based on an Annotated Learner Corpus. *The Modern Language Journal*, 97(S1), 11-30. doi: 10.1111/j.1540-4781.2012.01421.x
Zeldes, A., Lüdeling, A., Ritz, J., & Chiarcos, C. (2009). *ANNIS: A search tool for multi-layer annotated corpora*. Retrieved from <http://edoc.hu-berlin.de/oa/conferences/reS4Xo05sncZc/PDF/29i8VTIzYfT3M.pdf>

Geo- and sociolinguistics of plural acquisition by German pre-schoolers with and without immigration background

Eugen Zaretsky^{1,2}, Benjamin P. Lange³

¹ University Hospital of Marburg, Germany

² University Hospital of Frankfurt/Main, Germany

³ Julius Maximilian's University of Wuerzburg, Germany

Introduction

The distribution of plural allomorphs and avoidance strategies in the varieties of pre-schoolers' language cannot avoid being influenced by geo- and sociolinguistic variables that are linked to the quality and quantity of the language input. This study aimed to examine associations between German pre-schoolers' error patterns and some characteristics of the children, their families, and the districts where they lived.

Methods

A learner corpus with results of language assessment studies including documented error patterns of 696 German pre-schoolers (281 monolingual Germans, 381 bi/multilingual children, 34 unknown; 384 boys, 312 girls; median age 51 months) in the grammar subtest of the validated language test SETK 3-5 (18 plural items) was analysed retrospectively in regard to the associations of the error patterns with 46 demographic/sociological characteristics of 45 Frankfurt/Main districts where the children lived as well as with sociolinguistic characteristics of children and their families, documented in questionnaires for parents and day-care centre teachers.

Results

A number of weak, but statistically significant correlations were identified between characteristics of children's plural forms (e.g., percentage of comparatively simple *-(e)n* overgeneralizations, use of double or not existing plural markers, use of avoidance strategies and zero plural, total number of correctly used plural allomorphs, consideration of gender-associated tendencies in the distribution of plural markers in the target language) and district characteristics linked to average income, the percentage of unemployed inhabitants and immigrants. The distribution of overgeneralized plural markers hardly varied in the error patterns of the largest immigrant groups (Italians, Russians, Turks, Arabs, Serbs/Croatians) and was much more linked to the classification of children as speaking (a) only German, (b) German and (an)other language(s), (c) only (an)other language(s). Only in case of Russian

speaking children, a possible influence of the mother tongue was found, reflected in the preference for vocalic plural markers in German (cf. vocalic Russian plurals *-a*, *-e*, *-u/vl*). Children producing comparatively advanced error patterns (e.g., using all available plural allomorphs) scored higher on a group of questionnaire items related to sociability (e.g., the willingness to play with other children).

Discussion

Error patterns of the largest subgroups of immigrant pre-schoolers in Frankfurt/Main were almost identical but differed significantly—quantitatively, not qualitatively—from those of monolingual Germans of the same age, although comparable to the error patterns of younger Germans. This lag in the linguistic development cannot be called a (multi)ethnolect but, rather, a result of the low-quality language input caused by the geographical and social separation from monolingual Germans. Immigrant children, especially Turks and Arabs, dwelled in the districts with a high percentage of the low-income and/or unemployed inhabitants speaking two or more languages at home. Comparatively advanced error patterns were associated with the children's sociolinguistic characteristics which can be summarized as openness or sociability. Children who were geographically, socially, and „psychologically“ isolated from the monolingual Germans as well as children from families with very few ties to the German mainstream society demonstrated the simplest plural error patterns.

Author Index

- Čermáková
Anna, 20
- Abe
Mariko, 7
- Akasegawa
Shiro, 22
- Alruwaili
Awatif, 9
- Alves
Jessica Ceritello, 46
- Andresen
Melanie, 10, 89
- Anthony
Laurence, 12
- Baker
Paul, 12
- Banat
Hadi, 78
- Berger
Cynthia M., 65
- Biel
Albert, 13
- Boulton
Alex, 15, 23
- Bredenbröcker
Martina, 90
- Brezina
Vaclav, 14, 37, 92, 93
- Brzoza
Bartosz, 53
- Callies
Marcus, 1
- Charles
Maggie, 17, 47
- Chen
Alvin Cheng-Hsien, 18
Meilin, 94
- Chlumská
Lucie, 20
- Chujo
Kiyomi, 22
- Cobb
Tom, 15, 23
- Díez-Bedmar
María Belén, 59
- D'Arienzo
Meredith, 25
- De Metsenaere
Hinde, 26
- Elaine
Boyd, 92
- Ellis
Nick C., 65
- Fastrich
Bridgit, 28
- Fenogenova
Alena, 30
- Ferreira
José Pedro, 103
- Flowerdew
John, 94
Lynne, 34
- Frankenberg-Garcia
Ana, 36
- Gärtner
Tobias, 42, 102
- Gablasova
Dana, 14, 37, 92
- Garner

James Robert, 39, 41
 Gilquin
 Gaëtanelle, 2
 Golcher
 Felix, 81
 Gráf
 Tomáš, 43
 Griebel
 Reiner, 99
 Guedes
 Annallena de Souza, 46
 Hübner
 Julia, 96
 Hadley
 Gregory, 47
 Herry-Bénil
 Nadine, 109
 Hirschmann
 Hagen, 81
 Hu
 Hsueh-chao Marcella, 97
 Huang
 Ping-Yu, 48
 Jeaco
 Stephen, 50
 Kakoyianni-Doa
 Fryni, 52
 Kaul
 Thomas, 99
 Kemp
 Jenny, 101
 Keng
 Nicole, 74
 Kliner
 Leonid, 99
 Knorr
 Dagmar, 89
 Kogan
 Marina, 102
 Krumpen
 Monika, 99
 Kuzmenko
 Elizaveta, 30
 Lüdeling
 Anke, 4
 Lange
 Benjamin P., 115
 Leńko-Szymańska
 Agnieszka, 104
 Lewandowska-Tomaszczyk
 Barbara, 4
 Lopes Alves
 Anna Luisa, 108
 Manova
 Stela, 53
 Marcos Miguel
 Nausica, 55
 Marinov
 Sanja, 56
 McEnery
 Tony, 6, 92
 Minagawa
 Yasuyo, 87
 Monsen
 Marte, 112
 Nahm
 Jillian, 57
 Nihal
 Çalışkan, 106
 Oghigian
 Kathryn, 22
 Orfanò
 Bárbara Malveira, 46
 Oviedo
 Alejandro, 99
 Pérez-Paredes
 Pascual, 59
 Pezik
 Piotr, 60
 Pinto
 Paula Tavares, 108
 Przewozny-Desriaux
 Anne, 109
 Puga
 Karin, 62
 Römer
 Ute, 65

Rørvik	Nai-Lung, 48
Sylvi, 112	Tyne
Rica Peromingo	Henry, 80
Juan Pedro, 64	Tziafa
Rosen	Eleni, 52
Anna, 67	Vinogradova
Sáenz Herrero	Olga, 30
Ángela, 64	Vyatkina
Satake	Nina, 81
Yoshiho, 69	Whiteside
Schaeffer-Lacroix	Karin, 83
Eva, 70	Wiemeyer
Schroth-Wiechert	Leonie, 85
Sigrun, 102	Wolk
Schutzzeichel	Christoph, 28
Marc, 96	Wu
Sing	Shu-jung, 97
Christine S., 72	Xu
Smith	Hai, 93
Simon, 74	Yan
Staples	Xun, 76
Shelley, 76, 78	Yatabe
Strobl	Kiyomi, 87
Carola, 113	Zaretsky
Swatek	Eugen, 115
Aleksandra, 78	Zingano Kuhn
Tsai	Tanara, 103
Kuei-Ju, 79	
Tsao	