Plenaries ......................................................................................................................... 6

Susan Hunston .................................................................................................................. 6
Anne O’Keeffe .................................................................................................................... 7
Alex Boulton ....................................................................................................................... 8
Costas Gabrielatos ........................................................................................................... 9

Conference papers .......................................................................................................... 11

S1 Chris Turner and Hilary Nesi ...................................................................................... 11
S4 Jane Seely ..................................................................................................................... 12
S5 Valentin Werner, Robert Fuchs and Sandra Götz ....................................................... 13
S8 Gregory Hadley and Hiromi Hadley ............................................................................ 15
S10 Ana Frankenberg-Garcia .......................................................................................... 17
S11 Ana Frankenberg-Garcia, Robert Lew, Adam Mickiewicz, Geraint Paul Rees, Jonathan C. Roberts, Nirwan Sharma ................................................................................... 18
S18 Hosam Darwish .......................................................................................................... 19
S21 Yoshiho Satake .......................................................................................................... 20
S23 Elen Le Foll ................................................................................................................ 22
S26 Suresh Jampa ............................................................................................................. 24
S29 Szu-Yu Liu and Hsien-Chin Liou ............................................................................. 25
S30 Tzu Wei Yang and Hsien Chin Liou ......................................................................... 26
S31 Marie-Louise Brunner and Stefan Diemer ................................................................ 28
S33 Cristóbal Lozano, Amaya Mendikoetxea and Paul Rollinson ..................................... 30
S36 Henry Tyne ................................................................................................................ 33
S38 Maggie Charles ......................................................................................................... 34
S46 Nina Vyatkina ............................................................................................................ 36
S53 Ayman Alghamdi & Eric Atwell .............................................................................. 38
S54 Tanjun Liu ................................................................................................................ 41
S58 Huifen Lin and Pinshuan Lee ................................................................................... 43
S60 Claire Wolfarth, Claude Ponton, Catherine Brissaud .............................................. 45
S62 Claudia Wunderlich .................................................................................................. 46
S71 John O’Donoghue .................................................................................................... 47
S73 Luciana Forti ............................................................................................................. 50
S75 Sanja Marinov ........................................................................................................... 51
S78 Meilin Chen and John Flowerdew .......................................................................... 53
S80 Martin Weisser ......................................................................................................... 55
S82 Hege Larsson Aas and Sylvi Rørvik ......................................................................... 57
S85 Shelley Staples .......................................................................................................... 59
S173 Agnieszka Leńko-Szymanańska ................................................................. 61
S195 Geraint Paul Rees ..................................................................................... 62
S197 John Williams ......................................................................................... 64
S110 Alvin Cheng-Hsien Chen ........................................................................ 65
S107 Geraldine Mark and Pascual Pérez-Paredes ........................................... 67
S109 Rolf Kreyer ............................................................................................. 67
S110 Mariko Abe, Yusuke Kondo, Yuichiro Kobayashi, Akira Murakami and Yasuhiro Fujiiwara . 70
S111 Lynne Flowerdew .................................................................................... 71
S113 Awaatf Alruwaili ...................................................................................... 73
S114 Benet Vincent, Hilary Nesi and Daniel Quinn .......................................... 75
S115 Katarina Lazic and Maja Milicevic Petrovic ............................................ 76
S116 Reka R. Jablonkai and Neva Cebron ...................................................... 77
S120 Kiyomi Chujo, Atsushi Mizumoto, Yuichiro Kobayashi, Akira Hamada, Kathryn Oghigian .. 79
S127 Yukio Tono ............................................................................................... 81
S129 Hildegunn Dirdal and Stephanie H. G. Wold ........................................... 82
S130 Diane Pecorari ......................................................................................... 84
S131 Katerina Florou ....................................................................................... 85
S134 Francesca Perri ....................................................................................... 86
S137 Lan-Fen Huang, Tomáš Gráň and Nicole Keng ......................................... 87
S138 Katherine Ackerley ................................................................................. 89
S142 Hyeson Park ............................................................................................ 90
S149 Yi-Ju Ariel Wu ......................................................................................... 92
S150 Christoph Wolk and Bridgit Fastrich ....................................................... 94
S152 Zhaozhe Wang ....................................................................................... 96
S157 Christine Sing ......................................................................................... 97
S158 Tomáš Gráň and Lan-Fen Huang ............................................................ 99
S159 Patricia Tosqui-Lucks and Malila Carvalho De Almeida Prado ............. 101
S161 Maria Kunicovskaia and Natalya Morgoun ........................................... 103
S162 Dana Gablasova and Vaclav Brezina ..................................................... 105
S163 Kyoko Sugisaki and Michael Prinz ........................................................ 106
S164 Ji-Young Shin ......................................................................................... 108
S166 Francesca Seracini .................................................................................. 110
S170 Paul Wickens ......................................................................................... 111
S173 Łucja Biel .............................................................................................. 113
Conference 7/14 presentations ................................................................. 115
  S22  Rudy Loock.................................................................................. 115
  S27  Eva Schaeffer-Lacroix ................................................................. 117
  S68  Michael Pace-Sigge .................................................................... 119
  S112 Lynne Flowerdew ...................................................................... 120
  S139 Sylwia Twardo .......................................................................... 121
  S144 Yuying Hu .................................................................................. 122
  S145 Sylvain Perraud .......................................................................... 124
Conference posters ................................................................................ 126
  S3  Danyang Zhang............................................................................. 126
  S66  Lucie Chlumská .......................................................................... 128
  S67  Stefania Spina and Anna Siyanova-Chanturia ............................. 130
  S87  Adriana Picoral ........................................................................... 131
  S92  Lin Jiang ..................................................................................... 133
  S93  Barry Kavanagh ........................................................................... 134
  S106 Miki Hyun Kyung Bong and Masako Tsuzuki ............................. 135
  S126 Rezan Alharbi ............................................................................ 136
  S132 Klara Klimcikova, Vishal Bhalla and Aisulu Rakhmetullina ..... 137
  S141 Shona Whyte ............................................................................... 139
  S143 Dong Ok Lim ............................................................................... 141
  S153 Xiao Wang .................................................................................. 142
  S155 Duy Van Vu ................................................................................ 143
  S159 Yolanda Noguera ...................................................................... 144
  S165 Francesca Poli .......................................................................... 145
  S171 Xin Xu ........................................................................................ 147
  S172 Aleksandra Swatek .................................................................... 148

Workshops ............................................................................................... 149
  W1  Niall Curry and Olivia Goodman .................................................. 149
  W2  Olga Vinogradova, Stefania Spina, Luciana Forti, Ivan Torubarov, Nikita Login .............. 149
  W3  Silvia Molina, Plaza María del Mar Robisco, Martín Verónica Vivancos Cervero, Ana Roldán-Riejos, Paloma Úbeda Mansilla ................................................................. 151
  W4  Laurence Anthony ......................................................................... 152
  W5  Adriana Picoral, Shelley Staples, Ji-young Shin, Aleksandra Swatek .......................... 153
  W6  Vaclav Brezina, Dana Gablasova, Irene Marín Cervantes ....................... 154

Author index ............................................................................................ 156
Corpus Linguistics has a tradition of research into the patterning of lexis, inspired by Sinclair (1991) and with examples including Hoey’s work on Lexical Priming, Hanks’ work on Corpus Pattern Analysis, and Francis et al.’s work on Pattern Grammar. The challenge in making this work relevant to learners is to strike a balance between specificity – each word has unique characteristics – and generalizability; that is, between lexis and grammar. Meanwhile, Construction Grammar proposes a model of language that is driven by both lexis and pattern. Constructions exist at all levels of generality, some of them overlapping with the units identified by corpus research, a coincidence exploited by, for example, Ellis, Römer and O'Donnell (2016) in their work on language development. This paper will propose an alignment between the groups of words identified in the Pattern Grammar research on the one hand, and constructions on the other hand, leading to the potential identification of one thousand constructions that are relevant to learners. Some examples of these constructions (not one thousand!) and how they can be identified are given, and the relevance of this to reference materials for learners is discussed.

**Susan Hunston** is Professor of English Language at the University of Birmingham, where she has worked for the last twenty years. She previously held posts at the National University of Singapore and the University of Surrey and before that taught English as a Second Language and English for Academic Purposes. She carries out research in corpus linguistics and discourse analysis, in particular investigating academic discourse and the language of evaluation. She is the author of Corpora in Applied Linguistics (CUP, 2002) and Corpus Approaches to Evaluation (Routledge, 2011) and numerous articles.
Anne O’Keeffe

*Throwing the pigeon among the cats – Data-driven learning and the Second Language Acquisition interface debate*

Over the years, debate prevails in Second Language Acquisition studies as to whether there is any interface between explicit consciously learnt knowledge and implicit acquired tacit knowledge in language learning. Three overarching stances or ‘positions’ are taken: 1) No-interface position: those who hold that implicit, automatized acquisition only takes place at a sub-conscious level via comprehensible input and that there is no connection with explicit learning; 2) the Weak-interface position: those who see a degree of connection between implicit and explicit knowledge, under certain conditions; and 3) Strong-interface position: those who believe that explicit knowledge can become automatized and implicit.

On the surface at least, Data-Driven Learning (DDL) could be categorized as an explicit approach to learning, which within at least one of the aforementioned positions could be rendered futile. Alternatively, within a strong interface position, where explicit knowledge is seen to lead, at some stage, to automatization whereby the learnt forms can become part of the users’ long-term memory and fluent sub-conscious functionality, DDL has a strong case.

This keynote talk seeks to conceptually explore DDL in relation to this interface debate in SLA. It will examine practices within DDL in terms of what they demand cognitively of the learner and consider whether an awareness of the interface debate might enlighten how best to structure DDL processes, pedagogies and practices so as to augment the likelihood of interface between explicit and implicit knowledge.

Anne O’Keeffe is a Senior Lecturer in Applied Linguistics at Mary Immaculate College, University of Limerick. Her research output includes papers, chapters and books on Corpus Linguistics, Pragmatics and Media Discourse. These include Investigating Media Discourse (2006, Routledge), From Corpus to Classroom (2007, Cambridge University Press, with Michael McCarthy and Ronald Carter), English Grammar Today (2011, Cambridge University Press, with Ronald Carter, Michael McCarthy and Geraldine Mark), Introducing Pragmatics in Use (2011, Routledge, with Brian Clancy and Svenja Adolphs). She also co-edited the Routledge Handbook of Corpus Linguistics (with Michael McCarthy). Her most recent research on the Cambridge Learner Corpus has led to the online resource, the English Grammar Profile (with Geraldine Mark). She has also guest edited a number of international journals, most recently Corpus Pragmatics and she is co-editor of two Routledge book series, Routledge Corpus Linguistic Guides and Routledge Applied Corpus Series.
Corpus tools and techniques have been used for pedagogical purpose for around 50 years (McEnery & Wilson, 1997). The first academic publications in the area appeared in the 1980s (e.g. McKay, 1980), though the concept is largely associated with work by Tim Johns who published a string of papers on the topic before and after he coined the term ‘data-driven learning’ (DDL) in 1990. His work coincided with a new generation of corpus building and analysis, not least the COBUILD project in Birmingham where Johns was based (e.g. Sinclair, 1991). The TaLC (Teaching and Language Corpora) conferences were inaugurated at Lancaster University in 1994 and have been conducted every two years since then, each conference accompanied by the publication of selected papers. There have now been several hundred papers, books, chapters, proceedings papers and PhD theses in this area. Given that, it is legitimate to wonder how to go about making sense of results in the field, and indeed what DDL actually looks like today. These are the two issues addressed in this presentation.

For the first, Chambers provided the first genuine attempt to summarise some of the empirical research in 2007; most such narrative syntheses since then are open to charges of presenting partial coverage of the field and subjective interpretation of the evidence. Systematic trawls find that the body of empirical research in DDL now amounts to more than 300 individual publications; characteristics can be derived from coding them in categories such as publication type and date, design and analysis, countries and contexts, learner levels and needs, tools and corpora, target language and uses, and so on. The abstracts can also be analysed using corpus tools for further insights. A second type of synthesis is the meta-analysis such as that conducted by Boulton and Cobb (2017). While again only providing partial coverage (i.e. quantitative results only), this is at least systematic in collecting research corresponding to stated inclusion criteria, and in the analysis itself. The results briefly sketched out are highly encouraging – but how then are we to justify the “seeming mismatch between utility and uptake” (Ballance, 2007)? We therefore focus in on less successful instantiations of DDL in an attempt to see if they have anything in common, with possible suggestions for future good practice.

For the second, one crucial development is clearly the existence of many fast, efficient, user-friendly corpora and tools, some designed specifically with language learners in mind, as well as the ease of creating one’s own resources. Further, users today are far more familiar with computer tools, and ‘digital natives’ regularly use search engines for language queries on the web via computers or mobile devices. Nonetheless, a look at examples given in the literature from the early days as compared to recent publications shows that many DDL practices have remained broadly similar over the decades. The question then is: has DDL reached a stable state, or is there room for substantial evolution in the
future? This is highly personal question (see e.g. Tribble, 2015) and a genuinely open one; audience suggestions are warmly welcomed.

**Alex Boulton** is Professor of English and Applied Linguistics at the University of Lorraine and director of the ATILF (UMR 7118: CNRS & UL). Particular research interests focus on corpus linguistics and potential uses for ‘ordinary’ teachers and learners (data-driven learning). He has published and edited books and papers in these fields over the years, and is on various boards and committees: AFLA (vice-president), EUROCALL and TaLC; as well as journals such as ReCALL (editor), Alsic, ASp, CALL-EJ, Eurocall Review, IJCALLT, JALT-CALL Journal, Language Learning & Technology, and Al-Lisaniyyat.

---

**Costas Gabrielatos**

*Pedagogy-driven corpus-based lexicogrammar*

The talk will discuss an approach situated at the intersection of linguistic theory, analysis of learner language, frequency studies, and evaluation of pedagogical materials, with the aim of developing a body of corpus-based lexicogrammatical information for language learners, particularly on issues that are (deemed to be) problematic for language learners. That is, the approach is intended to complement research on DDL. The motivation for this approach is three-fold.

Currently, pedagogy-oriented corpus-based studies have one of two main foci. One research strand examines learner language, usually incorporating an explicit or implicit comparison to L1 use (e.g. errors or frequency of use), or seeking to establish correlations between L2 use and the learners’ L1. A second strand seeks to evaluate pedagogical materials (grammars, dictionaries, and coursebooks) in terms of the information, examples, and exercises they contain by comparing this information to L1 use. However, few studies so far have combined the two strands.

In linguistics, there are different views on the nature of lexicogrammar, which (explicitly or implicitly) tend to ascribe primacy to either lexis or grammar. In turn, particular theoretical frameworks, especially those popular among language teachers and materials writers, influence the content of pedagogical materials. The proposed approach is influenced by Halliday’s view of lexis and grammar as “complementary perspectives” (1991: 32), and his conception of the two as notional ends of a continuum (lexicogrammar), in that “if you interrogate the system grammatically you will get grammar-like answers and if you interrogate it lexically you get lexis-like answers” (1992: 64). More specifically, the proposed approach interrogates the system lexicogrammatically to get lexicogrammatical answers.
Language teaching still treats lexis and grammar predominantly in a compartmentalised fashion, evidenced by the existence of pedagogical grammars and learner dictionaries. Although pedagogical materials overlap in their coverage (grammars also provide some lexis-like information, and dictionaries also provide some grammar-like information), this is not done consistently, and the two aspects are not presented as being interconnected (let alone inseparable). Similarly, coursebook units tend to provide separate sections for grammatical and lexical information, the latter complemented with phraseological information -- usually in the form of (semi-)fixed expressions. As a result, learners need to consult both a grammar and a dictionary (or different sections of a coursebook) in order to construct a full picture of a lexicogrammatical item. The grammar-lexis division in current pedagogical materials can be seen as unavoidable in light of the size limitations of hard-copy volumes (and the attendant cost). However, these issues do not apply to electronic publishing, which offers possibilities for the production of comprehensive learner resources that combine the features of pedagogical grammars and dictionaries: pedagogical lexicogrammars. Such resources would also be easily updatable and expandable, and grammar/lexis-like information could be interlinked. In addition, entries could provide links to open-access corpora, enabling learners to examine instances of actual use -- a practice that would also allow for “serendipity” (Bernardini, 2000), that is opportunities for learners to discover language features other than the ones for which they accessed the pedagogical lexicogrammar. Using such a resource, learners would be able to access language information, and examples of actual use, starting at any point of the lexicogrammar continuum, and combine information as needed.

Costas Gabrielatos has been working in corpus linguistics since 2001. He came to linguistics via English language teaching (1984-1993) and language teacher education (1992-2001). Costas joined Edge Hill University in September 2012; previously, he worked as associate lecturer and researcher at Lancaster University and Liverpool University. His general research interests are in the development of corpus approaches to issues in theoretical and applied linguistics. More specifically, his work combines the following areas:

– Corpus Linguistics: compilation of topic-specific corpora, annotation techniques, metrics.
– Lexicogrammar: conditionals, modality, tense-aspect, construction grammar, lexical grammar.
– Corpora in language education: pedagogical lexicogrammar, analysis of learner language.
– Discourse-oriented corpus studies.
Conference papers

S1  Chris Turner and Hilary Nesi

Corpus research into Some and Any and its implications for pedagogical grammar description and teacher education

This talk will draw on data from both native-speaker and learner corpora in order to present a new teaching approach to some and any, with implications for pedagogical grammar book entries and teacher education courses.

It will begin with a short overview of the standard descriptions of some and any provided in popular and highly regarded pedagogical grammar books (e.g. Swan, 2016, Carter and McCarthy, 2006, Murphy, 2012, Parrott, 2010, Biber, Conrad and Leech, 2002), and then provide examples from the Oxford English Corpus that reveal major gaps and inaccuracies in these descriptions. Pedagogical grammars tend to treat clause type rather than lexical meaning as the main variable that determines the choice between some and any. They also largely neglect the uses of some in negative sentences, and oversimplify the pragmatic factors that determine the choice between some or any in conditionals, questions and several other clause types.

The talk will go on to discuss the possible effects of inaccurate and oversimplified grammar book descriptions, drawing on data from the Cambridge Learner Corpus which suggest that some learner errors with some and any may be due to discrepancies between what both the grammar books and the teachers say, and what learners encounter when reading or listening to authentic English usage. We will also draw attention to some problematic error mark-up relating to some and any in the Cambridge Learner Corpus, suggesting that grammar book descriptions may have clouded markers’ judgement of correct and incorrect usage, and discuss the implications of this finding for teacher education.

Departing from our analyses of the two corpora, we will propose new pedagogical grammar descriptions and a new classroom teaching approach for some and any, suitable for learners at different levels of proficiency. Both proposals are based primarily on a lexical phrase approach to grammar (Boers & Lindstromberg, 2009; Lewis, 1993, 1997) and on Leech’s prototype approach to language description (Leech, 1994). A long-term teaching approach should first provide elementary-level learners with a solid grasp of the main meanings and uses of some and any, and then gradually help higher-level learners towards a deeper, more nuanced understanding of the semantico-pragmatic principles which determine the choice between the two items. Throughout this teaching approach, learners should be introduced to key lexical phrases with some and any that both provide support for the semantico-pragmatic rules that are provided and offer an alternative, non-analytical means of acquiring a number of key distinctions between the two words.
We will also propose some data-driven activities for teacher training, to raise awareness of the inaccuracies and limitations of the grammar book rules. The talk will conclude that, while some descriptive simplification may be necessary for both lower-level learners and teachers with little prior knowledge of English grammar, there is no place at any point in the learning process for fundamentally inaccurate rules, such as those that current grammar books provide for *some* and *any*.

References

S4  Jane Seely

*A corpus-aided comparative analysis of native and non-native speaking English language teachers’ approaches to in-class spoken feedback*

Classroom Discourse (CD) and Teacher Talk (TT) have received much attention over the years across a range of research perspectives. Much of recent CD work draws on Conversation Analysis (CA) and, more recently, there has been growing synergy between CA and Corpus Linguistics (CL). In the area of Language Teacher Education (e.g. CELTA), trainee teachers are introduced to the concept of TT as something to be minimised and very little mention is made of the types of TT used in the classroom. In other words, the focus is on reducing the quantity of TT rather than on the nature of the talk itself.

This paper will report on a quasi-longitudinal mixed-methods study of TT which uses data from classroom interactions over 9 months and also draws on interviews with the teachers. This 150,000 word corpus comprises 15 Native (NS) and Non-Native English (NNS) speaking teachers, at varying career stages, ranging in age from 23 to 36. Specifically, this paper will compare and discuss findings
on teachers’ spoken feedback in the classroom in relation to the Native English-Speaking teachers and their Non-Native English Speaking counterparts. The results indicate that, regardless of career stage, the NNS teachers are more direct in their approach to corrective feedback, with more examples of direct repair and form-focused feedback than the NS teachers. Additionally, while the majority of teachers interviewed stated that they were robust in their positive feedback, the corpus data showed a disparity between NS and NNS feedback whereby the NNS teachers seemed to show more positive reinforcement than their NS counterparts. This brings to light the usefulness of using corpus data in teacher training, in a mixed methods format, not least of all to highlight where there is disparity between teacher beliefs about TT and how they actually use it in the classroom context.

S5 Valentin Werner, Robert Fuchs and Sandra Götz

Cross-linguistic effects or universal learning mechanisms? A case study on temporal expression

While researchers from the domains of learner corpus and second language acquisition research widely agree that the linguistic expression of temporal relations represents a central and highly complex area, thus deserving persistent interest (see, e.g., contributions in Ayoun 2015, McManus et al. 2017, Fuchs & Werner 2018), opinions diverge as to the impact of cross-linguistic effects (e.g. Leńko-Szymańska 2007; Shirai 2009) vs. universal learning mechanisms (e.g. Klein 1995; Bardovi-Harlig 2000) in this domain. We aim to contribute to the wider discussion by exploring the alternation between the Present Perfect (PP) and the Simple Past (SP) in EFL learner data (cf. Fuchs et al. 2016), reflecting typologically different first-language backgrounds. More specifically, we tackle the following research questions:

1) The PP (and its alternation with the SP) as a challenging structure: Which rates of uptake do learners show? How and where (i.e. for which variables) do they differ from native speaker usage?

2) SLA principles: Is the acquisition of temporal expression and the difference to native speaker usage influenced by (i) cross-linguistic effects, (ii) guided by universal principles (irrespective of the learners’ L1), or rather by an interaction of (i) and (ii)?)

3) Linguistic principles: Can the use of PP/SP be predicted by linguistic variables, such as priming effects or lexical aspect of the verb?

To this end, we will present the findings of a Contrastive Interlanguage Analysis (Granger 1996, 2015), applying a modified version of regression-based multifactorial prediction and deviation analysis (MuPDAR; Gries & Deshors 2014) to a sample of c. 25,000 time-reference forms produced by advanced learners of English. Our database are the German and Chinese components of LINDSEI (Gilquin et al. 2010) and ICLE (Granger et al. 2009) as well as native-speaker control material from
LOCNEC (De Cock 2004) and LOCNESS, thus also allowing us to assess differences between the spoken and written mode.

Through this approach we are able (i) to disentangle cross-linguistic effects from general learner patterns and (ii) to identify the specific factor(s) where the cross-linguistic effects are salient. A fine-grained picture emerges showing that learners approximate to native choices in some contexts (e.g. SP choice with certain verbal semantics in conjunction with definite time adverbials), but not in others (e.g. SP choice with indefinite time adverbials). Overall, it further emerges that, while error rates differ between the two typologically different learner groups, actual linguistic conditioning (i.e. the influence of individual linguistic variables on the learners’ choices) does not. This suggests that, when factors playing a part in native speaker data are taken as a baseline, cross-linguistic influence occurs in the larger picture (i.e. error rates) rather than in the details (i.e. linguistic conditioning of these errors). We will discuss our results with a special focus on their implications for language teaching, e.g. how focusing on particular linguistic factors (such as temporal adverbials) in teaching may prompt learners to make a native-like choice, while other factors (such as priming) need to be explored further.

References


Data-Driven Learning (DDL), developed in the 1990s by Johns (Johns, 1991), has been shown to be effective among graduate-level and upper intermediate learners (Braun, 2007; Charles, 2012; Granger, Hung, & Petch-Tyson, 2002; Sun & Wang, 2003). However, its impact among lower level undergraduate learners suggests that marginal success comes after only significant amounts of scaffolding and teacher effort (Boulton, 2009; Hadley, 2002; St. John, 2001). The challenges to using DDL with undergraduates relate to the linguistic difficulty of currently available corpora and the nature of what we term as a Classical approach to DDL, which in today's mediated world, may be aesthetically displeasing to undergraduates, and which requires analytical skills more appropriate for graduate-level learners (Hadley & Charles, 2017).

This study investigates the use of a flexible approach to DDL that seeks to stimulate greater receptive lexicogrammatical knowledge and faster reading speeds among undergraduate learners in a Japanese university extensive reading program. Called Contemporary DDL, the underlying principles of Johns (1991), Bolton (2009), Charles (2012) are synthesized and artfully packaged to undergraduates in order to avoid the sometimes boxy and data saturated appearance of the Classical approach. Examples of Contemporary DDL materials, together with its relative strengths and weaknesses when compared to a Classical approach, will be discussed before a presentation of the following investigation.
From April 2015 to July 2017, students from six extensive reading classes participated this study. For 16 weekly 90-minute sessions, an experimental group (21 students) used DDL materials created from a corpus developed from the Oxford Bookworms Graded Readers, which contained 186 books from all seven levels with a total of 1,715,160 tokens (17,670 word types). The control group (28 students) had no DDL input. All students in this study read a minimum of 200,000 words during the course.

Quantitative data from a C-test (Klein-Braley & Raatz, 1984) constructed from an upper-level Bookworms reader. A speed reading test by Quinn, Nation, & Millett (2007) was also selected. A pre-test post-test design was used, and dependent as well as independent samples t tests were used. Pre-test analysis found that the experimental group was statistically distinct from the control group in terms of having lower levels of second language proficiency. Post-test scores found that both groups improved significantly, with high impact factors. However, the experimental group improved more by entering the same statistical bands as the control group. Post-test findings also indicate that students using the DDL materials were reading more books and were reading faster than the control group.

We conclude that a Contemporary approach to DDL, which seeks to supplement, enhance, and guide students towards a deeper level of learning, can have a positive effect with lower level undergraduate learners. While there is a trade off when deviating from a Classical approach, with its sharper focus on Key Words in Context, we find that an informed application of Contemporary DDL is nevertheless more suitable to the needs of undergraduates, and one that contributes to improving receptive learning and lexicogrammatical proficiency better than extensive reading alone.

References


The collocation repertoire of EAP users

While corpora enable researchers to observe the conventional usage of collocations among a given community of users, learner corpora have made it possible to study how less proficient language users may sometimes deviate from such conventions. Studies such as Nesselhauf (2005), Durrant and Schmitt (2009), Laufer and Waldman (2011), Lu (2017) and Paquot (2017), among others, have drawn attention not only to common miscollocations in learner writing, but also to the fact that less proficient writers tend to underuse certain collocations and overuse others. However, corpora only provide information about the collocation choices that come to surface in finished texts. They do not normally capture the collocation options available to writers during the writing process. Proficient writers should have little difficulty in retrieving from their mental lexicon the collocations they need without disrupting the flow of their words. However, less experienced writers may struggle to find appropriate collocations to convey their meanings, or may have a more limited collocation repertoire to choose from.

The present study investigates the collocations users of academic English at a British university are able to recall. Using authentic academic writing frames, it compares the collocations available to academics, EAP tutors and students at PhD, MA and undergraduate levels, and examines also whether L1-English and Other-L1 writers differ in their performances. A controlled experiment where ninety participants were asked to fill in a set of ten gapped sentences that emulate the wordings academic writers are likely to use routinely in their work was used to elicit the data. More than one collocation could be used in each gap, and the participants were asked to fill them in with as many collocations as they could effortlessly recall. The lexical items supplied by the participants were then classified according to whether or not they constituted valid academic collocations, using the Pearson International Corpus of Academic English (PICAE) (Ackerman et al. 2010) as a benchmark.
The results indicate that experience of written academic discourse plays a more decisive role than being having English as L1 in the collocations effortlessly available to EAP users. It is concluded that corpus-based tools and resources designed to jog writers’ memories in relation to the use of suitable academic collocations should recognize the needs of less experienced users of academic English in general, rather than just those of writers whose L1 is not English.

References


Developing a tool to help EAP writers with collocations in real time

Corpora have given rise to a wide range of lexicographic resources aimed at helping novice users of academic English with their writing, from core academic vocabulary lists (Coxhead 2000; Paquot 2010; Simpson-Vlach and Ellis 2010; Ackermann and Chen 2013; Gardner and Davies 2014), to textbooks (e.g. Schmitt and Schmitt 2005; McCarthy and O’Dell 2008), and even a dedicated academic English learners’ dictionary (Lea et al. 2014). However valuable these resources may be, novice EAP writers may not be aware of them or may not be sufficiently aware of the lexical shortcomings of their emerging texts so as to trigger the need to use such resources in the first place (Laufer 2011, Frankenberg-Garcia 2017). Moreover, even if EAP users did wish to look up a word
while writing an essay, dissertation, research article or similar, doing so could interrupt their thoughts and distract them from getting their ideas down on paper (Lew et al. 2018).

The ColloCaid project aims to address this problem by developing a tool to help EAP writers with academic collocations in real time. To achieve this, we have begun working on the compilation of a lexicographic database that supports novice EAP users’ collocation needs (Frankenberg-Garcia et al. 2019). Parallel to this, we are developing a tool that will enable writers to instantaneously visualise collocations from within a text editor (Frankenberg-Garcia et al. 2019; Roberts et al. 2017).

In this paper, we outline the rationale underlying our lexicographic coverage and preliminary visualisation decisions. More specifically, we describe the criteria used for (1) choosing which collocation bases (nodes) to cover, (2) selecting collocates and examples from expert academic English corpora, and (3) integrating information on collocations with text editors.

References


---

Hosam Darwish

*Writer-Reader Interaction: Writer’s Stance in English L1 & L2*

Students are frequently assessed in terms of the requirements of English academic norms to which they do not belong and of which they may have little experience; one of these academic features is expressing an appropriate stance. Stance refers to the ways academics annotate their texts to comment on the possible accuracy or creditability of a claim, the extent they want to commit themselves to it, or the attitude they want to convey to an entity, a proposition or the reader. According to Hyland (2005), stance features convey three broad meanings: 1. evidentiality, e.g. hedges (may) and boosters (clearly), 2. affect e.g. attitude markers (interesting) and 3. presence, e.g. self-mention (I). Writers use these devices to convey their judgements, opinions, commitments and their relationship and interaction with their readers.

Adopting Hyland’s (2005b) Model of Interaction, corpora of 80 discussion chapters written by both MA postgraduate Egyptian students (English L2) at Egyptian universities and their peers of British
students (English L1) at UK universities, were searched both electronically using Text Inspector tool (Bax, 2013) and manually to identify more than 200 stance markers in students’ academic scripts. Moreover, the study explored the perceptions of 20 of the text writers’ (both Egyptian and British) about the functions of certain stance markers and the factors that could affect their understanding and use of these academic features. Characteristics of successful stance-taking were identified after interviewing four expert writers.

The quantitative results found no statistically significant differences in the total number of stance markers, boosters and self-mentions used by writers at the two writer groups, but the L1 corpus contained statistically significant more hedges and attitude markers than the L2 one. Furthermore, the L1 texts contained noticeably more types of stance markers than the L2 scripts. The conducted discourse-based interviews indicated that both L1 and L2 writers were aware of the functions of stance markers. However, some of the interviewees (both L1 and L2) had narrow or even faulty conceptions of certain stance markers, e.g. possibility vs probability devices and, other attitude markers, e.g. important and significant. These features of academic discourse were not made more conspicuous to them, and this could have affected their employment of these linguistic features. The findings revealed that in addition to the lingua-cultural aspect, writer’s personal linguistic preferences, supervisor’s and other lecturers’ feedback, previous education and instruction and writer’s self-confidence were key factors that have played a considerable role on students’ lexical decision-making. For instance, the L1 students’ underuse of types of stance markers may be attributed to their lack of confidence and their reluctance to use certain types of devices that they did not practise enough to employ. Unlike previous studies (e.g. Hyland, 2004), the results from expert writers suggested that the higher use of stance markers does not absolutely indicate to a higher rhetorical level of writing. The epistemological stance of the study and the contextual factors do play a significant role on the quantity and type of the used stance markers.

References

S21 Yoshiho Satake

How error types affect the accuracy of L2 error correction with corpus use

The strengths of corpora in language learning have been reported (Flowerdew 2010); however, error correction in DDL settings has only been explored in a few studies (Satake, 2016, Tono, Satake & Miura 2014) despite that appropriate error correction is necessary for improving L2 writing. Although
Satake (2016) found that a corpus and dictionaries contributed to correcting different types of L2 errors, more studies are needed to prove the effects of corpus use. Thus, this study verifies Satake’s (2016) results by using the data from different students to examine the specific effects of corpus use on L2 error correction. It focused on how error types affect the accuracy of L2 error correction with corpus use. Participants’ corpus use was compared with their dictionary use along with their non-use of both. They were given 20-minute instruction on how to search for a target word in a corpus and interpret the concordance lines. In addition to the research methods used in Satake (2016), the author used a chi-squared test and residual analysis to determine whether there was a significant difference in the number of corrections among the different reference resources and the different types of errors.

The procedure was as follows: (1) Timed essay task (25 minutes). 55 Japanese intermediate EFL learners in total wrote an essay on a topic given by the author without consulting a corpus and/or dictionaries. (2) The revision session (15 minutes). The students were given highlighted feedback for errors and corrected the highlighted errors, consulting the COCA corpus and English–Japanese dictionaries of their choice at least once each. The above procedure was repeated every week on nine and eleven occasions in 2014 and 2015, respectively. (3) The author collected participants’ essays and made error-annotated corpora to analyze the different effects of both types of references and the lack thereof. The author provided error tags, which had information about parts of speech, error types, and the reference material used for error correction.

The results show a similar tendency as Satake (2016), and the new findings indicate that corpus use promoted significantly more frequent and accurate corrections of omission errors. The strengths of corpus use were easy access to the exact target phrases and the information of frequency of co-occurrence words. COCA especially helped the participants correct such errors as omission errors and word order errors because in most instances the participants could find the exact target phrases, refer to the information on the frequency of co-occurrence of words, and use the information when they searched for the target word(s) to correct their errors. The advantage of the frequency information provided by the corpus was prominent, considering that dictionary use was not very useful in correcting omission errors since dictionaries have fewer example sentences and did not give students access to co-occurrence frequency information. The findings suggest that efficient corpus use for accurate error correction requires teachers to consider error types. So that DDL can be a practical option in L2 language classrooms, fine adjustments of DDL are needed.

References


---

**S23 Elen Le Foll**

“They were walking in a corridor when suddenly the mummy appeared” *A corpus-based study of narrative texts in secondary school EFL textbooks.*

Most secondary school foreign language courses follow commercially published textbooks. These textbooks can thus be considered “one of the primary sources of [foreign language] input in the classroom” (Tono, 2004, p. 45). It therefore follows that thorough evaluations of the authenticity of the linguistic content of these secondary school EFL textbooks is paramount.

Since the 1980s, corpus-based methods have been applied to investigate the linguistic features of Textbook English. These studies usually focus on one lexical-grammatical feature – such as collocations (e.g., Koprowski, 2005) or reported speech (e.g., Barbieri & Eckhardt, 2007), and compare the frequencies, functions and patterning of the chosen feature(s) to those found in a reference corpus of native speaker English. Taken as a whole, the results of these studies undoubtedly demonstrate that Textbook English constitutes a distinct form of English that, in many respects, differs greatly from authentic, naturally occurring English.

However, many studies exploring Textbook English base their evaluation on the authenticity of the linguistic content on data from general English corpora such as the BNC (e.g., Alejo González, Piquer Piriz, & Guadalupe Reveriego, 2010; Chen, 2016; Zarifi & Mukundan, 2012). Considering their high proportion of professionally written texts, it may be argued that these corpora can only be partially representative of school English learners' target language. Moreover, most methodologies agglomerate all of the text types found in textbooks, thus ignoring considerable differences between the different registers (e.g., letters vs. newspaper articles) and production modes (e.g., transcripts of dialogues vs. narrative texts) found in textbooks. So far, the few studies that have considered the variety of text types within textbooks have focused on the representation of spoken language in textbooks. Thus, Dieter Mindt (1987, 1992), Ute Römer (2005) and others compare textbook dialogues with corpora of native spoken or written-to-be-spoken language. However, to the present author's best knowledge, no Textbook English study has yet to hone in on any one specific *written* register of school EFL textbooks.
The textbook corpus compiled for this study consists of nine series of EFL textbooks (42 textbooks in total) published between 2006 and 2017 and currently used in secondary schools in France, Germany and Spain. The textbook data has been manually tagged for text type in order to create spoken, narrative, informative and instructional textbook language subcorpora. This paper reports on an analysis of the lexico-grammatical features of narrative texts featured in EFL textbooks.

In a first step, Multi-dimensional Analysis (cf. Biber, 1986, 1988) is applied to map out the lexico-grammatical specificities of the narrative texts featured in EFL textbooks. Quantitative and qualitative analysis of the frequencies, functions and collocations of linguistic features typically associated with narrative texts are carried out. The occurrences of these features in school EFL textbooks are compared to those occurring in a specifically assembled corpus of children and teenage British English fiction. This relatively small, yet highly specific, reference corpus is designed to be as representative as possible of students' target learner language within the specific register of narrative texts.

In light of the observed differences between the textbook narrative subcorpus and the youth fiction corpus, recommendations are made to improve the authenticity of the narrative texts students are exposed to via their school textbooks.

References


---

**S26 Suresh Jampa**

*The effect of using corpora as a resource to enhance the language ability of the teachers in rural India*

This study examines the effect of using corpora as a resource for enhancing the language abilities of the teachers in rural areas using mobile applications. The main aim is to collect data in three forms such as audio, video, and written texts on teacher’s language in the classroom from the teachers who have a very good flair in English. Tools such as interview protocols, questionnaires, were used for the purpose. The list of frequently used words, phrases and sentences arrived after analyzing the data was made available to the teachers of secondary and primary schools in rural areas. After giving access to the data through what’s app, time was given to refer, use, and practice in all possible contexts. They were encouraged to give online responses. Later questionnaires were administered and focused interviews were conducted to examine the effect of using this corpus to enhance the language abilities of the target group. The study is of qualitative nature as data was collected through questionnaires and interviews. It has implications to the teachers of English, teacher trainees, trainers etc.

Though India is a multi-lingual country, the medium of instruction in most schools and colleges in recent times is English. This may be due to the emergence of English as an International language or the government’s response to the public demand (NCF 2005). Though the language of teaching / medium of instruction is English, many school teachers in rural areas of India, have very limited flair in English. The training facilities available to them are very meagre. So, they continue with very less or no training. Though the medium of instruction is English, some teachers, hardly use English as the language/medium of communication. Teachers should be able to communicate in English. Failure to use language as the medium of communication in the classrooms, maybe considered as a lapse and the learners may be at some dis-advantage. So, teachers should be able to communicate well with the
learners while teaching. To address this problem, a need to create a corpus on the language used by the teachers in the context of teaching in the classroom was felt, so that it will help the needy, to access language for teaching, practicing and reference purposes.

References

Szu-Yu Liu and Hsien-Chin Liou

Exploring the relation between English writing motivation, task engagement, and uptake of corpus-informed corrective feedback: A longitudinal study

Writing motivation plays an important role in both traditional L2 writing instruction and corpus-enhanced writing contexts. However, very limited motivation research has been conducted in either context. Receiving feedback for writing is believed to enhance learners’ motivation (Cardelle & Corno, 1981). Drawing insights from the literature of corpus-informed corrective feedback (e.g., Tono et al., 2014) and recent writing motivation research (Waller & Papi, 2017), the study examined how writing motivation influenced students’ uptake of corrective feedback with corpus consultation over one semester. Task engagement which is regarded as motivated behaviors during tasks (Dörnyei & Kormos, 2000; Han & Hyland, 2015; Wang et al., 2015) was carefully documented over time so as to discover why and how the learners were (or not) motivated to use the corpus tools for correction.

By adopting a questionnaire on the students’ writing motivation and feedback orientation (Waller & Papi, 2017), the study documented 16 EFL students’ changes in writing motivation, task engagement with corpus use via three observations, and error reduction between essay drafts over one semester. The students wrote three multi-draft essay assignments plus one diagnostic essay. The instructor marked around five errors on their drafts before the students used two English-Chinese concordancers or COCA to correct errors plus other reference resources. By analyzing students’ records of revision and screen recording, the result of error correction was positive with more successful corrections with corpora. Corpus use increased as a sign of more engagement over the semester (from 8.9% to 61.2%), and was positively correlated with the students’ writing motivation. Comparison of the questionnaire data over the semester shows a slight increase of writing motivation and feedback seeking.

Close examination of two focal students with data triangulation indicate that their cognitive, behavioral, and emotional engagement was closely related to the effects of feedback processing aided by corpora, in spite of fluctuation of different learners’ engagement over time (cf. Han & Hyland, 2015). With more successful corpus-aided error corrections, student A commented that his writing ability had been improved after corpus use and learning in the semester. Using concordancers was
interesting and he was willing to spend time on using them as references by extending such cognitive support. With less corpus use and successful corrections, student B commented using concordancers was helpful, but she still liked to use Google translator. It was not easy to find patterns via corpus tools and the interface of corpus tools looked a little complicated, and she was not willing to spend time on using concordancers for correction, choosing to stay in her comfort zone. Although every student received the same pedagogy with similar reference tools in the same context, they had different engagement toward corpus consultation.

In conclusion, corpus-informed feedback processing can be beneficial for improving students’ writing. The process of using corpora to correct errors may increase student’ writing motivation simultaneously. Effective procedures are suggested in orienting students toward full engagement with corpus consultation which leads toward feedback uptake and acquisition of word patterns as well as language development.

References


S30 Tzu Wei Yang and Hsien Chin Liou

*Pattern hunting via DDL at the drafting stage: How does the proficiency of EFL college students make a difference?*

Corpus consultation and data-driven learning have brought huge impact on foreign language learning. For L2 writing, DDL for error correction or general writing purposes has attracted more writing
researchers’ attention in the past, compared with using DDL for pattern hunting (Kennedy & Miceli, 2010). Integrating corpus consultation to help students generate ideas by re-using vocabulary items they find in concordance lines in their speech or writing has not been examined sufficiently in various instructional contexts (exceptions, Geluso & Yamaguchi, 2014; Kennedy & Miceli, 2010, 2017). The project bridged the gap by investigating how EFL students compose essays in their drafting stage via pattern-hunting or “observe-and-borrow” acts (Kennedy & Miceli, 2017) with the assistance of corpus consultation.

We documented the differences of three groups of undergraduate EFL students (2 first-year and 1 second-year groups, n = 49) in an Asian college writing program. Since corpora were new to all participants, eight-week training of using corpora in a process-oriented writing class was provided. Then, they drafted one essay assignment by consulting COCA or two Chinese-English concordancers. We collected their essays, written records and video files of corpus consultation, responses to an after-writing reflective questionnaire, and their interviews.

Both quantitative and qualitative analyses and their triangulation were conducted on the data sources including average keywords consulted per essay, proportions of consulted words included in drafts, and profiles of key words in each group based on Nation's BNC-20 word bands (2004, using http://www.lextutor.ca/, VocabProfile). Both groups looked up around 1.35 word/per 100 words in essays and consulted more verbs than other categories and mainly K1 and K2 words but the sophomores looked slightly more K3 to K6 words. Unsurprisingly, the sophomores wrote longer (294.39 words/essay) and included more consulted words into their own writing (90.48%) using phrasal units (42.1%), whereas the freshmen, shorter (252.53) and included fewer (87.5%) mainly using word units (64.55%). Their consultation processes did not differ much and mainly relied on pattern defining strategies and looked up words for confirmation (Yoon, 2016).

Students’ perceptions indicate that they were positive about integrating corpora for drafting and that learning the new tools changed their ways of English learning and writing, but they needed more training of examining sentences for word patterns. Although the sophomores may not be more proficient than freshmen, they seemed to perform slightly better concerning corpus use. It can be attributed to their longer time of English learning. How proficiency interacts with corpus consultation skills warrants more future research. Verbs seem a troublesome category when L2 students have to show their productive vocabulary knowledge in writing. While our learners seem goal-oriented by incorporating about 88% of consulted words to fulfill their vocabulary needs while drafting, they take concordancers as tools similar to dictionaries with very limited exploration mentality of pattern hunting. Teachers need to model ‘creative use’ of exploring concordance lines for writing and provide more training opportunities for their students’ observe-and-borrow acts, if such merits are recognized for their writing process and language development.
References


S31 Marie-Louise Brunner and Stefan Diemer

“We still can convey our uhm ... meaning ... and that's okay then” – Introducing a new corpus of English as a Lingua Franca (ELF) Skype conversations and its use in ELF-aware language teaching

While language classrooms are still mostly shaped by native speaker ideals, Jenkins (2012) calls for a stronger role of English as a Lingua Franca (ELF) in the classroom. Arguments have since been made for ELF-aware classroom practices (Vettorel 2015) and ELF-informed teaching (Seidlhofer 2015), focusing on communicative effectiveness rather than correctness. We use examples from CASE (2018), a corpus of ELF Skype conversations between students from nine European countries, to develop an ELF-aware language teaching approach.

CASE was compiled from 2012-2017 at Saarland University and Trier University of Applied Sciences, Germany, in cooperation with partners from Belgium, Bulgaria, Finland, France, Italy, Spain, Sweden, and the UK. The corpus consists of dyadic Skype conversations taking place in an informal context, with participants discussing a wide range of academic and cultural topics. It provides video and/or audio, as well as a transcription component including pragmatic aspects such as the use of nonverbal, paralinguistic, and plurilingual resources. The first part of the corpus is now freely available to researchers. CASE has considerable potential as a teaching resource as it provides real-language data in which issues that relate to students’ everyday experience are discussed. We use examples from CASE to illustrate two main didactic approaches: raising students’ language awareness and teaching them concrete communicative strategies.

First we focus on the demonstration and discussion of ELF as a form of communication in the classroom. This allows students to recognize the inherent flexibility of ELF and illustrates the use of typical ELF strategies such as code-switching (Cogo 2009, Klimpfinger 2009), approximation...
(Mauranen 2012), innovation, non-verbal resources (Brunner et al. 2016, 2017). Corpus examples increase students’ language awareness and help them recognize how these strategies are used to advance communication, and that they do not have a negative effect (“let-it-pass”, Firth 2009). In a next step, we present ways in which students can be encouraged to apply selected ELF-associated discourse strategies such as explicitness, metadiscourse, co-operative solutions (Mauranen 2006, 2012), repetition/rephrasing (Kaur 2009, Mauranen 2012), definitions (Brunner 2017), and metalinguistic comments (Vettorel 2014). In addition to the methodological issues and reflections, there are also other challenges that have to be addressed in order to provide ELF with a place in the language classroom. Our paper thus concludes with a discussion of the role of ELF in the curriculum and the issue of evaluating language use that focuses on effectiveness rather than correctness. By increasing students’ ELF awareness and by providing them with concrete ELF-based communication strategies that could be integrated into the curriculum, the paper illustrates how ELF can be part of a language learning environment that relativizes the native speaker ideal maintained by many teachers and institutions and that allows students to focus on successful communication as the key aim of instruction, supporting creative and assertive language use rather than penalizing it. The paper also presents CASE as a new resource for integrating real-life ELF data into the classroom.

References


---

**S33 Cristóbal Lozano, Amaya Mendikoetxea and Paul Rollinson**

*The design and use of corpora for SLA: CEDEL2 and WriCLR*

The main aim of SLA research is to build models of the underlying representations of learners at a particular stage in the process of L2 learning and of their developmental process. The central source of evidence for this is the language produced by learners, whether spontaneously or through data elicitation (Myles 2005: 374; cf also Myles 2015). The success of SLA research relies crucially on the validity and reliability of these procedures.

This paper (i) provides a critical overview of the use of corpora in SLA; (ii) describes in detail two comparable learner corpora that have been designed specifically for the study of SLA: WriCLE (*Written Corpus of Learner English, http://wricle.learnercorpora.com/*) and CEDEL2 (*Corpus Escrito del Español como L2, http://cedel2.learnercorpora.com/); (iii) presents the research carried out with these corpora, and (iv) points out the way forward for corpus use in SLA.

Both corpora are available online, have been designed according to the same standard design principles recommended by Sinclair (2005) (see Rollinson & Mendikoetxea 2010 for WriCLE and Lozano 2009a and Lozano & Mendikoetxea 2013 for CEDEL2) and have clear advantages over other available learner corpora:
a) They are relatively large-scale learner corpus (c. 750,000 words to date, aiming at 1 million words).

b) As the corpora are designed primarily for SLA purposes, they contain texts from learners at different stages in the acquisition process, allowing for cross-sectional research, and proficiency level has been determined using standard tests. This is essential to conduct reliable studies of L2 acquisition and interlanguage development (see Tono 2003).

c) Both corpora are comparable to equivalent native speakers’ corpora: WriCLE can be compared with LOCNESS (Louvain Corpus of Native English Essays, see https://uclouvain.be/en/research-institutes/ilc/cecl/locness.html, and CEDEL2 contains a similarly designed Spanish native speaker subcorpus. This allows for the contrast of interlanguage data against the native norm under equally comparable conditions.

d) While CEDEL2 contains texts from L1 English – L2 Spanish learners, WriCLE contains texts from L1 Spanish – L2 English learners. These language pairings permit detailed analyses of transfer phenomena in both directions, together with the investigation of language-specific vs universal influence in L2 acquisition.

e) For each learner, both corpora contain precise and detailed background information (e.g., proficiency level, age of first acquisition, length of exposure, learning environment, language use patterns, etc.), which is essential to conduct L2 research concerning not only interlanguage grammars, but also critical period effects, language use patterns, etc.

The corpora have been used to study word order, anaphora resolution, subject realization, collocations, etc., and to explore hypotheses about cross-linguistic transfer and deficits at the interfaces (The Interface Hypothesis as postulated by Sorace 2000, 2005) (see, for example, Lozano & Mendikoetxea 2011; Mendikoetxea and Lozano in press; and Lozano 2009b, 2016, as well as all the publications cited in http://cedel2.learnercorpora.com). More pedagogically oriented studies have also been conducted for WriCLE (see e.g. Mendikoetxea, Murcia and Rollinson 2006 and the TREACLE project http://treacle.es/index.html). In the conclusion, we point out the way forward for the use of corpora in SLA.

References


Mendikoetxea, A. & C. Lozano (in press). From corpora to experiments: methodological triangulation in the study of word order at the interfaces in adult late bilinguals (L2 learners)


Applying corpus use across the curriculum: analysis of small corpora by pre-service teachers as a means of engaging with student feedback

This paper concentrates on the use of corpus analysis in teacher training. Set in a low-tech environment, working with future teachers of French as a Foreign Language (who are not seasoned corpus linguists), it looks at how students reinvest basic corpus techniques in the form of an assignment requiring analysis of student feedback on assessment.

Students (n=20) received 12 hours of training in the use of concordancing software and online tools to analyse texts and corpora in an initial course basically geared towards developing pre-service “corpus literacy” (Heather & Helt 2012; Zareva 2017). Among the stated learning outcomes of this course was the ability to analyse any given text or corpus using basic lexicometric means, and also to tease out ‘typical’ elements. While this course was assessed with a specific assignment, the possibility for transfer of knowledge and methods was targeted in a second course. The second course (12 hours), somewhat different from the first, was concerned with assessment methods and issues arising in language testing. Part of the assignment for this course involved analysis of a small corpus made up of student feedback on different types of evaluation. The texts making up the feedback corpus (9500 words) were the anonymous answers given by previous cohorts to two simple questions: describe in detail your best exam/assignment ever (4600 words); describe in detail your worst exam/assignment ever (4900 words). No specific mention was made of the transferability of corpus techniques to this second course and students were simply asked to engage with the data, looking at how likes and dislikes were conveyed in order to better understand problems and feel-good factors (and their impact as perceived through the language used to describe them) with different types of assessment.

This paper looks first at the context in which this practice was developed, discussing the rationale and objectives. It then goes on to outline some of the basic features of the small feedback corpus before looking critically at the types of observations and conclusions offered by the students engaging with a task where corpus techniques are a means to studying issues pertaining to methods of evaluation and assessment and the impressions they make on people. It is argued that transferability of knowledge and methods of corpus analysis (cf. Boulton 2011) is a useful component in the (future) teacher’s “tool box” (Römer 2006: 105), allowing systematic exploration and depth of expertise in areas of study that may not initially invite the use of corpora (cf. Jackson 1997) and that do not relate to specific linguistic objectives in language learning. The issue of fairly low-tech, low-expertise methods is discussed (cf. Tyne 2012) as is that of the relevance of small corpora (Ghadessy et al. 2001). The question of corpus literacy in pre-service teacher training (Leńko-Szymańska 2017; Zareva 2017) is raised.
Maggie Charles

First steps in DIY corpus use: What questions do students ask and what outcomes do they achieve?

Research on the pedagogic use of corpora in academic writing has recently highlighted the importance of student consultation processes (e.g. Pérez-Paredes et al., 2011; Pérez-Paredes et al., 2013; Yoon, 2016a, b). That research typically provides students with access to one or more large general corpora, and student searches are prompted by the demands of writing an assigned paper (Yoon, 2016a, b) or carrying out a given grammar task (Pérez-Paredes et al., 2011; Pérez-Paredes et al., 2013). The present study takes a somewhat different approach, focusing on questions determined by the students themselves and addressed using a small do-it-yourself (DIY) corpus in their own discipline.

The paper reports on data from an EAP course for doctoral students, in which participants built a corpus of research articles in their field for use in editing their thesis. After an introductory session on corpus work and the AntConc software (Anthony, 2014), each student built an initial corpus of 10-15 research articles (about 100,000 words) and interrogated it using questions they devised themselves.
on language issues or problems that concerned them. Students recorded these issues/problems (called here ‘queries’) and their outcomes on worksheets, which provide the data employed here. This study reports on the students’ queries and the outcomes they achieved in their first consultation of their tailor-made corpus.

Data are available for 63 students, all with L2 English; 49% studied natural sciences, 27% social sciences and 24% arts/humanities. Students recorded 206 query-outcome sequences, which were analysed drawing on categories established in previous research (Frankenberg-Garcia, 2005; Kennedy and Miceli, 2010; Park, 2012; Yoon, 2016a). Queries were analysed as either ‘verifications’ (checking a given item), or ‘elicitations’ (ascertaining an unknown item); further analysis coded the focused linguistic feature (e.g. phrase, noun, preposition). Outcomes of the queries were analysed as ‘satisfactory’ (information retrieved answered the query); ‘partially satisfactory’ (student indicated dissatisfaction with the information retrieved); or ‘unsatisfactory’ (no hits or abandoned).

Results showed that the majority of queries were verifications (82%); the most frequent types concerned phrases (33%), prepositions (21%), verbs (14%) and nouns (13%). Despite the small size of the corpora, students answered 84% of queries to their own satisfaction (83% of verifications, 89% of elicitations). Examples from the worksheets are given below.

Satisfactory verification of a preposition  
Query: *existing procedures lack of or in selectivity* [av prep]  
Outcome: *lack of selectivity* [s]

Partial verification of a noun  
Query: not clear about how to use the word *mediation* [av N]  
Outcome: *mediation of* can be used. But I was criticised by my examiners for my use of nominalisation *mediation of* and still not sure what is wrong with it. [p]

Unsatisfactory elicitation of a phrase  
Query: alternative words for *readily computable diagnostic tool* in *these indices for readily computable diagnostic tool*. [ae phr u]  
Outcome: not sure how to do this. [u]

This paper reports in more detail on the queries and outcomes data and discusses their implications for facilitating the early stages of corpus consultation. It argues that the examination of such query-outcome sequences has a role to play in devising effective tasks for novice corpus users in academic writing.

References


Nina Vyatkina

Data-driven learning beyond concordancing for improving the breadth and depth of L2 vocabulary knowledge

Teaching and learning L2 vocabulary is a prominent topic in Instructed Second Language Acquisition research (ISLA), and studies of non-corpus teaching methods have shown that form-focused instruction (Peters, 2014; Webb & Kagimoto, 2011), repeated exposure to target vocabulary (Laufer & Rozovski-Roitblat, 2011), visual input enhancement (Peters, 2012; Sharwood Smith, 1993; Sonbul & Schmitt, 2013), and high involvement load (Kim, 2008; Hulstijn & Laufer, 2001) all facilitate L2 vocabulary learning. It is then not surprising that Data-Driven Learning (DDL), a corpus-based teaching method that exhibits all these characteristics, has been found to facilitate the acquisition of L2 vocabulary, especially collocations (e.g., Chan & Liou, 2005; Daskalovska, 2015; Gordani, 2013; Rezaee et al., 2015). A recent meta-analysis has confirmed that DDL is effective and more efficient than some traditional teaching methods for teaching vocabulary (Boulton & Cobb, 2017). However, there still are some unaddressed gaps in both ISLA and DDL vocabulary research. In particular, few ISLA studies have explored the development of the depth of L2 vocabulary knowledge, i.e. aspects of words beyond the basic L1-L2 translation, which is especially important for developing advanced L2 proficiency. In DDL research, most studies have focused on the use of concordancers, whereas the effectiveness of other corpus tools still remains underexplored. Furthermore, there is a dire need for
both ISLA and DDL research on target languages other than English. This study addresses these gaps by combining the methodology from both research strands: ISLA and DDL.

The participants were US university students with a high-intermediate German proficiency, and the target structures were lexical and morphological aspects of German verb-noun collocations. The study compared the effectiveness of computer-based and paper-based DDL in a single-group, pretest-posttest design. In the computer-based method, participants searched a large open-access German corpus to discover and learn L2 collocations with the help of a suite of corpus analysis and visualization tools: word clouds, lists ranked by the association strength, and concordance lines. In the paper-based method, participants worked with teacher-prepared materials that included corpus printouts. The time on task in both conditions was the same. The multilevel modeling analysis results show that both teaching methods led to significant gains, thus confirming findings about the efficacy of different DDL types from previous research (Boulton, 2010, 2012; Vyatkina, 2016). The novel finding of the study is that DDL was effective not only for the increase of target words and collocations in learner production but also for improved morphological accuracy, and that the hands-on method was more efficient for the latter. This finding is especially important for inflectional languages like German (Boers & Lindstromberg, 2012; Stengers et al., 2011). The study also revealed that learners with a larger overall vocabulary size gained more from DDL. The practical implication of this study is that high-intermediate students can effectively use different corpus tools, not only concordancers, for enlarging their L2 lexicons. Using this method would foster independent learning and relieve the teacher from time-consuming and paper-consuming preparation of printed corpus-based materials.

References


---

**S53  Ayman Alghamdi & Eric Atwell**

**An Arabic corpus-informed list of MWEs for language pedagogy**

The phenomenon of multiword expressions (MWEs) in languages has attracted the attention of researchers in various disciplines (e.g., linguistics, psychology, language pedagogy LP and natural language processing NLP). Hence, this phenomenon has been researched from a number of different scientific angles. A considerable amount of research has emphasised the major role of MWEs in the process of analysing and understanding languages. From the applied linguistic perspective, many
studies have emphasised the crucial importance of including formulaic language and MWEs in second language learning and teaching. Several researchers have highlighted the fact that the mental lexicon is not merely represented by single orthographic words, but rather it incorporates longer multiword Expressions (e.g., Pawley and Syder, 1983; Sinclair, 1987; Wray, 2002; Nesselhauf, 2005). Other research (e.g., Durrant, 2008; Martinez, 2011; Ackermann and Chen, 2013) have attempted to develop different MWEs lists, which can be used as a pedagogical tool in language teaching and learning such as, material design, curriculum developments and language testing (e.g., Biber and Barbieri, 2007; Ellis and C, 2010; Martinez, 2011). Thus, this study reports the construction of an Arabic corpus-informed list of MWEs for language pedagogy.

A hybrid model was adopted for extracting AMWEs from the corpus that combined automatic and manual extracting methods, based on well-established quantitative and qualitative criteria that are relevant from the perspective of language pedagogy.

Figure 1: Diagram of the hybrid framework for extracting a pedagogical listing of AMWEs

The corpus used in this experiment is the ArTenTen corpus (Arts et al., 2014) which contains more than 7.4 billion tokens, the corpus was automatically analysed by two different toolkits for SA morphological and linguistic disambiguation, the first one was the Stanford Arabic Parser (SAP) (Manning et al., 2014) and the second one was MADAAMIRA toolkit (MA) (Pasha et al., 2014) for Arabic morphological and shallow syntactic analysis. The corpus content was compiled from various web domains by the SpiderLing tool for web scribing. Table 1 provides the basic information about the ArTenTen corpus.
Table 1: Basic information about ArTenTen corpus

To the best of our knowledge, this is the largest available SA corpus with acceptable quality and detailed information about the corpus preparation and compiling processes. Most available SA corpora are limited in terms of their size or the scope of SA representations. However, when it comes to corpus linguistics, these two criteria of corpus construction considered as the core elements in any corpus evaluation task (McEnery and Gabrielatos, 2008; Corpas Pastor and Seghiri, 2010). ArTenTen corpus represented different SA variations, and it was divided into 28 sub-corpora according to the most common domains in which the web crawler targeted during the corpus compiling process. The crawler tool used more than 116k domains to ensure the comprehensive representations of SA, these domains mainly from Arabic-speaking countries and also includes several others countries which had large number of SA websites.

The pedagogical implications of this list are estimated to facilitate the inclusion of AMWEs in the process of learning and teaching Arabic, particularly for non-native speakers.

1 The ArTenTen corpora can be accessed through SketchEngine website: https://www.sketchengine.co.uk.

2 This tool is available on the following link: http://nlp.fi.muni.cz/trac/spiderling

References


Evaluating the effect of data-driven learning (DDL) on the acquisition of academic collocations by advanced Chinese learners of English

This study aims to explore the effect of data-driven learning (DDL) on the acquisition of academic collocations by advanced Chinese English as a Foreign Language (EFL) learners in a Chinese university. Collocations, prefabricated multi-word combinations, are considered to be a crucial component of language competence which indicates the central role they should play in language teaching and learning. However, collocations remain a challenge to L2 learners at different proficiency levels, and particularly a difficulty to Chinese learners of English (e.g. Fan, 2009; Granger & Bestgen, 2014). From a pedagogical perspective, collocations have so far attracted only limited attention in language teaching in the Chinese language teaching classroom. This study, therefore,
focuses on the effectiveness of the teaching of academic collocations to advanced Chinese learners of English, using a specific pedagogical approach, the corpus-based data-driven learning approach (DDL). Although a considerable number of corpus-based studies in the pedagogical domain have indicated that the corpus-based data-driven learning approach (DDL) has been argued to offer an effective teaching method in language learning (e.g. Nesseulhauf, 2005) and it may be beneficial to learners, a DDL approach has so far not become one of mainstream teaching practices and the large-scale, quantitative studies carefully evaluating the effectiveness and assessing the benefits of DDL in the acquisition of academic collocations are limited in number when compared to a different method of teaching of collocations (Bouton, 2010; Boulton & Cobb, 2017; Chambers, 2005). This study, therefore, sets out to examine the contribution of DDL and another method of teaching of collocations (using of online dictionaries) to the acquisition of academic collocations, with regards to the receptive and productive knowledge of academic collocations, compared with the traditional teaching approach.

The study used data from 120 Chinese students of English from a Chinese university and employed a quasi-experimental method, using a pre- test-and-post-test (including delayed test) control-group research design to compare the achievement of the use of DDL and online dictionary with the traditional teaching approach in teaching academic collocations to the Chinese EFL learners. One of the experimental groups used #Lancsbox (Brezina, McEnery & Wattam, 2015), an innovative and user-friendly corpus tool. The other experimental group used the online version of the Oxford Collocations Dictionary. The results were analysed for the differences in collocation gains within and between the three groups. Those quantitative data were supported by findings from surveys and semi-structured interviews exploring learners’ attitudes towards different approaches and tools and linking their attitudes with the test results. Results indicated that the majority of learners in two experimental groups believed that corpus and collocations dictionary consultation were useful in learning academic collocations and improving their academic writing. The findings contributed to our understanding of the effectiveness of DDL for teaching academic collocations and suggested that the incorporation of technology into language learning can enhance collocation knowledge.

References

S58 Huifen Lin and Pinshuan Lee

The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention

Research on data-driven learning (DDL) has generally suggested that DDL can facilitate vocabulary acquisition (Bernhardt & Ellis, 1993; Gardner, 2013; Johns, 1991; Nation, 2001; Schmitt, 2000) and retention (Craik & Lockhart, 1972); however it still has some limitations: the difficulties and a great amount of time needed to make an inference. A call for a combination of DDL and the traditional teaching approach to compensate for the limitations was proposed (Balunda, 2009; Boulton, 2010; Chambers, 2007; Chan & Liou, 2005), but few studies (Kaur & Hegelheimer, 2005; Lin & Lee, 2015) have answered the call. To address aforementioned limitations, the current study combined DDL with the deductive approach employed in the traditional teaching approach and compared the effect of the deductive and inductive approach under DDL context on vocabulary acquisition and retention.

Additionally, conflicting interaction effect between teaching approach (i.e. the inductive and deductive approach) and language proficiency was found by few studies (Shaffer, 1989; Wang, 2002). Because of the few and conflicting findings and the high possibility for vocabulary size to influence inductive DDL learners’ vocabulary acquisition, the present study examined the interaction effect between teaching approach and learners' vocabulary size on vocabulary acquisition and retention.

Twenty-seven (N=27) participants with high-intermediate reading proficiency or above were equally distributed into the inductive and deductive DDL group based on their large or small vocabulary size, and then received the modified VKS delivered as the pretest, immediate posttest, and delayed posttest to assess the acquisition and retention effect of two teaching approaches.

The results showed that both the inductive and deductive approaches were effective to facilitate learners' vocabulary acquisition and retention under DDL context, but the no significant differences were found in vocabulary acquisition and retention between the two teaching approaches. Additionally, the deductive approach appeared to be more efficient in vocabulary acquisition. Furthermore, no significant interaction was found between vocabulary size and teaching approach on vocabulary acquisition and retention in DDL context.
The results that deductive DDL was effective and efficient in vocabulary acquisition suggested it not merely maintained the advantages of inductive DDL, but also improved the limitations of it. EFL teachers are encouraged to adopt the deductive DDL to facilitate learners' vocabulary acquisition.

References

Balunda, S. A. (2009). Teaching academic vocabulary with corpora: Student perceptions of data-driven learning. Faculty of the University Graduate School in partial fulfillment of the requirements for the degree Master of Arts in the Department of English, Indiana University.


S60 Claire Wolfarth, Claude Ponton, Catherine Brissaud

Which Method to Develop a Natural Language Processing Tool to automatically analyze First Language Learner Corpora?

The purpose of our project is to build a computer aided longitudinal corpus of texts written by children between 6 and 11 (first years of acquisition of writing). Our corpus will contain at least about 3600 texts produced by 596 French pupils coming from 40 French primary schools. Several years ago, large primary school corpora of French did not exist, and when such corpora were built, they were not available online.

Today, more projects aim to create such corpora (Elalouf, 2005; Garcia-Debanc and Bonnemaison 2014; David and Doquet, 2016). However, those corpora are neither longitudinal nor associated to natural language processing (NLP) tools, but some methods from NLP related researches could be adapted (Wolfarth, 2017). After a short presentation of the natural corpus, we propose a method, giving the detailed processing chain, to create a tool for linguistic analysis purposes.

Regards to the specificity of our corpus, the approach we adopted uses annotation (Doquet et al., 2017). Then we developed an original approach based on multi-level comparisons between the literal transcription of each text and a normalized version of each of them. We chose this approach because we intend to use it again in other didactic contexts like dictation, copy and rewording.

Our method aim to align each single produced form of each text, with the normalized version. For that, several operations are required:

1) Morphosyntactic analysis and postagging: The normalized version is analyzed with TreeTagger, a tool for annotating texts with part-of-speech and lemma information. (Schmid, 2007);

2) Phonological conversion: The orthographic form is converted into a phonological representation with the tool LIA_PHON, a text-to-phoneme converter (Béchet, 2001);

3) Alignment: On the basis of the phonological representation, the literal transcription of the text and the normalized version are aligned with a tool that we have designed.

As a result of this three step process, for each form of each text of the corpus, we can give the normalized form, the grammatical type and the lemma. This tool elaborated with NLP method is not a 100% efficiency tool, and an evaluation of this tool is required. Nevertheless, we can already extract global trends in our corpus. To give an example, thanks to this alignment, we are able to extract all verbs of our corpus, and to observe variation of orthographic errors, according to the level of the child. We can also determine the frequency of errors that impacts the phonology, compared to the
frequency of the errors that have no phonological incidence. Alignment between each literally transcribed grapheme and the normalized grapheme, is in progress for the whole corpus.

Thus, this corpus and associated treatments could permit further linguistic analysis. We also plan to elaborate, based on the results of our corpus, didactic resources and teaching sequences.

References


S62 Claudia Wunderlich

Creating a subject-specific corpus and academic word list for business informatics

Creating subject-specific word lists from specialized corpora has proved to be of high relevance for learning and teaching of ESP including EAP (Nation 2016, Coxhead 2018). The research in this paper presents results of a corpus study into the language of Business Informatics, a discipline combining principles of computer science and business studies originating in Germany, from where it has spread to other countries as a degree programme at leading universities. In Germany, undergraduate students of Business Informatics usually study English for several semesters to prepare for an international career and also to enable them to read and understand books and journal articles as well as write their bachelor theses in English. The corpus compiled here and its yields are used for data-driven learning and creating corpus-informed materials for the English courses in the Business Informatics programme.
to optimize the learning of the relevant words and lexical bundles. A corpus of 2,000,000 words of written academic language is analysed applying a combination of the methodologies both of Coxhead (2000) and Gardner and Davies (2013) to identify the flemmas peculiar to the academic and technical language of Business Informatics. The corpus consists of journal articles from the leading international journals of the discipline such as ACM Computing Surveys, Information Sciences, and Information and Management reflecting the topics currently central to the discipline including data security, Internet of things, manufacturing, supply-chain management, and AI.

The study confirms the previous findings according to which subject-specific academic word lists deviate significantly from the above-cited general academic word lists. To achieve optimal reading comprehension in Business Informatics, this newly created list should therefore be given preference to increase the students’ knowledge of the words they are most likely to encounter in texts. Overall, the list contains approximately 550 words and is divided into ten sub-lists that need to be learned in addition to the Oxford 3000 word list the students should be familiar with by the time they start their university English course in Bavaria. Together with separate lists of acronyms (such as ERP, JIT, CAD), collocations, and lexical bundles consisting of up to four consecutive words, these sub-lists provide an excellent basis for a task-based English course with a lexical syllabus and enable setting long-term goals for vocabulary learning and the systematic acquisition of the most relevant academic words. The sub-lists can be used to inform the design of the various modules of the course and memorized with the help of flash cards in paper or electronic form. The corpus at the basis of the present study is part of a larger, comprehensive corpus of English for Business Informatics currently under construction.

References

S71 John O'Donoghue

An analysis of the use of structural and functional lexical bundles in L2 academic writing corpora.

If we accept the claim that lexical bundles “provide basic building blocks for constructing ... written discourse” (Biber et al. 1999: 184), analysing non-native speakers’ preferred bundles may reveal the
focus of their own writing strategies, while comparing such bundles with those of native-speaker writers may indicate what NNS writers tend to overuse or underuse. This paper examines the 50 most frequent lexical bundles in two corpora of non-native-speaker academic writing (L1 German speakers) at undergraduate and postgraduate level. These recurrent word combinations from 86 bachelor’s and 112 master’s dissertations submitted to the Berlin School of Business and Law were analysed according to the structural and functional categories established in *The Longman Grammar of Spoken and Written English* (Biber et al. 1999) and subsequently employed by Biber 2009; Biber, Conrad & Cortes 2004; Cortes 2002, 2004, and Hyland 2008. The largest group among the top 50 bundles in both corpora was classified as PP-based bundles (those beginning with a preposition followed by a noun-phrase fragment, e.g. *on the other hand*), comprising 44% of all bundles in undergraduate texts and 48% in postgraduate ones. Noun phrases with post-modifier fragments were the second most frequent group (e.g. *the scope of this*), at 20% and 18% respectively. Comparing these results with previous research on L1 (Biber 1999, Cortes 2002, 2006) and L1 and L2 academic writing (Chen and Baker 2010; Ādel and Erdmann 2012) we see that certain structural bundles were underrepresented in both corpora when compared to native-speaker writing, most noticeably: anticipatory it + VP/adjective + complement clause, e.g. *it is important to*; passive verb + PP fragment, e.g. *can be seen as*; (VP+) *that* clause fragment, e.g. *can be said that*; and (verb/adjective +) *to* clause fragment e.g. *to be able to*. In functional terms, the results show that a majority of bundles from both corpora were devoted to referential expressions (many of them intangible framing attributes, Biber et al. 2004), approximately a quarter to discourse organizers and an eighth to stance expressions. Thus we see that these writers were primarily concerned with structuring their experience and secondarily expressing “textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse” (Cortes 2004: 401); they were less interested in expressing epistemic evaluations or their attitude towards their own propositions (cf. Durrant 2015).

While acknowledging the high quality of writing that both corpora contain, these findings indicate some overuse of the ‘preposition + noun phrase fragment’ pattern which is closely linked to overuse of referential expressions and clear underuse of stance expressions by these non-native speakers. Far from being restricted to German speakers, this seems to follow a pattern in the literature; compare Hyland and Milton (1997) on Chinese students expressing qualification and certainty and Pérez-Llantada (2014) for similar underuse of probability and evaluative markers by Spanish writers. One strategy for teaching practice might be to encourage students to develop their own stance by exploring some of the neglected structures identified above, including (VP+) *that* clause fragment and (verb/adjective +) *to* clause fragment. Exposing NNS students to the wide range of expressions indicating stance, especially epistemic and modality bundles, and alerting them to appropriate hedging strategies including “the focused instruction of formulaic sequences” (AlHassan & Wood 2015) might help them create more balanced academic texts.
References


Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. Linguistics and Education. 17, 391-406.


L1 congruency in the evaluation of data-driven learning effectiveness: a study on Italian verb-noun collocations

Data-driven learning (DDL) aims at fostering noticing and awareness-raising processes in second language learning, most typically through the guided discovery of patterns in concordance lines (Johns, 1991; Sinclair, 2003). A recent meta-analysis of empirical studies aimed at evaluating the effectiveness of the approach reveals that the results reached so far are overall promising, at least as concerns English at predominantly advanced proficiency levels (Boulton & Cobb, 2017).

This study investigates the role that L1 congruency plays in evaluating the effectiveness of DDL. It analyses collocational proficiency data collected at four points in time over a period of 12 weeks from two groups of students: an experimental group, which followed a DDL approach, and a control group, which worked on traditional learning materials. The learning and teaching context is that of Italian language courses attended by pre-intermediate level Chinese students looking to enroll at Italian universities upon successful completion of their one year Italian language course. The learning aims were 64 verb-noun collocations, evenly distributed in 8 one-hour weekly lessons, and selected by combining error analysis based on the Longitudinal Corpus of Chinese Learners of Italian (LOOCLI) with DICI-A, a Dictionary of Italian collocations for L2 learners based on the Perugia Corpus (PEC), an Italian reference corpus (Spina, 2010, 2014).

The paper combines the theme of L1 influence in second language learning with the investigation of DDL effectiveness, in the hope of shedding some light in both domains of research. Results will be interpreted in the light of current debates about the role of collocations and L1 influence in language learning (Nesselhauf, 2005; Bestgen & Granger, 2014; Wang, 2016; Gablasova, Brezina, & McEnery, 2017).

References


Is there a relationship between self-regulation, lexical proficiency, and attitudes towards corpus use?

Since the results of any corpus use will, among other necessary skills, largely depend on the user’s willingness to engage in its usage, this study aims to research whether there is a relationship between students’ attitude towards corpus use, their self-regulation in vocabulary learning, and their lexical proficiency.

The concept of self-regulation has recently been largely embraced by SLA researchers (Dörnyei, 2005; Tseng et al., 2006; Oxford, 2017). Its introduction into SLA research was particularly encouraged by the criticism of language learning strategy paradigms (Oxford, 2017). This study applies Dörnyei’s model of strategic learning which derives its taxonomy from the framework of motivational control strategies (Dörnyei, 2001, 2005). In particular, the study applies the SRCvoc questionnaire in which the given taxonomy is adapted to the task of vocabulary learning (Tseng et al., 2006). Since “the SRCvoc does not measure strategy use but rather the learner’s underlying self-regulatory capacity that will result in strategy use” (Dörnyei 2005: 184), i.e. students’ aptitude towards vocabulary learning, and not the actual strategies, it has not been embraced willingly by all learning strategy researchers. However, Gao (2006, in Rose, 2012) sees these two approaches as compatible and not competing because they are measuring the beginning and the end product of the same event, the strategy use being the outcome of the aptitude, the initial incentive to use them.

Browsing corpora being one of such strategies indicated as potentially helpful in vocabulary acquisition leads us to assume that there would be a positive relationship between the measure of self-regulation and students’ attitude towards corpus use. Consequently, the research question 1 is: Do
students who achieve higher scores on SRCvoc, the measure of self-regulated vocabulary learning, have a more positive attitude towards corpus use?

A questionnaire measuring students’ attitude towards corpus use was constructed particularly for this purpose and administered to a group of participants who are not language students, but they take a language for specific purposes course as part of their study programme. As part of this course students engage in an exercise of building and searching a small specialised corpus with the aim of informing one of their written assignments.

The second measure that is correlated with students’ attitude towards corpus is their lexical competence. Thus, research question 2 is: Do students with higher lexical competence have a more positive attitude towards corpus use? Since lexical competence is a multidimensional construct, in order to answer this question, the study measures lexical diversity, lexical density or productivity, and lexical sophistication, the behavioural constructs corresponding to the theoretical constructs of size, width and depth of vocabulary knowledge (Read, 2000; Bulté, Housen, Pierrard & Van Daele, 2008; Pavičić Takač & Buljan, 2017). A range of available operational constructs/statistical measures is used to estimate students’ lexical competence based on their productive use of language in an assigned piece of writing.

References
The association between postgraduate students’ learning styles and their evaluation of corpus use in L2 academic writing

Computer-readable corpora have undoubtedly transformed research into English (Biber, et al., 1999; Sinclair, 1991, 2001, 2004a, 2004b; Tribble, 2013). The use of corpora has also been spread to the English language classroom, where learners are encouraged to search corpora to discover language patterns in real-life English by themselves. This new learning approach, also known as data-driven learning (DDL) (Johns, 1991a, 1991b), has demonstrated many benefits (Frankenberg-Garcia, 2005; Flowerdew, 2008; Boulton, 2011; Yoon, 2011; O’Sullivan, 2007). Its drawbacks, on the other hand, have also been reported (Cobb, 1997; Ädel, 2010; Charles, 2011); as Flowerdew (2009) has pointed out, for example, DDL might not be suitable for all types of learner. However, with the exception of a study by Boulton (2009), who measured DDL learning styles on four scales (active/reflective, sensing/intuitive, visual/verbal, sequential/global), little research has been carried out exploring the role of learner variables in the effective implementation of DDL. This study takes Boulton’s research a stage further and investigates the correlations between 10 learning styles, three other learner variables, and learner perceptions of DDL.

The students in this study (117 postgraduate research students) attended a corpus-based research writing workshop in computer labs. Prior to the workshop, a learning style questionnaire adapted from Cohen, Oxford and Chi (2002) and Ehrman and Leaver (2002) was distributed in order to measure learners’ learning-style preferences with regard to the learner variables mentioned above. After the workshop, the participants completed a post-workshop survey focussed on their perceptions of the affordances of corpora in English L2 research writing.

The post-workshop survey reveals participants’ highly positive evaluation of the workshop and the usefulness of corpus tools in research writing. Correlation analyses of results from both surveys suggest that, with one exception, in general, learning styles do not have a strong impact on learners’ predispositions to use corpora in research writing. The exception was the leveler-sharpener dimension (Spearman correlation of: $r = 0.183, p = 0.049$), with sharpeners (those seeking distinctions and storing different memories separately) tending to be better disposed to corpus-assisted learning than levelers (those being likely to reduce differences and clump materials together).

Learners’ overall preference for using computer resources in language learning also corroborates with their openness to the new tools introduced in the workshop. A significant correlation was identified between students’ use of computer resources (frequent user vs. non-frequent user) and their comments on the useful aspects of the workshop (Fisher’s exact test: $x^2 = 15.626, df = 8, p = 0.013$). A further
significant correlation was identified between learners’ school level and their dissatisfaction with certain aspects of the workshop; however, the relationship appears not to be linear.

Findings from this study shed new light on the long unresolved question about who would benefit more from the DDL approach and may also help to relieve some of the concerns about the validity of corpus applications in English language teaching. The study shows that DDL may be suitable for a wide range of learners having different style preferences.

References
Profiling learners through pragmatically and error annotated corpora

This study presents a novel way of analysing learner language based on data from LINDSEI China that has been annotated both pragmatically and for errors. The materials used for the study consists in 20 randomly extracted interviews from the corpus, focussing on one particular sub-task/genre of LINDSEI.

To be able to perform the multi-level analyses, the data was first annotated using a new XML annotation scheme that introduces a separation into functional (speech-act) units (cf. Weisser in press). This categorisation provides a more realistic basis for quantifiability/countability of (non-
standard) features than e.g. suggested before in Callies (2013, p.18), where the turn is assumed to be a suitable unit of comparison, or Götz (2015), where comparisons are carried out based on frequencies normed to 100 words, as it makes it possible to quantify errors in relation to the number of communicative functions a learner is attempting to carry out.

Annotation and analyses were carried out in a modified version of the Dialogue Annotation and Research Tool (DART; Weisser, 2016a). In the pre-processing phase, the files were first split into appropriate functional units, then annotated for errors manually, and finally enriched with initial pragmatics-relevant annotations fully automatically. During the post-processing phase, the error annotation scheme was consolidated, and the pragmatic annotation manually corrected. The relevant features were then extracted and normed by the highest common denominator in terms of number of units (cf. Weisser 2016: 175) to obtain comparable per-unit results across all speakers.

The error annotation scheme is loosely informed by prior research into error analysis/coding (Jain 1973; Richards 1973; Dagneaux et al. 1996 & 1998) but represents errors as empty XML elements with 10 main error type attributes, agreement (AGR), aspect (ASP), coherence (COH), disfluency (DISFLU), pro-drop (DROP), lexis (LEX), number (NUM), phonology (PHON), specification (SPEC), structure (STRUCT), followed by a description attribute and an optional idiomatic ‘correction’. However, unlike in prior schemes, higher emphasis is placed on identifying features of textual/temporal coherence and the effects of various types of discourse marking.

The results so far indicate that the combination of pragmatics-related and error annotation not only makes it possible to identify the most common error types produced by Chinese learners (cf. Chuang & Nesi 2008) easily, but also that the majority of errors produced by the learners affect the coherence of the interaction through various features such as the use of constructions involving hypothetical instead of realis forms, ‘pseudo-anaphoric’ references, or phatic or redundant logical connectors.

Another highly prominent feature is the number of discourse markers or response-signals that are used (em)phatically, i.e. without really contributing to the interaction at all.

References
This paper presents the results of a study investigating the use of lexical and non-lexical fillers (e.g. well and eh, respectively) – or “words and phrases used to fill pauses, cover for hesitations, gain time, and provide smooth transformations in breakdowns” (Dörnyei & Scott, 1997, Table 1) – with the aim of exploring interlanguage fluency variations and the potential for transfer of fluency behavior from native language (NL1) Norwegian to interlanguage (IL) English. There is a scarcity of studies of interlanguage fluency that take into account the learners’ behavior in their own native language (but see e.g. Aas & Rørvik, 2017; De Jong et al, 2015; Rose, 2015). By mapping out speakers’ individual fluency styles (Fillmore, 1979), it may be possible to account for a portion of the “fluency gaps”
(Segalowitz, 2010) typically found in comparisons of fluency variables in IL and target language speech data (e.g. Belz et al., 2017; Götz, 2013). As De Jong et al. (2015) concludes: “it would be futile for an L2 speaker to strive for using very few filled pauses in his L2 when he tends to be an “uhm”-er in his L1” (p. 239).

The study seeks answers to the following research questions:

1) is the IL speakers’ use of fillers reflected in their NL1 performance, in terms of equivalent frequency, forms and/or functions?

2) is the degree of variation in the NL1 data comparable to the variation found in the IL data?

3) how “nativelike” in terms of frequency and use are the preferred fillers in the IL data?

The material was taken from the (forthcoming) Norwegian component of the Louvain Database of spoken English Interlanguage (LINDSEI) (Gilquin et al., 2010), and from six comparable interviews with the same speakers in their native Norwegian. All fillers in the data were manually identified and categorized. To explore the final research question, additional material from the LOCNEC corpus of British English interviews (ibid.) was used.

The preliminary results show considerable variation within the NL1 and IL language groups in terms of filler frequency. Three of the speakers have a higher filler frequency in the IL, but for the three remaining speakers there is only a minimal difference between the NL1 and IL. In both language varieties, a vast majority of the fillers used are non-lexical, and all but one speaker uses more lexical fillers in Norwegian than in English. The primary function and position of the fillers used in both varieties is that of hesitation or planning in the middle of a speaker turn. Other functions observed in both varieties include “a diminished intent to continue” (Tottie, 2015, p. 241), where the filler, while avoiding silence, is also seen as signaling to the interlocutor an intention to hand over the turn.

Our results indicate that speakers do transfer patterns of filler usage from their native language to the interlanguage. The study thus highlights the importance of including NL1 material in studies of interlanguage speech.

References


S85 Shelley Staples

**Corpus-based curriculum development in ESP: needs analysis, materials development, assessment and evaluation**

This presentation discusses the development of corpus-based curriculum for ESP, with a focus on two underresearched areas: health care communication and the use of corpus materials for pronunciation. Three aspects of corpus-based curriculum development are explored: corpus-based needs analysis; corpus-based materials development; and corpus-based assessment and evaluation (Flowerdew, 2012; Tono, 2011). These methods can be applied to developing any ESP course, and thus one outcome of the presentation will be a greater understanding of how corpora and corpus methods can be used for each of these three aspects of curriculum development. However, while there have been a number of corpus analyses in written ESP contexts, there are many fewer in spoken ESP contexts, and the field is particularly limited in its applications of corpus findings to pronunciation instruction as well as health care contexts.
First, there is a clear gap in the understanding of linguistic needs of second language health care providers, particularly for clinical interactions. Notable exceptions include Bosher and Smalkowski (2002), Cameron (1998), and Crawford and Candlin (2013), all of which use qualitative discourse analysis. In part 1, this presentation briefly reports on a quantitative corpus-based analysis of 104 nurse-patient interactions that was conducted to identify needs of nurses in clinical interactions, with a focus on the findings related to pronunciation (pitch range, tone choice, and prominence/sentence stress). Key differences were found between international and U.S. nurse discourse in the use of these features (along with lexico-grammatical, interactional, and fluency features and nonverbal behavior).

In part 2, the presentation provides an overview of a curriculum for a Pronunciation for Nurses course, developed in part from the corpus analysis, focusing on examples of corpus-based materials development from the corpus described above. In part 3, the presentation provides a discussion of the corpus-based assessment of participants’ progress and an evaluation of the Pronunciation for Nurses curriculum, including pre and post-tests, interviews with nurse participants, interviews with ESL teachers, and course evaluations. Nurse participants improved their use of all of the pronunciation features targeted in the corpus-based materials. In their evaluations and interviews, nurse participants particularly mentioned the usefulness of the focus on intonation, stress, pitch range, and sentence rhythm based on the findings from the corpus analysis. The main challenge identified from the instructor interviews is the need for additional teacher training. The presentation will end by discussing aspects of the curriculum to be expanded to incorporate other pronunciation, lexico-grammatical, interactional, and nonverbal features that were found from the corpus-based needs analysis. Participants will leave with ideas of how the principles discussed in the presentation can be applied to other ESP contexts and other aspects of language use.

References


S89 Jenny Kemp

Working towards a discipline-specific vocabulary core (DSVC) for postgraduate International Law

This paper reports on a study that investigates the vocabulary that postgraduate (LLM) International Law students need in order to cope with the reading requirements of their course. Law texts are widely regarded as challenging to read, both for first and second language speakers of English (Northcott 2009), and yet to date there has been no corpus study which covers the wide variety of text types encountered by students. Existing law corpora focus on primary legal documents, such as legislation (e.g. Williams 2007) or law reports (e.g. Marin and Rea 2011); secondary sources, such as journal articles, have largely been ignored. The DSVC International Law corpus has been compiled with the aim of addressing this knowledge gap and providing educators with a Discipline-Specific Vocabulary Core (DSVC) that will help students whose first language is not English. Currently standing at about two million words, the DSVC International Law corpus includes 5000-word text samples from 12 domains of International Law (e.g. Company Law). The texts fall into three broad communicative function categories: Prescriptive (e.g. treaties, directives), Descriptive (e.g. textbooks, commission reports) and Hybrid (e.g. law reports). In an attempt to ensure at the corpus-building stage that the corpus was representative of the texts that LLM students need to read, three important steps were incorporated. Firstly, a survey was carried out of the LLM modules at 21 UK institutions, which informed the domain structure. Secondly, module reading lists were consulted: most texts sampled in the corpus are specifically mentioned on reading lists, and some are common across institutions. The third method was to consult legal experts: for each domain, two law academics were asked to review the list of texts for their area of Law and to evaluate its representativeness; adjustments were then made to the corpus. The DSVC International Law corpus can therefore be considered typical of the texts that learners are exposed to.

The purpose of this research is not to develop a list of legal terminology - lexis already found in legal dictionaries and likely to be glossed by law lecturers and textbook writers - but to identify the frequent, pervasive vocabulary which is essential for understanding. This vocabulary can be much more discipline-specific than it initially appears (Hyland and Tse 2007). Preliminary analysis carried out on the corpus shows that there are indeed lexical items often considered 'general academic' yet which tend not only to be more frequent in law texts, but also to have highly specific usage and patterning, often helping to structure text. Examples are the collocation *parties/agree* and the
prepositional phrases *in (the) light of* and *in accordance with*. The implications for teaching are that a much greater emphasis should be placed on raising students' awareness of the discipline-specific behaviour of lexis they are likely to encounter frequently in their reading. This can be done through appropriately scaffolded tasks, examples of which will be illustrated.

**References**


---

**S95 Agnieszka Leńko-Szymańska**

**The English Vocabulary Profile and lexical profiles of texts written by tertiary-level students**

The English Profile project (EP, http://www.englishprofile.org/) has aimed to supplement the Common European Framework of Reference (CEFR, Council of Europe 2001) by providing detailed information on what language learners can use at each of the six CEFR levels. In this way it attempts to offer a clear benchmark for teaching, learning, and assessment of English language learners. Within the EP project, the Reference Level Descriptors take the form of lists of criterial features, defined by Hawkins and Filipovic (2012: 11) as “properties of learner English that are characteristic and indicative of L2 proficiency at each of the levels and that distinguish higher levels from lower levels.”

The English Vocabulary Profile (EVP, Capel 2010, 2012) – part of the English Profile project – has sought to establish which words and phrases are commonly known by English L2 students at the six CEFR levels (A1-C2). It takes the form of a learner-corpus-based database containing lists of words, their meanings, and frequent collocations and colligations, that learners of English typically know at each level. However, so far the extent to which the information in the EVP can be applied to sketch lexical profiles of texts written by EFL learners at different levels has not be extensively investigated (cf. Leńko-Szymańska 2015). This paper presents the results of a study, whose aim was to explore the
usefulness of the EVP for assessing (the lexical aspects of) written production of students of English at the tertiary level.

Thirty argumentative essays written by Polish, Austrian and Spanish students of English were drawn from the International Corpus of Learner English (Granger et al., 2003). Each word in every essay was assigned a CEFR level according to the information available in the EVP. Obvious lexical errors (an incorrect form, meaning, collocation and colligation) were also tagged. As a result of this procedure a lexical profile was created for every text, which consisted in the proportions of A1-C2 words and lexical errors. In addition, the essays were assigned the CEFR levels by three raters. The same raters assessed the vocabulary of these texts – both range and control – using relevant CEFR descriptors.

The result of one-way MANOVA points to statistically significant differences in the use of A1-C2 vocabulary by students at different levels. A high positive correlation was found between students’ use of C1 and C2 vocabulary and their marks for lexical range as well as between students’ errors and their marks for lexical control. Finally, a high positive correlation was found between the CEFR levels assigned by raters and the grouping of essays generated in a cluster analysis based on their lexical profiles.

The study constitutes but a first step in defining lexical profiles of learner texts at different proficiency levels. In a larger perspective, this paper contributes to the discussion of empirical validation of CEFR levels (Wisniewski, 2017).

References
Discipline-specific academic phraseology: corpus evidence and applications

Wordlists such as Coxhead’s (2000) New Academic Wordlist and, more recently, Gardner and Davies's (2014) New Academic Vocabulary List have dominated corpus-based methods for the selection of vocabulary for English for Academic Purposes (EAP) courses and materials. Although, in general, reaction to these lists has been positive, two principal criticisms have emerged: Firstly, as general- academic resources, these lists cannot reflect differences in vocabulary needs between academic disciplines. Secondly, as lists of single words, they cannot take phraseological concerns into account. In this context, the phraseological turn evidenced by development of academic formula lists, be they general (e.g. Simpson-Vlach & Ellis, 2010) or discipline-specific (e.g. Hsu, 2014), represents a positive development. However, these phrase-lists are often selected on the basis of variants of a frequency- based n-gram approach widely known as the lexicalbundle approach (Biber, Johansson, Leech, Conrad, & Finegan, 1999). On this approach, strings of text, usually of three or four words in length, occurring above a given frequency threshold are extracted from a corpus. This paper argues that the positive development represented by the phraseological turn in EAP vocabulary selection is limited by this distribution-based method of phrase extraction and, moreover, that the generalist approach to vocabulary selection in EAP obfuscates important differences in vocabulary meaning between academic disciplines.

These beliefs are examined via an experiment designed to test the collocational behaviour of verbs in a bespoke eight-million-word corpus comprising academic research articles from the disciplines of history, microbiology, and management studies. The experiment employs a combination of Corpus Pattern Analysis (Hanks, 2004, 2013), a lexicographical technique for mapping meaning onto text, and statistical techniques to compare the syntagmatic patterns of frequently occurring verbs across disciplines. The results demonstrate that, in many cases, the prototypical meaning of a given verb varies according to the discipline in which it is found. These differences are manifested in several ways including: Variations in the prototypical semantic type of verb collocates, variations in semantic prosody, and variations in the syntactic arrangement of verbs and their collocates. Going beyond the lexical level, the findings also provide further insight into differences in intertextuality, reporting practices, and move structure across disciplines, as well as shedding light on some salient uses of metaphor. More generally the results demonstrate that, in order to fully appreciate these differences, a means of vocabulary selection which accounts for both syntactic and semantic concerns is necessary.

In addition to evaluating current approaches to the selection of vocabulary for EAP, and providing insights into differences in language use between academic disciplines, the results also have potential practical applications. A proposal for exploiting the data collected in the experimental procedure in a mobile EAP lexicographical resource is advanced. It is envisaged that such a resource would promote
student interest in investigating corpora of academic texts and encourage reflection on the academic writing process.

References


---

**S99 John Williams**

`Definitely [sic] worth it': engaging undergraduate corpus linguistics students in real research

This paper was inspired by the author’s experience of teaching corpus-based lexicology on a final-year undergraduate unit entitled Researching English Vocabulary. As this title implies, there is an expectation that students should carry out actual lexicological research.

Not surprisingly, this expectation presents some challenges. The unit is taught over a single term with only two contact hours per week: a lecture and a hands-on session in a computer lab. Most of the students have had little practice with corpora up until this point. Thus, it is all too easy for the lab classes to descend into ‘how to’ sessions, with insufficient time left to discuss the linguistics behind the data, the ‘why’ questions (cf. Baker, 2009, p.75), or the process of doing research. This can lead to student dissatisfaction and disappointing feedback.

Thus, the unit lecturers decided to make some important changes in 2017-18. One was to align more of the lab tasks with typical research activities, eg. early in the unit, students were asked to replicate Aitchison’s (2004, pp.128-9) research into the word ‘disaster’ on the basis of concordances sampled from larger, more up-to-date corpora.

The main focus of this paper is on how this approach was applied towards the end of the unit, when students were choosing topics for their assessed research reports. Students have a tendency to choose unsuitable topics, so the idea was to guide them towards a predetermined programme of research
(though with still the option to choose their own topic). The starting point was a list produced by Oxford Dictionaries of the ten most common ‘misspellings’ in English (Oxford Dictionaries, n.d.). The number one item on this list, namely ‘publically’, has been the subject of a longstanding interest on the part of the present author (Williams, 2012ab). In the lecture, a methodology was presented for comparing the collocational distribution of ‘publicly’ and ‘publically’, inspired by the ‘concordance-keywords’ method (Taylor, 2010). The analysis revealed a significant finding, namely that although, in the enTenTen13 corpus (Sketch Engine, 2013), ‘available’ was the most frequent lexical collocate of the standard spelling ‘publicly’, it was a keyword (Anthony, 2006), ie proportionally even more frequent, in the concordance-corpus for ‘publically’ when compared against the equivalent corpus for ‘publicly’. This finding generated a hypothesis, namely:

"Non-standard spellings of lexical words will be associated disproportionately with the most frequent lexical collocate of the corresponding standard form."

In the lab sessions, the students were able to test the hypothesis by applying the methodology to the other ‘misspellings’. Although the hypothesis was not confirmed in many of the cases, interesting findings were reported for ‘definitely’, ‘government’, and ‘receive’. In a short evaluation activity at the end of the session, students expressed great satisfaction at being able to participate in ‘real’ research, which they then had the option to write up for their assessment - motivated partly by the promise to be mentioned as joint authors if their findings were eventually published in a journal article.

References

---

**S101 Alvin Cheng-Hsien Chen**

*Beyond two-word sequences: assessing phraseological development in L2 texts using forward transitional probabilities*

This study evaluates L2 phraseological development using a directional association measure, Delta P (DP), and addresses two questions: (1) whether learners develop their phraseological competence as their proficiency level grows? (2) how is this development mediated by the directionality of associative learning and length of phraseology? DP effectively measures the native-like intuition in word selection based on forward- or backward-directed temporal relations. We assessed the formulaicity of bundles from two-word to five-word sequences in L2 texts collected in the International Corpus Network of Asian Learners of English V2.0 (ICNALE) (Ishikawa, 2013). In ICNALE, the L2 proficiency level was defined through external standardized tests and mapped to the categorical labels in the Common European Framework of Reference for Languages (CEFR).

We assigned each L2 text eight scores of DPs, measuring the average forward and backward DPs for word sequences from bigrams to 5grams. DPs of word sequences were estimated based on two representative native corpora, i.e., the British National Corpus and the Corpus of Contemporary American English. Mixed-design multilevel linear models were used to analyse the variation of DP with learner proficiency as a between-subject factor and directionality of association and length as within-subject factors. Our analysis shows that formulaicity increases with learner proficiency. When the directionality is considered, learners tend to show a steadier increase in forward phraseological competence as their proficiency grows; backward phraseological competence, however, develops more evidently in more advanced levels. This difference may be reduced when longer bundles were considered. These findings provide a more holistic understanding of L2 phraseological development.

---

**S107 Geraldine Mark and Pascual Pérez-Paredes**

*Examining high frequency adverbs in learner and native speaker language: some implications for spoken EFL learning and teaching*

Most research in learner adverb use has shown evidence of both overuse and underuse of adverbs along with a lack of register awareness, in particular in written genres. Studies that examine adverb
use in spoken language and those that consider pragmatic aspects are not abundant (excepting Pérez-Paredes, 2010; Aijmer, 2011; Gablasova et al., 2017). Similarly, studies comparing the use of adverbs across different tasks and L1 speakers in the same task or research condition are rare. This paper describes the findings of a study comparing high frequency adverbs in English spoken communication across four L1 datasets. The study draws on the Chinese, Spanish and German components of the LINDSEI database (Gilquin, DeCock and Granger 2010) and the first language English data in the LOCNEC extended corpus (Aguado et al., 2011). Our research is based on comparative analysis that draws on the following principled decisions: (1) Participants across all four datasets completed exactly the same three spoken tasks; (2) taking a combined quantitative and qualitative approach, the study examines the contexts of use and underlying pragmatic functions of adverbs across each different task, adopting thus a task-perspective; (3) our choice of adverbs was based on the frequency of use. We’ll base our presentation on the analysis of the adverbs actually, just, maybe, obviously, perhaps, probably, really, and well. In this paper we will discuss the implications of our study for the teaching of adverbs in EFL classrooms. These implications can be grouped in four areas: (a) the role of tasks; (b) the role of L1; (c) language patterning, including collocation; (d) pragmatic considerations.

References
S109  Rolf Kreyer

Collocations in learner English – the true-longitudinal perspective

Learner-corpus linguistic research has made a lot of progress over the last few years regarding the analysis of collocations in learner English. However, research so far has focused mainly on fairly advanced learners (usually students of English language and literature) and true-longitudinal studies on the development of phraseological competence are few and far between. At the same time a growing body of quasi-longitudinal research on collocations across different proficiency levels documents the interest in this kind of research. The present study wants to contribute to this line of research by analysing the use of collocations in true-longitudinal data, namely the Marburg corpus of Intermediate Learner English (MILE; Kreyer 2015), which contains about 750,000 words of written material of German learners of English as they progress from grade 9 to grade 12 in a German secondary school.

In particular, the paper looks at verb-noun collocations with four highly frequent ‘delexicalized’ or ‘light’ verbs, namely MAKE (n=2216), TAKE (n=939), GIVE (n=978) and DO (n=4041). ‘Collocation’, here, is understood in the phraseological rather than the statistical sense, i.e. as lexical combinations that can be distinguished from idioms on the one side and free combinations on the other (see, among others, Nesselhauf 2005) and “are characterized by restricted co-occurrence of elements and relative transparency of meaning” (Laufer & Waldman 2011: 648).

For each of the four verbs above and their noun-collocates the study explores the following:

1. number/proportion and types of well-formed as well as ill-formed collocations, e.g. make a contribution or *make an experience across grades and across individual pupils
2. the correct collocate, e.g. have an experience vs. *make an experience, for each ill-formed collocation
3. possible L1 influences, e.g. *make homework for German Hausaufgaben machen, for each ill-formed collocation
4. possible L2 influences, e.g. *take advantages vs. take advantage, for each ill-formed collocation
5. persistence of/changes in deviant collocations across grades and individual pupils

On the basis of these analyses the present study wants to shed light on three aspects of collocations in learner language. Firstly, what can we learn about the development of collocational competence in learners of English at an intermediate level of proficiency? In particular, to what extent do learners follow general patterns of acquisition (if at all), to what extent do they show individual learning paths? Secondly, what error-analytical insights can be gained from the data at hand? More specifically, what can we say about the etiology of deviant collocations? When do L1-contrasts lead to collocational errors, when do they not? What exactly is the role of the L2 in this respect? Can we
account for particularly persistent erroneous collocations? Finally, given the questions above, what are implications for language teaching?

References

S110 Mariko Abe, Yusuke Kondo, Yuichiro Kobayashi, Akira Murakami and Yasuhiro Fujiwara

Initial findings from a longitudinal learner corpus: a year-long development of L2 speaking performance

Despite a dramatic increase in learner corpus studies in the last two decades, most of the studies target written production by advanced learners (Paquot & Plonsky, 2017) and are cross-sectional or pseudo longitudinal in design (Meunier, 2015). The emphasis on written production by advanced learners is a problem because we cannot gain a comprehensive picture of language development by studying the language use of advanced learners alone and because spoken performance is considered to reflect learners’ implicit knowledge better than their written production. The predominance of cross-sectional research has also been an important limitation in learner corpus research and second language acquisition in light of repeated calls for the investigation of longitudinal development in individual learners (e.g., Larsen-Freeman & Cameron, 2008; Housen, 2002). Our research project fills the gaps by compiling and analysing a longitudinal corpus of secondary school students’ L2 speech samples. Based on the data collected over one year, our talk presents the initial findings of the study investigating the longitudinal L2 speaking development.

The data exploited in the study include speech samples by 120 upper secondary school students collected three times over one year from the same group of learners. Their English proficiency was estimated as 0 to B1.1 level of the Common European Framework of Reference (CEFR) benchmark. For data collection, we employed Telephone Standard Speaking Test (TSST), a monologue speaking test carefully designed to assess learners’ free-spoken production in terms of vocabulary, grammar, and pronunciation. Three certified raters holistically scored the speech samples based on various criteria such as function-based ability, sentence structure, accuracy, and content. In the three batches
of data collected over one year, the holistic scores spanned across five oral proficiency levels out of an eight-point scale and the learners’ overall score tended to rise across the year.

In order to identify the features characterizing L2 speaking development, we examined various speech and textual features that are believed to correlate with spoken performance of L2 learners. We found that (i) the average length of the spoken text best distinguished lower and higher proficiency learners, (ii) the mean length of utterances (MLU) increased from 11.23 at Level 2 (0 in the CEFR) to 26.49 at Level 6 (B1.1) as proficiency rose, (iii) fillers are used less by higher proficiency learners, (iv) oral proficiency is also discriminated by word bigrams representing such syntactic patterns as subject + be-verb, verb + personal pronoun, verb + that-clause, preposition + personal pronoun, determiner + noun (cf. Crossley et al., 2011).

These findings based on the longitudinal corpus indicate the textual and spoken features that characterize learners at particular developmental stages and can be applied to language teaching in the classroom in the future.

References

---

**S111 Lynne Flowerdew**

*Discourse-based perspectives on the application of ESP learner corpora to writing pedagogy*

ESP learner corpus research of writing remained an under-developed field until quite recently (see Author 2015), and it is only in the last few years that learner corpora of specific written genres have been integrated directly into classroom pedagogy. The purpose of this presentation is twofold: to take
stock of recent advances in ESP learner corpus research and pedagogy for writing and to propose some avenues for further enquiry. Discourse-based perspectives will be the focus of attention.

A few accounts describe how learner corpora have been directly integrated into various online teaching tools. Two of these take a Swalesian genre-based approach with a focus on move structures and their linguistic features, thereby combining the top-down approach of genre analysis and the more bottom-up approach associated with corpus enquiries. For example, Birch-Bécaas and Cooke (2012) describe a DDL program which uses a small corpus of NNS drafts of research article introductions, together with their revised versions processed pedagogically with annotations to highlight key discourse moves and their linguistic realisations. Another learner-corpus initiative which takes an explicitly Swalesian genre-based approach is reported in Nordrum and Eriksson (2015). Nordrum and Eriksson annotated with rhetorical moves a specialised corpus of data commentaries drawn from published research articles and master’s theses, represented by different levels of quality, with a view to having students engage in formative self-assessment practices facilitated by DDL activities. Other pedagogic initiatives addressing discourse-level concerns include those by Charles (2016) and Wärmsby et al. (2016). Charles’s doctoral students compiled corpora of their own writing and corpora of research articles in their own field using the AntConc tools. The Concordance Plot allowed students to track content, ideas and terminology within a single chapter and compare usage across chapters of their theses. An ongoing large-scale learner corpus project is that reported in Wärmsby et al. (2016), tasked with capturing writing as a process. To this end, several drafts, student self-reflective papers, and teacher and peer feedback addressing higher order concerns such as linkage of points across thesis chapters are planned for inclusion.

A few other accounts, while primarily research-based, make valuable suggestions for pedagogy. One such study is that by Miller and Pessoa (2018) on the analysis of junior and senior information systems project reports using Kaufer’s theoretical framework of genre incorporated into an NLP tool (Docuscope). Parkinson’s (2017) study using corpus methods supplemented with interviews also investigates learner writing at different levels of acculturation into a vocational genre (builder’s diaries produced by novice carpentry students and those written by more experienced students). Content analyses are employed by Sing (2016, 2017) to investigate how advanced business students ‘technicalise’ at a functional level in a corpus of seminar papers.

Some suggestions for further expansion of the field are offered, e.g. employing more multi-method approaches for text analysis (for example, combining the Swalesian approach to genre with Kaufer’s theoretical perspective on genre).

References


---

**S113 Awatif Alruwaili**

**What do in-service teachers know and think about corpora after short training?**

This paper aims to bridge the gap between corpus linguists’ enthusiasm about the language pedagogical potential of corpus linguistics (CL) on one hand, and the reality of English language teaching in a foreign context on the other. Several authors have also confirmed the key role that teachers play in applying the uses of corpora in language teaching (e.g., Frankenberg-Garcia, 2012; Leńko-Szymańska & Boulton, 2015). The present study sought to widen the existing perspective on
using corpora in language classrooms given previous research’s promising results on the importance of investigating teachers’ attitudes towards the uses of corpora. This study is particularly interested in ways to transform classrooms into learning environments that truly facilitate the use of corpus-based approaches and methods for learning English in a foreign context.

This paper aims to explore in-service teachers’ dispositions towards the use of corpora in language classrooms. To this end, the first phase in this research involved designing a training course to show language instructors possible ways of using corpora in the classroom. The course units were always presented in the context of language learning activities, rather than covering theoretical and pure linguistics analyses. The training course consisted of two sessions, each of which ran for one hour and 30 minutes. The course content consisted of three units: teaching about corpora, exploiting corpora to teach language and teaching to exploit corpora.

The design used in this study is based on mixed-methods research, involving an exploratory angle. Data were gathered through questionnaires, and post-course semi-structured interviews. This paper evaluates 56 in-service teachers’ dispositions towards the use of corpora in language classrooms. The analysis of data revealed that the questionnaire produced an excellent measure of internal consistency, as the reliability test of teachers’ attitudes towards CL in the classroom results in \( \alpha = .90 \). The teachers displayed moderately positive attitude towards the use of corpora in the classroom, which is considered a good start to integrating CL in the classroom. This study identified several factors that may facilitate or hinder the use of corpora in the classroom, as discussed from the teachers’ perspectives; these factors can be categorised into themes related to 1) the bureaucratic system (curriculum, policymakers, availability of resources and facilities, number of students per class and teaching load); 2) the training needs (existence of an expert or supporter, computer literacy and training on linguistic analysis); 3) the teachers themselves (teachers’ views about their roles, motivation, and willingness to use corpora); and 4) the learners themselves (learner level, motivation). The abovementioned factors help to clarify why there is a lack of corpora use, and the discussion of these factors provides deeper investigation into the classroom and understanding of why corpora is not fully integrated into the EFL classroom. This barrier can be eliminated by more proactive efforts, such as more training, free resources and ready-made teaching materials.

Reference


Exploiting corpora to provide guidance for academic writing: the BAWE hyperlink project

Although the idea that corpora offer great potential for language learning has existed at least since Johns (1986), this potential has yet to be fully realised in mainstream teaching. The lack of take-up is as true for academic writing as it is in other areas of language learning.

One obstacle to the pedagogic use of corpora may be the apparent complexity of corpus interfaces, which discourage novice users. Moreover, even where relatively user-friendly resources exist, such as SkELL, Just the Word, or Word and Phrase, the corpora from which these tools take their data may not be appropriate for the learners concerned. A further barrier is that learners may not know what it is that they want or need to discover from the corpus data, while teachers lack the time to help them with their individual requirements.

This paper will report on an initiative to overcome these challenges to realising the potential of corpora in pedagogy. Inspired by Johns’ DDL approach (1991a, b), the initiative offers teachers a time-saving way to help their students engage with corpus data relevant to their particular writing needs.

The process we have adopted builds on a technique introduced by Nesi, Gardner and Kightley (2015) for the British Council Writing for a Purpose project. We first identify recurrent problems affecting intelligibility and/or communicative force in the electronically submitted writing of undergraduate and master’s students, for example collocation errors. Concordances are then retrieved from the British Academic Written English (BAWE) corpus, using the Sketch Engine open access interface. These concordances are refined to illustrate typical uses of the relevant item. A hyperlink is then inserted into the student’s work so that they can review these typical uses and correct (or improve) their own work.

As well as being within the spirit of DDL, this approach has the advantage of using a corpus of the type of writing the students themselves aspire to produce, in terms of genre, discipline and level of study (Durrant 2013), rather than mega-corpora of ‘general English’ (SkELL) or academic corpora of research articles (Word and Phrase). A second advantage is that the students do not have to deal with daunting corpus interfaces; we provide links to pre-filtered concordance lines. A third advantage is that, because these links are permanent, they can be collected in a database of recurrent errors for re-use by other colleagues and writing tutors, anywhere in the world.

This presentation will report on the creation of this database (currently under construction at http://bit.ly/BAWEhyperlink). We will reflect on the types of errors we have found and how we have created and classified the resulting resources, and the challenges surrounding the need to achieve a
balance between re-usability and usefulness for the individual learner. We will also present feedback we have collected from users – both students and teachers.

References


S115 Katarina Lazic and Maja Milicevic Petrovic

Creating pedagogically useful lists of biotechnical academic formulas

Corpus-derived lexical bundles such as *allows us to* or *for each of the* have recently been grouped by a number of authors in lists of pedagogically useful English academic formulas (see Simpson-Vlach & Ellis, 2010, for a general list, Fox & Tigchelaar, 2015, for engineering). Typical criteria applied in the creation of such lists are that the formulas must be highly frequent in a relevant corpus of native academic production, significantly more frequent in academic than in non-academic discourse, and present in a wide range of genres and/or sub-disciplines.

The present paper reports on a study with a similar goal and a partly modified methodology. Our focus is on lists of frequent four- to six-word lexical bundles in biotechnical writing in English, with pedagogical usefulness for L2 learners in view. Longer bundles are chosen as they often encompass shorter bundles and tend to be more difficult for non-native speakers to identify. The distinctive characteristics of our approach are that it is L1-specific, under the assumption that L1 formulas can influence the choice of L2 formulas, and that it involves multiple lists, corresponding to multiple teaching objectives.

Three comparable corpora of biotechnical research articles were used in the process of list creation: CoBNEA (*Corpus of Biotechnical Native English Articles, 1,525,469* words; divided into sub-corpora from four sub-disciplines – forestry, landscape architecture, wood processing, ecological engineering), CoBNONEA (*Corpus of Biotechnical Non-Native [L1 Serbian] English Articles, 157,179* words), and CoBSA (*Corpus of Serbian Biotechnical Articles, 126,275* words); CoBNEA was the source of
formulas to be included in pedagogical lists, while CoBNONEA and CoBSA were used for comparison purposes. The formulas were extracted using AntConc (Anthony, 2011) and manually classified according to structural and functional criteria, following Biber, Johansson, Leech, Conrad, & Finegan (1999) and Hyland (2008), respectively.

Two main lists were generated, one of 39 formulas exclusive to the native writing, and another one of 29 formulas without translation equivalents among the formulas used in Serbian; both lists highlight the sequences that seem particularly problematic for L1 Serbian writers. Additional lists were created to target formulas of specific structural and functional types, again with a focus on those found to be underused in non-native writing, e.g. a structural list of passive formulas (such as been shown to be), and a functional list of text-oriented formulas (such as as part of the). A more general list of 34 formulas shared by at least three biotechnical sub-disciplines was also compiled, as well as a unified list of 121 formulas, produced by merging the smaller lists (overlapping sequences were combined into single formulas, e.g. of this study was to and of this study was).

We show how the obtained lists can contribute to biotechnical academic writing instruction in the L1 Serbian context, pointing also at how they can be used more generally in teaching English for academic and specific purposes, while the method used for creating the lists can easily be replicated to other disciplines and languages. We conclude with suggestions for classroom-based activities that exploit the created lists.

References

---

S116 Reka R. Jablonkai and Neva Cebron

*Autonomous language learning with the help of corpora: the case of a corpus-based ESP course*

Several studies proposed ways to directly apply corpora in language teaching and there is also a growing body of literature analysing the effectiveness of such a direct approach of corpus applications...
in language teaching. Previous studies found that corpus use enhanced learners’ language awareness, it complemented existing reference works for learners, provided native speaker insights for teachers of languages other than their L1, and was applied successfully for error correction in L2 writing (Boulton & Cobb, 2017; Cobb & Boulton, 2015; Johns, 2002; Römer, 2011; Flowerdew, 2013; Gaskell & Cobb, 2004; Gavioli, 2005; Yoon, 2008; Yoon & Hirvela, 2004). Most research into direct application of corpora in language learning has focused on EAP contexts and used corpora mainly for writing in L2. However, so far there have been only a handful of studies that investigate how corpora are can be used as tools to facilitate learner autonomy (Charles, 2012; 2014; Lee & Swales, 2006). The present study extends earlier research by focusing on corpus use during and after a subject specific ESP course for autonomous language learning in general and vocabulary learning in particular. The study aimed to answer the following research questions:

RQ1: How do learners perceive corpus use in general?
RQ2: Do learners use their DIY corpora autonomously after the course?
RQ3: Do learners use online corpora and corpus analysis tools autonomously after the course?
RQ4: What do learners use online corpora and corpus analysis tools for autonomously?

The present study applied a mixed-method approach with content analysis of written initial feedback, analysis of interview and questionnaire data. Participants of the study were a group of 14 students who took part in a 30-hour corpus-based EU English ESP course focusing on topics relating to the European Union as part of their studies at a Slovenian university. At the end of the initial 30-hour course, students were asked to provide written feedback. Six months later, in order to establish autonomous use of corpora among participants, a questionnaire was administered, and selected students participated in a focus-group interview.

Overall, findings show that students found corpora very useful. Their anonymous feedback after the course and their comments in the focus group suggest that they considered corpora helpful not only for the purposes of the corpus-based ESP course but saw long-term potential in corpus use for language learning. Similarly, the results of the questionnaire indicated that students’ overall evaluation of corpora was positive. Regarding the autonomous use of DIY corpora and available corpora, findings reveal that only a few students used their DIY corpus autonomously, but almost all of them used available corpora on their own. Students predominantly consulted both types of corpora to look up collocations and to check the exact usage of words. Based on the results tentative principles of design and implementation of corpus-based ESP courses are proposed for future syllabus design with direct and autonomous corpus use. Preliminary results of a follow-up study with another group of students being carried out at the moment will be presented as well.

References
S120 Kiyomi Chujo, Atsushi Mizumoto, Yuichiro Kobayashi, Akira Hamada, Kathryn Oghigian

Evaluating the appropriateness of the Sentence Corpus of Remedial English to benefit beginner level EFL students

Recent meta-analyses in data-driven learning (DDL), have suggested that the use of corpora can be effective and efficient (Cobb & Boulton, 2015: 494; see also Mizumoto & Chujo, 2015); however, Chujo et al. (2007) found that the majority of e-texts available on the web were rated at an advanced level. Japan’s 2015 school survey results found that most of Japanese learners were at A1 to A2 levels.
on the Common European Framework of Reference for Languages. Although the use of authentic materials is debated, we agree that “[t]hroughout this debate the fact that learners should aim at a good level of linguistic authenticity in their communicative competence is much more consensual” (Meunier & Reppen, 2015: 500). To address these issues, we developed an educationally modified bilingual corpus, the Sentence Corpus of Remedial English (SCoRE).

SCoRE was first introduced in Multiple Affordances of Language Corpora for Data-driven Learning (Leńko-Szymańska & Boulton, 2015), and we have just completed the fourth phase of development. It currently contains twenty-two categorized grammar items with 10,460 level-specific, semi-authentic sentences (101,528 words) written to satisfy particular pedagogical considerations, i.e., appropriateness and usability, and fair use for copyright issues. Each English example sentence has a Japanese translation. Users can search sentences by grammar item and by proficiency level (beginner, intermediate or advanced). It includes a grammar/keyword browser, concordancer, quiz function and downloadable data for copyright-free materials creation. SCoRE has been field tested for two years in a low proficiency Japanese EFL university course aimed at remediating previously identified grammar issues. Participants showed improvement in proficiency for the targeted grammar items in case studies, and reported that the corpus was useful, easy, and enjoyable to use. Those field tests demonstrate that beginner level EFL learners can also take advantage of the benefits of DDL when using an educationally modified bilingual corpus.

The purpose of the current study is to evaluate SCoRE objectively for its “appropriateness” as an educational corpus. To do so, evaluation focused on (1) the difficulty level of the sentences as concordance lines; (2) the coverage of grammatical words and phrases indexed in three ELT grammar books; and (3) the semantic content distribution.

Although this study is only the first of a series of corpora evaluation studies for appropriateness and usability, preliminary indications are that SCoRE is appropriate and 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


---

**S127 Yukio Tono**

*CEFR-Jx27: Developing corpus- and CEFR-based pedagogical resources and e-learning systems for 27 languages*

This paper is an interim report on the project called the CEFR-Jx27, whose purpose is to develop corpus-based pedagogical resources for 27 European as well as Asian languages. The materials are all aligned to a set of can-do descriptors developed for the CEFR-J, and all the vocabulary and phrase lists were created in line with the original English materials. Thus the materials are bi-directional, which means one can learn any set of foreign language in a quite similar way using the common CEFR-J based dataset.

The unique feature of this project lies in its methodological claims. We argue that this project can provide a systematic method of creating pedagogical language resources for a given language by using a series of CEFR-based linguistic resources originally developed for English. This paper will describe (1) the procedure of the construction of the English resources, (2) the conversion of English resources to each language using NLP techniques and (3) the development of e-learning environment for learning and teaching these languages.

The original framework used for this project is called the CEFR-J (Negishi, et al. 2016). It is a modified version of the Common European Framework of Reference for Languages (CEFR), with total 12 CEFR sub-levels validated statistically following the same method as the original study by Brian North. After the construction of the CEFR-J, a Reference Level Description project called the CEFR-J Profile project was launched. In the project, INPUT corpora (ELT textbook corpora) and OUTPUT corpora (Japanese EFL learner corpora) with CEFR levels were compiled and machine learning techniques were used to identify features which contribute to the discrimination of different CEFR levels. In this way, the CEFR-J Wordlist, Grammar Profile and Text Profile were created. Also
in order to utilize can-do descriptors for actual task design, possible expressions and vocabulary were selected from ELT textbook corpora and arranged as the CEFR-J Phrase list.

The CEFR-J English Wordlist (A1 to B2 levels) and Phrase list (A1-A2 levels) were used as the base lists and then they were converted into 26 other languages using Google Translate. Native speaker instructors went through the translation and evaluated the quality of translations. At the same time, the frequency lists of words from web corpora were provided to supplement the Wordlist. In this way, most grammatical words which cannot be directly translated from English were found in the corpus-based frequency list and supplementary lists were prepared for each language. Overall, this method was much quicker than translating each word manually. By the end of 2017, we finished preparing the Wordlists for most of the 27 languages, and a half of the Phrase lists were ready for the next implementation stage.

The three types of e-learning systems were newly created for implementing these resources into actual teaching: (1) flash-card vocabulary building smartphone apps, (2) translation tutoring application with error detection functions, and (3) web-based spoken and written learner corpus building sites.

References

S129  Hildegunn Dirdal and Stephanie H. G. Wold

L2 development of -ing clauses

The English morpheme -ing features in many studies of progressive marking (e.g. Housen, 2002; Robison, 1990, 1995; Rocca, 2007; Rohde, 1996). Other uses of -ing have received less attention. The present study focuses on the development of -ing clauses in L2 English by L1 Norwegian learners, building on recent work by Wold (2017), who finds minimal use of -ing outside finite contexts among 11-year-olds and widespread use among 15-year-olds. A closer look at the use of -ing according to proficiency level reveals a path from no use outside finite contexts (A1 level), via a few uses after aspectual verbs (A2 level 11-year-olds), increased use after perception verbs and in adverbial functions (A2 level 15-year-olds), and finally an equal proportion of progressive and non-progressive uses (B level 15-year-olds). This hints at a spread from more to less progressive-like contexts, but more knowledge is needed about the longitudinal development.

We use data from TRAWL: a longitudinal corpus of written texts from Norwegian school children, currently under construction. Data collection has started in years 5–13, and most pupils have currently been followed for one and a half years. However, a subset has been followed over three years (years
8–10). We use a combination of longitudinal case study of some of these pupils and a quasi-longitudinal study of texts from all years to answer the following questions:

- Are there common stages in the development and use of -ing clauses?
- Are there individual differences in the paths taken?
- Can L1 influence be detected?

Following a usage-based and cognitive perspective, multiple factors are assumed to influence L2 learning, making distinct individual paths possible. However, similar educational practices and properties of the English language itself, may lead to Norwegian learners showing evidence of common stages, possibly also based on their particular L1 background. Norwegian verbal nouns in -ing cannot project clauses, and present participle (V-ende/-ande) clauses are severely restricted compared to English -ing clauses (Dirdal 2017).

Preliminary longitudinal results from four students seem to confirm Wold’s findings that -ing clauses appear after the progressive structure is mastered, i.e. when the auxiliary BE is no longer omitted. Intermediate stages are less clear: once -ing clauses appear, they are used in a variety of syntactic functions, but mainly selected by particular verbs or prepositions. There are very few cases of preposition-less adjuncts, and only in year 10 do we find examples of -ing clauses used as subjects. A gradual increase in the use of -ing clauses is found only for the two students who use the lowest numbers to begin with. Individual differences seem to be quantitative rather than qualitative, but may also have to do with timing relative to the mastery of other linguistic features. At no stage do the learners seem to restrict their use of -ing clauses in a way that mirrors the syntactic functions, internal syntax and verbs found in Norwegian present participle clauses, but there may be an underuse of -ing clauses that could be related to L1 influence.

References


Recent years have seen a rapid growth in the number of programmes taught through the medium of English in countries where English is not a first language for the majority of the population. The popularity of English-medium instruction (EMI) is due in large part to its enable role in internationalisation. A second driver for EMI is the belief that it will produce incidental language learning outcomes; that is, by being exposed to English, students will become more proficient users of English. At the same time, to have good pre-conditions for success with academic study through the medium of a second language, students must have a level of proficiency which enables them to engage in receptive tasks such as listening to lectures and reading textbooks to a satisfactory level of comprehension, as well as performing tasks which call on productive skills, such as giving presentations and producing various kinds of texts for assessment purposes. Good skills in English are therefore an enabler of EMI, while improved skills are an expected outcome.

One important area of academic literacy is vocabulary, and a good knowledge of academic vocabulary in turn underpins the ability to read, write, speak and listen at university. In other words, both receptive and productive vocabulary skills are necessary. This paper will report the results of an investigation into the receptive and productive academic vocabulary knowledge of students in the EMI environment. The context in question is a prestigious technological university in Sweden at which the ordinary language of instruction at the master's level is English. Tests of English academic vocabulary were administered to approximately 500 students to determine their knowledge of academic vocabulary at various frequencies. In order to tap into productive academic vocabulary knowledge, a corpus was compiled of academic writing in English produced for assessment purposes by similar students at the same university. The corpus consists of 80 texts and approximately 720,000 words. Both the writers and the test-takers were second language users of English, some of whom were local students with Swedish as their first language, while others were international students from a variety of language backgrounds. This diversity reflects the composition of the EMI classroom in Sweden and thus has ecological validity. The corpus was profiled for academic vocabulary. The findings were then compared with the results on the test of receptive vocabulary, to establish the extent to which the students’ receptive and productive vocabularies differed.
The results reveal distinct patterns when the students’ receptive and productive vocabularies are compared. Pedagogical implications will be discussed, and include the suggestion that to the extent that vocabulary is the subject of explicit instruction, teaching could usefully be aimed at encouraging productive use of a broader range of lexis.

---

**S131  Katerina Florou**

*Parallel-learner corpora or another way to confirm teachers’ intuitions*

It is commonly accepted among linguists that as soon as big corpora became accessible the use of empirical data in language research has shown a remarkable increase. The research hypotheses are not restricted to the native linguistic code using monolingual general/native corpora but also have been extended to learner and parallel corpora. This study reflects the intention to verify some intuitions based on daily work in the foreign language classroom using a mixture of parallel and learner corpora. Experienced L2 teachers have been able to predict with some accuracy the rhetorical strategies transferred from L1 to L2 in their students’ writing. The interference may be more apparent to a teacher who shares the same L1 language with his/her student (Pereda et al. 2014). In this case, the Greek language is the mother tongue of both the teacher and the students in a classroom that the Italian language is being taught. In order to investigate this type of interference, two parallel corpora have been created. The first is in Greek and is composed of Greek students’ essays in specific topics (narration and academic speech), and the second is in the Italian language composed by the same students who have written/translated essays of the same size and topic in the foreign language.

The research aim, in this case, is the study of the transfer that has been observed in the use of auxiliary verbs. Auxiliary verbs in Greek language do not fully correspond to the Italian. Furthermore, this topic is usually stressed in L2 books focusing in grammar rules. The Greek student frequently feels a need to transfer spontaneously the use of auxiliary from his/her mother tongue instead of using the form as it is originally used by the native speakers. In accordance to the above observation, this research will analyze the errors in the use of auxiliaries and compare the results with the use of the auxiliaries in the L1 corpus.

The results have revealed different types of errors divided in two main categories: Errors caused of the overuse of rule- presented in the didactic tool- and errors due to translation strategies, for example replacement in order to achieve specific communication results. Although the conclusions have demonstrated the students’ difficulty in the use of auxiliaries, the same conclusions show the way to affront such difficulties (Granger, 2015), using interference as a tool in cases of similarities of L1 and L2, and as an example to avoid when transfer creates more serious problems than misuse.
S134 Francesca Perri

The Mistakes Laboratory: digital didactics, error analysis and corpus linguistics in secondary school teacher training

This paper describes a proposal for a training course addressing English as a Foreign Language (EFL) teachers and focusing on the introduction of Corpus Linguistics in their didactics and on the digitalization and creation of learning materials in particular. It mainly aims to support the development of intermediate learners’ written production.

The paper reports on a training course that involved five secondary schools, twenty teachers of English as a Second Language, and a hundred intermediate students, in northern Italy. The training course was based on the creation of a Mistakes Laboratory. The educational approach adopted drew on a multifaceted coexistence of various theories such as connectivism, experiential learning, project-based learning and, most of all, TEDD educational model. TEDD (Technologies for Education: diversity and devices) is an advanced course held by the Department of Engineering and Information Technology at the University of Trento, Italy, intended for graduated and teachers. The Mistakes Laboratory was designed as the main project of the author’s six-month TEDD internship.

The paper provides an overview of the Mistakes Laboratory organized into three parts. Firstly, a series of semi-guided activities drove participants to familiarize with corpora through a Digital Didactics Initial Training (DDIT) as an introduction to the following stages. Secondly, participants were involved in a Mistakes Session. The subsequent Error Analysis enabled the teachers to group the students’ errors in the translation of a text which was tailored to adapt to the intermediate language level indicators and descriptors, classified according to the Common European Framework of Reference. Thus, eight classes of intermediate students were provided with a digitalized feedback comprising grammatical references and concordances shown from two software programs (Sketch Engine and GLOWBE). Thirdly, by taking into account the findings of the ML, teachers planned,

References


created and shared a set of digital materials to innovate their teaching practices, supported by the use of corpora, and a kit of digital tools, selected according to their educational functions.

The adoption of good digital practices is a fundamental step to promote the introduction and use of Corpus Linguistics in EFL teaching and learning processes. It is by analyzing the students’ written production that significant suggestions for language methods and pedagogies can be acquired to improve both teaching practices and the students’ performances, with the support of educational technologies and corpora. In fact, dealing with corpora not only promotes the learners’ critical thinking and orienteering versus wild browsing for second language ready-made translations or examples, but it also triggers a growing confidence with digital tools and activities. From this perspective, the Mistakes Laboratory stands as a crucial learning space where teachers may collaborate and create lesson plans and digital materials, build their digital repository, share findings and difficulties, and flip their lessons in a vibrant community of inspirational digital practices.

The paper will explain reasons, limits and resources of the Mistakes Laboratory as a training course for EFL teachers who aim to contribute to the intermediate learner’s written production with the support of corpora and the creation of digital didactics.

S137 Lan-Fen Huang, Tomáš Gráf and Nicole Keng

Assessing spoken learner corpus data using the CEFR scales and error rates

Contrastive interlanguage analysis (Granger, 2002) exploits learner corpora to compare the language produced by learners with that of native speakers and that of learners from different first-language backgrounds. One of the largest spoken learner corpora, the twenty-sub-corpus Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin et al., 2010), takes a somewhat problematic approach to defining proficiency; it is not explicitly attested but simply inferred from the educational background; thus, 3rd- or 4th-year English majors are expected to be advanced. The reality may, however, be different and this may lead to the misattribution of learner usages to L1 transfer and affect research implications.

This paper first demonstrates how the ‘can-do’ descriptors in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, pp. 28-29) is adopted to assess learners’ proficiency levels in the Czech (n = 50) and Taiwanese (n = 50) components of LINDSEI. The credibility of CEFR, which is empirically under-explored (Alderson, 2007; Hulstijn, 2007), is evaluated by comparing the rating results of two trained raters.

In terms of holistic scores, the two raters agree exactly in 44 per cent of cases. Further analysis shows the close similarity between them when performing global assessment (r = .889). The cross-tabulation
shows that in one-third of the samples Rater 2 scores higher than Rater 1 and they agree more often at the higher levels. Yan (2014) also finds that rater alignment is particularly difficult at lower levels and recommends training that focuses on rater disagreement. The post hoc assessment results in a division into four groups: B1 (n = 9), B2 (n = 51), C1 (n = 38), and C2 (n = 2). While most of the Czech learners (n = 36) are at C1 level, most of the Taiwanese learners (n = 39) are at B2. This distribution shows the unreliability of using learners' educational backgrounds to infer their proficiency in English.

The study goes on to ask how far the results of aural accuracy evaluations carried out by two trained examiners match the results of a post hoc error analysis of LINDSEI-CZ, which has been error-tagged using the Louvain Error Tagging Manual (Dagneaux et al., 2008). The accuracy levels from B2 (n = 13) and C1 (n = 35) to C2 (n = 2) are confirmed by the decline in the error rates (2.01, 1.22, 0.45 errors respectively per hundred words). There is a strong negative correlation between the perceived accuracy levels awarded by Raters 1 and 2 and their error rates, $r = -.53$, $p < .0001$ and $r = -.5$, $p < .0001$ respectively.

This paper illustrates how post hoc proficiency ratings for learner corpora may be undertaken and shows that global assessment ratings by trained raters provide highly reliable results which may be further complemented by linguistic accuracy ratings based on error analysis.

References


Concordancing for genre-specific writing: an investigation into students’ learning strategies

This study explores students’ use of strategies for learning and using genre-specific phraseology in written texts. The ability to use the vocabulary and lexico-grammatical patterns of a particular text type can allow language learners to adhere to the conventions of a discourse community, meaning that their writing is more likely to resemble that of expert writers. A data-driven approach to learning (Johns 1990) can allow teachers and learners to identify such patterns, enhancing the students’ potential to use the appropriate phraseology.

While the benefits of data-driven learning (DDL) approaches are widely recognised, there is some debate as to whether a paper-based or computer-based approach, or what Boulton (2012) refers to as hands-off or hands-on, is more effective. Vyatkina (2016) found that both were equally effectively, while a recent meta-analysis by Boulton & Cobb (2017) concluded that the hands-on approach appears to be have higher success rates. A recent study by Ackerley (2017) led to the question of whether, in a writing task produced under exam conditions, students tended to reproduce phrases identified in paper-based DDL tasks or in hands-on tasks, or indeed whether they preferred to use phrases that they had identified in single text (non-corpusbased) tasks. A further question is whether they are able to access a corpus to retrieve appropriate phraseology independently of guided learning activities, for example under exam conditions. This study, then, aims to investigate students’ preferences for learning phraseology. That is, when students have the choice of using phraseology learnt through corpus-based tasks, do they choose to use those learnt through hands-off tasks or hands-on tasks, or neither of these?

This paper aims to answer these questions in order to lead to greater understanding of effective approaches to DDL in language courses. It describes a study carried out on corpus use by a large class of first-year language majors at an Italian university. Both paper-based ‘noticing’ tasks and hands-on tasks using the concordancing software AntConc to investigate a specialized corpus were integrated into classroom lessons and computer lab sessions to raise students’ awareness of the text type’s particular phraseology. In the end-of-course exam, students were required to write a text similar to those studied, with instructions to use specific key words selected from those encountered in the DDL activities. The students were permitted (though not obliged) to access the corpus using AntConc during the exam to help them produce genre appropriate phraseology.

Analysis of the students’ texts as well as responses to a post-exam questionnaire provide insight into the students’ choices of phraseology; whether their choices were based on language learnt through hands-on tasks, hands-off tasks or neither; and their attitudes towards the corpus-informed tasks. The
final part of the study looks into a possible correlation between students’ attitudes towards use of corpora and their exam scores. The implications of the findings of this study on corpus-based approaches in the classroom are discussed.

References


---

**S142 Hyeson Park**

*Relative clauses in a learner corpus: exploring the interaction of text type and L2 proficiency*

The usage patterns of syntactic constructions are known to interact with text types/registers. An implication of such an interaction for L2 learning is that learners are expected to acquire not only formal features of a target construction, but also the compatibility of the target construction with various text types. The purpose of this study is to investigate L2 learners’ sensitivity to the interplay of linguistic construction and text type, using English relative clauses (RCs) as an example. Analyses of NS corpora have observed variable distributions of English RCs across registers (Biber et al, 1999; Kam & Park, 2008; Roland et al, 2007); subject gap RCs are frequent in written text, especially in academic prose over fiction/news, whereas in conversation, object gap RCs are more common. The choice of relative pronouns also exhibits generic variation; ‘that’ is most common in fiction, ‘who’ in news, ‘which’ in academic prose, and the zero-form in conversation. The reduced RCs are a characteristic feature of academic texts (Master, 2001).

Against the backdrop of these findings, we examine L2 writing to see whether similar generic differences are found in learner data, and how L2 proficiency contributes to this pattern. We analysed a part of the Yonsei English Learner Corpus (YELC) (Lee & Chung, 2012), which consists of one million words compiled from essays (narrative and argumentative) produced by around 3,000 Korean college freshmen. The essays were graded into seven levels. We selected 40 essays from each of the
seven levels and each text type, resulting in a sub-corpus of 560 essays (99,161 wordstotal-28,893 for narrative and 70,268 for argumentative essays). The argumentative text, requiring diverse cognitive skills and persuasive rhetoric, is expected to contain linguistic features associated with an academic text to a greater degree than the less formal narrative writing (Ravid & Berman, 2010).

Our preliminary analysis has revealed some interesting features of RCs in L2 data. First, the average frequency of RCs was similar between the two text types: 13.1(narrative) vs.12.7(argumentative) (per 1,000 words). The frequency increased from 5.6 (low level) to 15.1 (advanced level) in the narrative, and 5.5 to 12.3 in the argumentative writings. Our expectation that the argumentative essays would contain more RCs was not corroborated. The most frequent relative pronoun in the argumentative essays was ‘who’(34.8%), followed by ‘that’(24.3%) whereas the frequency order in the narrative essays was ‘that’(29%), ‘zero-form’ (19.5%), and ‘which’(15.5%). The low and intermediate level learners relied heavily on one or two relative pronouns (either ‘that’ or ‘who’), while such a tendency was not shown at the advanced level.

As was observed in Ganger (1998), the reduced RCs was barely visible in the lower level writing, with most of the tokens found in the higher level argumentative essays (narrative: -ing (3.2%), -ed (3.9%); argumentative: -ing (8.4%), -ed (7.2%)).

Overall, the learners’ production of the RCs in the two different texts evidenced that they were progressing toward the NS usage pattern, though the progression was error-ridden and non-linear.

References

Introducing CLARIN resources

CLARIN is a European Research Infrastructure that has been established to support the accessibility of language resources and technologies to researchers from the Humanities and Social Sciences (Krauwer and Hinrichs, 2014). CLARIN’s vision, mission and design are aimed at findability, accessibility, interoperability and re-usability of its resources, tools and services to support researchers in the Humanities and Social Sciences (SSH) (de Jong et al., 2018). CLARIN ERIC currently has 19 member and 2 observer countries which provide numerous language resources and tools through a network of over 40 certified data centres.

The goal of this talk is to introduce CLARIN to the community of researchers, teachers, practitioners and software developers working with corpora for language teaching and learning. In the first part of our talk we will give an overview of the most relevant families of resources available in the CLARIN infrastructure, such as newspaper corpora, computer-mediated corpora and parallel corpora. In particular, we will focus on a recent survey of L2 corpora, i.e. corpora of written or spoken texts produced by learners of a second language.

In the second part we will present CLARIN’s (www.clarin.eu) most relevant web services and applications for the TALC community. The most important one is the Virtual Language Observatory (VLO) portal which enables searching for resources in the CLARIN infrastructure and provides a uniform display of metadata (Van Uytvanck et al., 2012). The Federated Content Search functionality (Stehouwer et al. 2012) enables scholars to search with a single query in multiple, quite diverse resources without having to download them or master any specialized concordancer. The Language Resource Switchboard (Zinn 2016) helps users to connect possibly distributed resources with the tools that can process them by listing all applicable tools for a given resource, specifying the tasks that the tools can achieve, and running the selected tool without the need to install or modify it.

Finally, we will demonstrate CLARIN’s depositing services through which researchers can store their own resources in a sustainable repository at a CLARIN centre, thereby archiving it in a reliable manner, making it available and more visible to the community as well as easily citeable through a persistent identifier.

References

Discovering the collocation use of change of state verbs through corpus consultation: learners’ use and performance

The study investigated how learners discover language patterns through corpus consultation to enhance the use of change of state verbs (e.g., reduce, accelerate) in their writing, with the focus on how to accurately use the noun and adverb collocates. Research has showed that Data-Driven Learning (DDL) is effective in enhancing language use in writing through error-correction approach, the other approach of DDL, namely how learners autonomously utilize corpora and discover patterns to prepare for writing has been little explored even though its importance has reached wide scholarly recognition (Boulton, 2017). Also, although collocations with synonymous constitutes are a big source of learners’ difficulty in writing (e.g., Laufer & Waldman, 2011), little has been known regarding the effective instruction through DDL. Adapting two concepts from Kennedy and Miceli (2017)- pattern hunting and pattern refining, in which learners utilize corpora to explore language and content ideas as well as checking the collocation usages, the current study intends to bridge the gaps by examining how 35 college students from Taiwan utilize Corpus of Contemporary American English (COCA) to discover content ideas about changes and collocation patterns about 20 easy and 10 difficult near-synonymous change of state verbs (Frodesen & Wald, 2016) in an 18 week Freshman English class. Data was obtained through triangulation including (1) learners’ performances: pre-test and post-test writing (2) learners’ corpus consultation from behavior logs and video files (3) learners’ attitude from questionnaires and interviews. A mixed-method approach including quantitative statistics and qualitative analysis was used. Results showed that most learners successfully discovered collocation patterns of change of state verbs even though some difficulties in corpus consultation emerged. Regarding implementing patterns into writing, most learners showed...
positive attitude toward borrowing new collocation patterns into their writing and found the collocation use in their writings has improved; some learners found it difficult to “foresee” or “settled with” the patterns they discovered prior to writing. Also, as writing is a dynamic process, the collected patterns may end up being irrelevant and unused as “a waste of time and efforts” spent through the corpus consultation process because of the shift in idea flows in the writing process. On the other hand, the attitude about how corpus consultation assisted in content idea was controversial. While some learners successfully embedded the newly found content ideas with substantial coherence built with the original themes planned in essays; some learners rejected including new ideas because they were unable to elaborate on unfamiliar themes and afraid of “making more writing errors” due to their unfamiliarity with new themes. Also, a tension was found between the “borrowed patterns” and the originally planned content. As new patterns “inspired” their content ideas, some learners found their creativity “limited” by the new ideas.

(1 Twenty easy change of state verbs include climb, drop, enlarge, expand, extend, fall, gain, grow, intensify, lose, lower, multiply, peak, raise, rise, reduce, shrink, sink, spread, swell. Ten difficulty change of state verbs include accelerate, contract, decline, diminish, escalate, plunge, proliferate, skyrocket, slash, soar.)

S150 Christoph Wolk and Bridgit Fastrich

Custom concordancers for the ESP and linguistics classroom

Language corpora have promising applications in both ESP and linguistics classrooms: they can bridge the gap between learning about and actually engaging in research, they allow learners to autonomously formulate and check their own theories about how certain text types work, and – since they involve naturalistic language use – the practical applications of the classroom material tend to be clear (cf. e.g. Gavioli 2005; Götz & Mukherjee 2006; Boulton & Cobb 2017).

Despite these clear benefits, corpus work has not yet seen widespread adoption in either ESP or linguistics teaching (Römer 2006; Boulton 2011). We believe that this is at least partially caused by the tools available for corpus linguistics. Classic general-purpose concordancers like AntConc (Anthony 2014) or Wordsmith Tools (Scott 2016) are powerful, but often also unfamiliar to learners. This and the initial setup required provide a challenge both for student motivation and for classroom time management. The very fact that these software packages are usable with any corpus design also means that they generally cannot take the unique characteristics of the corpus into account; making it inconvenient, for example, to restrict analyses to certain subsets of the corpus. Web-based concordancers such as CQPweb or the interface to the BYU corpora (Hardie 2012; Davies 2008-), on the other hand, do not require students to install additional software, but they are usually either
difficult to set up and/or use, or do not allow the teacher to supply their own corpus, limiting the possibilities for application.

We propose a third option: corpus-specific custom concordancers as web applications. These concordancers should be similar in appearance and usage to modern web pages that students are familiar with, and thereby lower the initial hurdle. They should be easy to pick up and reward autonomous exploration, but also allow a range of more advanced tool to facilitate specific research questions and projects. To this end, we introduce Shinycon, an open-source, cross-platform (Windows, Mac, Linux, mobile) framework for building custom concordancers. This framework is built with the package Shiny (Chang et al. 2016) for the programming language R (R Core Team 2016) and yields concordancers that can be hosted on the web or run on local computers. R is frequently used as a statistics toolkit within linguistics departments, so the construction of custom concordancers can often leverage existing expertise. Our primary aim, however, is to even allow users with minimal programming knowledge to construct simple concordancers for specific kinds of data. In fact, given a suitable corpus, our tools can create such concordancers automatically. In our talk, we will demonstrate some of the features of ShinyConc-based concordancers using webapps of various specialized corpora (e.g. television dialogues, corporate mission statements, advertising platforms) we have employed in our own teaching, and show how users can leverage this software for their own pedagogical or research-oriented corpora.

References


Novice second language (L2) writers likely perceive “good academic writing” as impersonal (Tang and John, 1999, Hyland, 2002, and Shen, 1989). Yet research has shown that every linguistic and rhetorical choice a writer makes, including, e.g., the presence/absence and moves of self-mention, potentially reveals the writer’s authorial identity (Ivanič, 1998). The implication that academic writing is dialogic in nature as manifested through strategic self-mention is still overshadowed by other linguistic issues in L2 writing pedagogy. Therefore, in this presentation, I will report on the findings of a corpus-driven descriptive inquiry into the notion of authorial identity operationalized as the use of first person pronouns in a corpus of 126 research-based argumentative papers written by L2 students (100% Chinese, female: 56, male: 70, mean age: 18) enrolled in first-year L2 writing courses (word tokens: 133,266). Specifically, I will look into how L2 writers practice self-mention by comparing the frequencies of first person pronouns (I, me, we) used in the corpus with a larger corpus (texts: 2,268; word tokens: 1,961,271) containing other genres (narrative, proposal, interview report, and literature review) produced by the same group of writers, and with published research articles examined in Hyland (2001). I will also define and characterize the five qualitatively coded and quantitatively measured rhetorical functions of the subjective singular case “I” used in the corpus (reporter, architect, narrator of personal experiences, conceder, and opinion-holder). The results show that the Chinese L2 writers in this study used subjective singular first person pronoun “I” less frequently in argumentative papers (0.52%) than in narratives (5.02%), proposals (2.41%), and interview reports (1.58). Additionally, L2 argumentative papers likely carried a narrative tone, as indicated by the lower ratio of subjective/objective case of pronoun (5.78) than that in Hyland’s (2001) corpus (14.00). A comparison of rhetorical functions reveals that nearly half (49.35%) of “I”s in the corpus are used as “narrator of personal experiences,” 26.12% as opinion-holder, 14.14% as reporter, 9.38% as architect, and 1.01% as conceder. Lastly, I will provide pedagogical suggestions
for college-level L2 writing teachers on more effectively initiating students into academic discourse communities through facilitating textual presentation of their discursively appropriate and stylistically idiosyncratic authorial identities.

References


---

**S157 Christine Sing**

**Using corpora for task-based ESP writing instruction: The case of source-based business writing**

Course design in English for Specific Purposes (ESP) settings tends to be underpinned by three main principles: genre, discourse community and task (Swales 1990, 2004). Even though the concept of task has been shown to be vital for L2 teaching and learning (e.g., Ellis 2003, 2004) in general, particularly however for business writing (Tardy 2012), research into task-based language teaching has only recently shifted its focus from spoken to written communication (Ortega 2012: 405). Little is thus known about how task relates to writing, including pedagogic genres such as L2 assignment writing.

The present corpus analysis of an ESP writing task will help to address this research gap. In focusing on intertextual practices, it will be shown that the writing of business students is deeply influenced by both task representation and requirements. While intertextuality covers a great variety of language phenomena, including direct quotations, copy-and-paste jobs or paraphrasing (e.g. Petrić 2012, Pecorari & Shaw 2012, Davis & Morley 2015), this study focuses on citations as well as (inter)textual functions of source use. Specifically, it will be argued that, in an ‘open task’ such as the one in question, these student writers tend to rely on expert models in appropriating source use; crucially this appropriation also includes the level of textual organisation.
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 these naturalistic data, the study is grounded in an integrated approach combining corpus analysis and computer-assisted textual analysis (CATA).

The findings show that there are considerable individual differences in source use as manifested in terms of usage patterns, dispersion and frequency. It would thus seem that intertextual strategies, while pervasive throughout the corpus texts, have limited range and tend to cluster in specific sections of the papers, thus pointing to both localized and global intertextual practices. Overall, business students use more direct than indirect quotations, which is in line with both disciplinary conventions (Hyland 1999, Charles 2006, Petrić 2012) and the conventions of pedagogic genre, in which students use considerably more verbatim quotes than experienced writers (Ãdel & Garretson 2006).

Some of the issues emerging from these findings relate specifically to ESP writing instruction, suggesting a strong influence of two interrelated factors, one task-related and the other teaching-induced. This study is therefore of value to practitioners, wishing to blend task- and genre-based writing instruction by means of data-driven learning (Gavioli 2005) or personal corpora (Charles 2015), and curriculum designers, aiming to address student needs in terms of a task-based needs analysis (Lambert 2010).

References


False starts and self-corrections in learner and native English

Spontaneous spoken production typically contains a high number of performance phenomena such as repeats, filled and unfilled pauses. These elements may be seen as disruptive of the continuous flow of language but they may as well be viewed as means of obtaining time for planning subsequent
utterances, which is of particular importance to foreign language speakers. Much has been written about the use of filled and unfilled pauses (see e.g. García-Amaya, L. (2015)) and other disfluencies (e.g. Gilquin & De Cock, 2011) by L2 learners. Recently, attention has also been brought to the use of repeats in learner English (Gráf, 2017). However, fewer studies exist (see e.g. Götz, 2013) of the closely related phenomenon of false starts and self-corrections, which occur frequently both in learner and native spontaneous spoken language. The present study aims to explore the frequency of false starts and self-corrections in native and learner English at different levels of proficiency.

The data for the study come from the Czech and Taiwanese subcorpora of LINDSEI (Gilquin et al. 2010), each of which comprises 50 15-minute interviews with advanced speakers of English, 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. False starts and self-corrections were identified semi-automatically using a computer script and the three corpora were tagged for disfluencies using the author’s own disfluency tagger. A total of 4,422 false starts, and 1,123 self-corrections were identified. The groups of speakers and the different proficiency levels (ranging from B1 to C2) were then compared using ANOVA, log-likelihood and correlation tests.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Tokens</th>
<th>Instances of false starts</th>
<th>Instances of self-corrections</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINDSEI_CZ</td>
<td>95,969</td>
<td>981</td>
<td>426</td>
</tr>
<tr>
<td>LINDSEI_TW</td>
<td>83,437</td>
<td>2515</td>
<td>591</td>
</tr>
<tr>
<td>LOCNEC</td>
<td>122,214</td>
<td>926</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 1. Numbers of tokens in the three tasks.

Significant differences were found between the three groups. The Taiwanese learners, who were mostly at B1 and B2 levels of proficiency, produced the largest numbers of both false starts and self-corrections. The Czech learners, who were mostly at C1 and C2 levels, produced fewer of them but still more than the native speakers (see Figure 1). It appears that both groups of learners overuse self-corrections compared to native speakers, but the most striking difference is the Taiwanese learners’ highly frequent use of false starts. The paper will also show the analysis of the type of false starts and self-corrections with regard to their length and position within clauses and constituents and a detailed breakdown of correlations at the several different levels of proficiency.

Figure 1. Boxplots showing the distribution of self-correction rates and false-start rates (expressed in occurrences per hundred words on the y-axis).
The results suggest that speakers at lower proficiency produce false starts more frequently but that self-correction rates do not necessarily decrease so rapidly with increasing proficiency. It would appear that the frequency of false starts is a strong predictor of fluency and one which makes comprehension harder.

References


---

**S160 Patricia Tosqui-Lucks and Malila Carvalho De Almeida Prado**

**Pragmatic features of radio communications between pilots and controllers: raising students’ awareness through corpus analysis**

After verifying that pilots and air traffic controllers’ insufficient language proficiency has been a contributing factor to several incidents and accidents of civil aviation, ICAO (International Civil Aviation Organization), a specialized United Nations agency which regulates civil aviation worldwide, requested a proficiency level to be accredited in the licenses of these two professionals when operating internationally. To obtain this license, applicants are required to take a proficiency exam prescribed in the Manual of Implementation of Proficiency Requirements (ICAO 2010). Since
the primary concern of ICAO is aviation safety, it is of utmost importance that training and testing professionals from ICAO member states make sure that the courses and tests they use to implement the policy accurately reflect and assess the performance of pilots and air traffic controllers in critical workplace circumstances.

The language of air traffic control is distinct from other English for Specific Purposes contexts because it has a restricted set of functions and a phraseology with reduced syntax and vocabulary for routine actions but, at the same time, involves shared information concerning the aircraft in the area, the parameters of the airport or airspace, and expected actions in the flight (Moder & Halleck, 2009).

In order to study and describe Aviation English from the Corpus Linguistics perspective, and to use this description for teaching purposes, we have compiled and transcribed two spoken corpora with radio communications between pilots and air traffic controllers: one with speakers that operate in the Brazilian airspace exclusively, and another with communications from all over the world, with international traffic handling problem-solving situations. By analyzing data of these corpora, we have come up with a number of pragmatic features that may affect communication and therefore should be addressed in the aviation English classroom (Prado & Tosqui-Lucks, 2017). Research on spoken corpora has brought about some issues that Pragmatics has been dealing with for a long time, such as the use of hedges to express politeness (Adolphs, 2008). The combination of Corpus Linguistics (CL) and Pragmatics, or Corpus Pragmatics (Aijmer & Rühlemann, 2015), provides an insight into the automatic statistical processing of frequencies that allows the grouping of certain chunks according to their pragmatic function (Adolphs, 2008).

Using this methodology, we will analyze some pragmatic features of the abovementioned corpora related to: frequent hedges to express politeness; chunks mistakenly used; transition from scripted to spontaneous language, and to some interaction strategies such as avoidance, complaining and apologizing, all of which are identified in the problem-solving contexts. Based on the language investigated, we will propose activities to be applied in the classroom (Viana & Tagnin, 2010, McCarten, 2010), in order to draw students’ awareness to elements that will allow them to enhance pragmatic competence (O’Keeffe, Clancy & Adolphs, 2011) and to understand radiocommunications as a means of sharing the responsibility of assuring safety in the skies.

References


S161 Maria Kunilovskaya and Natalya Morgoun

*Do translation textbooks address real-life translation problems: evidence from corpus-based error analysis*

In this work we align perceived translation difficulties discussed in the textbooks with the observed translation errors and suggest a corpus-informed approach to translator education. The aim of this research is to generalize about the methodologies adopted in professional translator education in Russia (English-to-Russian) and describe difficulties that learner translators fail to overcome in their actual output. One of our tasks is to find a way to generalize about translation problems, drawing from the translation error annotation data. The research is based on texts from Russian Learner Translator corpus, a large multi parallel bidirectional English-Russian corpus. On the one hand, we analyze the content of 50 textbooks, published in Russia from 1963 to 2017 for students of translation. On the other hand, we report the distribution of translation errors, tagged in multiple student translations of informational texts (633 translations (263 K words in total) to 75 sources (42 K words), total number of error tags is about 13,5 K). The annotations are produced in an online annotation environment over the period of 4 years as part of assessment procedures within an advanced practical translation course for university students majoring in translation. The annotation process follows formal annotation guidelines and is based on a customized error typology of 30 individual error types arranged in 7 categories. The general statistics on error categories (shown in Figure 1.) does not provide enough information with regard to the source of the translation difficulties that students face, therefore we...
turn to parallel data analysis. To identify recurrent issues in student translations we perform contrastive analysis of sentence-level translation units, grouped according to the categories of errors tagged. To reveal problematic areas we identify source sentences and their parts that provoke most errors in multiple translations and analyze their structure and lexical setup. We build frequency lists for annotator notes and analyze error attributes which optionally characterize the importance of the tag (“gravity of the error”) and explain why it is added. The results of this research are supposed to shed light on the major reasons behind the errors and supplement findings based on comparative analysis of learner translations and non-translations (translationese studies).

We find that translation problems experienced by most learner translators are of non-linear nature and do not involve processing discrete source language elements such as determiners, gerunds, modals or passive voice. We argue that the structural cross-linguistic differences, which are often targeted in translation textbooks, are less relevant to the observed translation problems. In our data most errors originate from lack of text-level functional translation competence, poor text comprehension skills and inadequate production skills in the target language rather than grammatical or lexical language competence. This work has been partly supported by the Russian Foundation for Basic Research within Project No. 17-06-00107/18.

![Figure 1 Distribution of error types in EN<->RU translations (in %)](http://www.rus-ltc.org/download/RusLTC_error_typology2016.pdf)

1 [http://www.rus-ltc.org/about.html](http://www.rus-ltc.org/about.html)

Mastering the social dimension of language use is as important for successful communication as mastering the grammar and the lexicon; whenever we use language to communicate, we convey not only the content of the message but also a complex array of social meaning (Bardovi-Harlig, 2012, 2013). Subjectivity in language refers to how speakers express “their perceptions, feelings and opinions in discourse” (Scheibmann, 2002) and to the linguistic features and structures that enable “self-expression in the use of language” (Lyons, 1994; Benveniste, 1958). The expression of subjectivity is an important component of pragmatic ability as it is closely related to how speakers communicate politeness (e.g. boosting or downplaying one’s involvement) or their stance in interaction (Reilly et al., 2005). In order to contribute to our understanding of how spoken, interactive production develops in learner language, the paper investigates the subjective involvement and pragmatic stance expressed in the ‘I + verb’ construction by speakers of different levels of English proficiency. The study, in particular, it focuses on two prominent lexical categories of verbs (Biber, et al, 1999; Scheibman, 2002; Levin, 1993) that occur in this construction and express speaker’s stance – emotive (e.g. love, need, wish) and cognitive/epistemic (e.g. believe, think, suppose) verbs.

The study uses the Trinity Lancaster Corpus (TLC) of spoken L2 production (Gablasova et al, 2015) based on examinations of spoken English conducted by Trinity College London (a major examination board). L1 Spanish and Italian speakers aged over 20 years (to control for the effect of cognitive maturity on expressions of subjectivity) were selected from the TLC to represent three proficiency levels of the Common European Framework of Reference: B1 (183 speakers), B2 (170 speakers) and C1/C2 (102 speakers). All speakers participated in two interactive speaking tasks (conversations) which together lasted approximately 10 minutes. Using MonoConc Pro (Barlow, 2000), all ‘I + verb’ constructions in speakers’ production were identified and the verbs in these constructions were categorised as emotive, cognitive or other (e.g. material, auxiliary) verbs (e.g. Biber et al. 1999; Scheibman, 2002). The ANOVA was used to compare the frequency of each verb category across the three proficiency levels.

The findings show a very clear and statistically significant trend in the use of the ‘I+ verb’ construction. With the increase in proficiency, the frequency of emotive verbs decreased while the frequency of the epistemic verbs increased considerably. The study also identified the most frequent cognitive and emotive verbs and the trends in their use according to the proficiency level of L2 users. The study contributes to a larger discussion of the effect of lexico-grammatical competence on the development of pragmatic competence (e.g. Schauer, 2013; Kasper & Rose, 2002) and discusses the findings from the perspective of second language pragmatic ability.
References


---

S163 Kyoko Sugisaki and Michael Prinz

“The brindled university language” - building a corpus from handwritten historical documents for teaching linguistics in practice

University lecture notes have hardly been studied by German linguists so far, although they provide insight into the communicative conditions of university tuition and the academic languages in general (Prinz 2017). In this abstract, we (a) report the development of a corpus of 18th and 19th century lecture notes which was carried out during a seminar at the University of Zurich and (b) discuss the pedagogical implications for teaching historical linguistics.

Although there are many research projects (e.g. Verbmobil project in Reithinger & Kipp 1998) in which Linguistics students are engaged in transcribing and annotating texts for the creation of a corpus, there are few projects (e.g. RIDGES: Register in Diachronic German Science, cf. Springmann...
et al. 2018) whose focus is on analyzing and annotating real texts with an annotation guideline within a Bachelor’s or Master’s study of linguistics. In our seminar of German historical linguistics, we have developed a lecture corpus from various disciplines with the students. Our „Diachronic Lecture Corpus“ consists of samples (6000 tokens per lecture) of handwritten university lecture notes from the Age of Enlightenment to the 19th century (lectures delivered by Kant, Hegel, Lichtenberg and others). The corpus offers an empirical basis for linguistic investigation such as:

– communicative needs of different academic faculties and disciplines
– certain academic forms of communication (e.g. scientific controversies, cf. Gloning 2018)
– multilingualism and language choice in academic communication
– written and spoken language in the history of academic communication
– scientific terminology

Although most historical lectures exist only in the form of manuscripts, some of the notes chosen for the corpus are available as printed or digitized editions. However, these editions are often not the exact copy of the handwritten text, i.e. some of the linguistic properties such as writing variations are lost. Therefore, the students were taught to decipher and transcribe the original handwritten texts, comparing them with the printed and digitized texts. The transcription was conducted in form of a simplified XML format so that some textual properties such as headers and character types could be annotated during the transcription. The transcribed texts with XML mark-ups were then automatically segmented into tokens (Sugisaki 2017) and converted to Microsoft Excel documents for further annotations. In the seminar, the following annotation was performed by the students:

– normalized word form
– types of compound forms
– text-linguistic and pragmatic properties such as intertextuality and person deixis
– sociolinguistic aspects of multilingualism and code-switching by means of the “biscriptality” (Kurrent vs. Antiqua) of the text (Bunčić et al. 2016).

For the students, we have provided a guideline for the annotation and assistance during the procedure. At the end of the seminar, the Excel documents with the annotations were compiled to NoSketch Engine (Rychlý 2007), so that the students can use the corpus for their term papers. In the presentation, we will further discuss pedagogical challenges of creating a corpus with students, such as learning to read historical handwritten texts or the proper use of corpus-linguistic tools, and some prospects for learning linguistics in practice.

References

Stance in a second language first-year college writing course: genre, process, and writers type

Stance is central to academic writing, functioning as a linguistic and rhetoric tool to convey a writer’s personal feelings and attitudes on, and evaluation of the proposition (Gray & Biber, 2012). The linguistic and rhetorical realization of a writer’s assessment expresses a writer’s expectations towards readers and proposes a specific role for readers. In this regard, stance animates written texts as a community where social interactions and negotiation of meanings occur between writers and readers (Hyland & Guinda, 2012). These dynamic and complicated features of stance create challenges for novice writers, especially second language (L2) writers who are not accustomed to academic writing genres and convention or/and English language use. However, previous writing research on stance has focused on established, expert writers in academia (Hyland, 2005) but paid little attention to L2 novice writers in pedagogical settings (Staples & Reppen, 2006). Thus, the current study investigates
a corpus of L2 writing collected from first-year undergraduate writing courses for their use of stance. Specifically, the study is centered on the variation of using stance, in terms of lexico-grammatical features, by different writing genres, writing revision stages, and writers’ profiles. For interpretations, the varied distributions of the current corpus were compared to general patterns of stance in academic writing from previous studies (Biber, 2006; Biber et al., 2011).

Using Biber’s (2006) lexico-grammatical stance framework and his computerized tagger, 852 students’ essays were automatically coded. For analysis, normed frequencies of each stance device were compared among 1) different genres (three course assignments, literacy autobiographies, synthesis papers, and argumentative essays), 2) revision stages (first and final drafts), and 3) writer profiles (gender, TOEFL total and subskill scores, first language, and majors). The results from descriptive analysis revealed that in general few changes were made in the patterns or frequencies between different assignments or drafts. However, regarding the variation associated with writer profiles, I observed relatively larger differences between groups with different TOEFL writing scores. Particularly, the difference between genres was higher in a group with higher writing scores. In addition, the results indicated that L2 students in a first-year writing course predominantly relied on modal verbs, specifically can, when expressing stance while published authors used more diverse grammatical structures, including a that- or to-complement headed by a stance noun (e.g., We cannot rule out a possibility that…). The presentation will discuss statistical significance from factorial ANOVA results. The findings suggest the need to incorporate instruction of diverse, genre-based stance use into L2 first-year college writing curriculum to enhance their academic written communication.

References


Corpus-based research into language learning has highlighted how phraseology represents a source of considerable difficulty in foreign language learning (cf. Osborne 2008). However, as Granger and Meunier (2008: 247) point out, “although phraseology has received more attention, […] it would be misleading to claim that there is uncontroversial consensus about its role in pedagogy”. With reference to translation, research has drawn attention to collocations as a potential source of problems for translators (cf. Hatim and Mason 1990: 204) and phraseology has been identified as a key element to determine the quality of a translated text (Colson 2008: 201).

The present paper reports on research carried out on a small learner corpus of specialized texts in the area of economics translated from Italian into English by Italian postgraduate students specialising in the area of international management and attending a translation course as part of their university curriculum. The research aims to 1) identify the key problem areas for the students in relation to the translation of economic terms 2) investigate whether corpus-based practice can help learners overcome these problems. An initial, qualitative analysis of the translations done by the students showed that most of the specialised terms had been translated correctly by the students, while the collocations were often incorrect. The qualitative analysis was then followed up by a quantitative analysis of the corpus which adopted corpus linguistics tools and methodology for the analysis of the collocational patterns involving discipline-specific terms. The analysis confirmed that, while the specialised terms did not pose particular problems for the student, mistakes often occurred in the translation of phraseological patterns such as NOUN+NOUN, ADJECTIVE+NOUN, VERB+NOUN and NOUN+VERB constructions. Examples include the wrong translation of the Italian economie di mercato as market *economics, and of the Italian dare la foro as *give employment. The research therefore proceeded to explore whether the introduction of corpus-based practice in the translation course would help the students in mastering phraseology involving specialised terms. For the purpose of the study, a monolingual corpus of specialised texts in the area of management and economics in English was built and the recurrent collocational patterns of economic terms were researched and presented to the students in the form of tasks based on a “data-driven learning” (Johns 1991) approach. Class discussion and self-correction of the students’ own translations based on the data from the corpus (cf. Watson Todd 2001) was also encouraged. The students’ translation tasks were monitored throughout the course in order to verify whether there was an improvement as regards phraseology in the translated texts.
The study showed that, as well as helping the students learn frequent collocational patterns in their area of specialisation, the corpus-based tasks also increased the students’ level of awareness of the importance of phraseology for the naturalness of the translated texts. This also led to an improved use of the English monolingual dictionary in the search for appropriate collocates for the specialised terms in the translations. The research provided evidence, therefore, that specifically designed corpus-based teaching materials are engaging for the students and effective in helping them in a particularly critical area related to specialised translation.

**References**


**S170 Paul Wickens**

**DIY corpora of students assessed writing: engaging year 3 applied linguistics students.**

The paper reports on a CL assignment in year three of a BA degree in English Language and Communication which asks students to construct, analyse and critically reflect on a corpus of their own assessed writing from years 1 and 2. Students work together to build a class corpus and are asked to compare their individual writing in groups of three (using AntConc) and to use the class corpus (on Sketch Engine) and the BAWE corpus as wider reference corpora. They have three pre-defined queries based on three set readings around voice and stance (Hyland 2002, Groom 2005, Hyland 1999). The assignment frames this within a broader question asking them to consider whether there is a “tension that student-writers often experience between what they feel they want to say and what
they feel they are allowed to say in their academic writing." (Lillis 1999 p73). The study currently has writing from three cohorts totaling 900,000 words.

The paper will briefly outline the rationale for the assignment in terms of student engagement in year 3 modules on SFL, CDA and CL. I argue that in researching their own writing practices and setting this within a critical frame the task can ‘make real’ abstracts concepts when talking about text as social practice and it allows for consequential discussion of their own lived experience of ‘doing being a student’ (Gee 2014). It also is timed to prompt them to consider their writing as they step up into level 6 and dissertation writing.

The paper will go on to report on the students’ corpus findings and interpretations from their class presentations. The study takes as its jumping off point the students’ own insights into their data and the discussions they have about both inter and intrapersonal variations in their writing as a cohort. They identify relatively common concerns in CL in academic writing around disciplinary and genre variation (Nesi and Gardner 2012) particularly around students doing joint honours. However, they also identify substantial interpersonal variation in the cohort even within the same disciplinary area and genres. As Gablasova et al (2017) point out whilst task and genre variation are often key, there is a need to track ideolectal variation. The corpus design also allowed students to focus on temporal variation over the two years of study. Further, contextual insights were also possible due to the insider knowledge students could bring to bear such as the influence of individual lecturers on student writing.

Finally, drawing on the students assessed critical reflections on the task and follow up interviews, the paper evaluates the insights students have gained from a corpus approach to their practice of assessed writing at university.

References


Teaching EU English to national judges: terminological collocations in EU Competition Law

The objective of the paper is to report on a study of terminological collocations in EU Competition Law, conducted for the purposes of English language training for national judges from five countries (Poland, Italy, Croatia, Spain, Greece), organised as part of a larger project funded by the European Commission’s DG Competition. The study involved a corpus-driven identification of key terms, analysis of their collocational environment, compilation of a glossary and preparation of e-learning exercises for the judges taking part in the project.

Collocations and other phraseological units are well known to cause problems to non-native speakers of languages as well as to native speakers who are semi-experts in specialised languages (e.g. translators). These problems become exacerbated in the EU context, where EU institutions use English as their main procedural language. English used in the European Union tends to be regarded as a distinct hybrid variety of English (Modiano 2017, Doczekalska 2009), known under a number of names: EU English, Euro-English, and Eurish. In most cases, EU texts are written by non-native speakers of English (cf. Wagner et al. 2002: 70; Tosi 2002: 178; Gardner 2017: 150; Hanzl & Beaven 2017: 141). Secondly, documents are prepared and drafted in a multistage and multilingual manner (Doczekalska 2009: 360) with a text going back and forth, with the help of translators, through various languages and back to English. As a result, English undergoes multiple types of mediation through translators and non-native speaker authors; hence, an extreme filtering through other working languages. On the other hand, EU English also reflects attempts to build a neutral legal meta-language (Šarčević 2010: 34–35), ‘a go-between’ which has been ‘reinvented’ to facilitate multilingual translation (Pozzo 2012b: 1198, Crystal 2003: 182, Robertson 2010a: 3, 6). This background raises a number of questions as to the naturalness of collocations of neutralised and deculturalised legal terms.

To answer these questions, we compiled the EU Competition Corpus with the help of legal scholars. The corpus comprises key EU competition legislation, judgments and guidelines. It was uploaded to SketchEngine, where it was processed, including POS tagging and sketch grammar, to enable term and n-gram extraction against a number of pre-loaded EU and non-EU corpora available on the platform. The lists were next sorted manually and reduced to over 100 terms. Collocations of these terms were extracted for the purposes of the compilation of a glossary and were analysed, pointing out areas of interest for trainers, e.g. salient patterns, neo-classical compounds, frequent participial premodification, negative prosody, an increased variation of collocations and their patterns.
served as a basis for preparing collocational exercises for the judges. The study points to the usefulness of corpora in teaching languages for special purposes, in particular hybrid EU English, contributing data from the underresearched but important group of learners with specific professional needs.

References


Conference 7/14 presentations

S22 Rudy Loock

How to make translation students aware of the insufficiency of grammatically correct translations through corpus data


This presentation aims to explain the corpus-based methodology used with advanced translation students (master’s level) in an English-French comparative grammar class in order to help them provide translations that do not only respect grammatical rules but that also take into account
grammatical usage, something they generally find difficult to achieve. Our 3-step approach is to be laid out:

1. Use of quantitative data from a comparable corpus of original English and original French data for a specific linguistic feature.
2. Use of qualitative data from a parallel corpus of English texts translated into French by professional translators so that students can observe how they deal with the translation of sentences containing the linguistic feature.
3. Use of the knowledge acquired from observations in steps 1-2 to translate a new batch of sentences with the linguistic feature.

The aim of step 1 is to make students aware of the existence of usage differences between the two languages in spite of grammatical equivalence (e.g. passive voice, existential constructions, which exist in the two languages with similar discourse functions), leading generally to a significant difference in frequency of use. The aim of step 2 is to help students find solutions to translate such linguistic features in a non-literal way, that is by resorting to another grammatical construction although a literal translation would have been perfectly grammatical. Step 3 consists of one or two translation exercises where students need to translate sentences by using the information collected from both the observation of the comparable corpus and of the parallel corpus. The general aim is to help students write natural-sounding, idiomatic translated texts, by having them use both “manufactured” and “do-it-yourself” (DIY) corpora (Bernardini & Ferraresi 2013).

Our presentation will provide concrete examples of linguistic phenomena that require such an approach for English-French translators (translationese-prone phenomena due to a difference in usage between the two languages), as well as feedback from the students, who find the approach generally interesting but time-consuming.

References


Language learning from an audio description corpus

The last years have been particularly profitable to the development of inclusive sociocultural practices like the production of audio descriptions, that means, additional soundtracks informing blind or visually impaired people about visual events considered necessary for the understanding of the story told by a film. Audio description scripts have been analysed with corpus techniques (Salway, 2007; Arma, 2011; Matamala & Villegas, 2016), and didactic applications of this highly standardised text genre are presented by Burger (2016) as well as by Ibáñez Moreno and Vermeulen (2017). Some of the methods and tools used for creating audio descriptions (AD) have qualities which go beyond their primary objective: writing an AD script and recording it with a free online tool like Youdescribe (2017) may help foreign language learners conceptualise complex language features such as the linguistic expression of space and visual perception. In addition, in France, the foreign language learning syllabus applicable to year 11 (15-16 years old students) focuses on solidarity and encourages the use of digital tools. It can be expected that producing an audio description gives such learners the opportunity to deal with linguistic, technological and societal challenges.

I invited a cohort of future German teachers enrolled into a French master program (n=14, divided into four groups) to design secondary education lesson plans containing at least one component of an audio description. One of the groups, aiming year 11, produced a complete audio description for an episode of the web soap Jojo sucht das Glück [Jojo’s pursuit of happiness] (n.d.). I used their script and their recording on Youdescribe as a starting point for designing corpus-based learning activities.
which match the needs of German learning audio describers, Cefr level B1/B2 (Common European Framework of References for Languages, 2001). To achieve this objective, I created the specialised corpus Buettenwarder (2017) containing 69 text files of AD scripts produced by team members of the television broadcast company NDR Fernsehen from August 2013 to December 2016. They describe 69 episodes of the German series Neues aus Büttenwarder (Eberlein 1997-2017) telling rural stories located in northern Germany. The data consist of 336,723 tokens and are stored on TXM, a corpus tool for textual analysis (Heiden, 2010). I obtained the right to use the files for research and teaching purposes, data mining and tagging included, and to share them via Ortolang (2015), a platform linked to the CLARIN network. In this corpus, I explored prepositions, verb particles, present participles, compound words, spatial verbs, and verbless sentences. These are linguistic features supporting AD standards (Clark, 2001; Rai, Greening & Petré, 2010) like being precise, short and objective when describing audiovisual events.

My presentation contrasts selected text parts of the learner AD script, containing errors or not, with comparable Buettenwarder corpus data. The resulting observations are used as a basis for shaping learning activities implying direct corpus use by pre-tertiary learners of German.

References
University students are often provided with specialist courses in academic writing. Writing English for academic purposes requires all students to become familiarized with two key lexical forms: a) technical terminology and phrases that are usually connected with the topic written about and b) words, phrases and key collocations that are typical of the formal register of academic writing.

By introducing students to basic techniques of corpus-building and the use of a concordancer tool, these teaching session aims to achieve the following:

Raise students’ awareness of electronic aides available

Develop students’ ability to make appropriate word choices

Develop students’ range of vocabulary

S68 Michael Pace-Sigge

*Topic-targeted academic writing: how to make use of a template corpus*

University students are often provided with specialist courses in academic writing. Writing English for academic purposes requires all students to become familiarized with two key lexical forms: a) technical terminology and phrases that are usually connected with the topic written about and b) words, phrases and key collocations that are typical of the formal register of academic writing.

By introducing students to basic techniques of corpus-building and the use of a concordancer tool, these teaching session aims to achieve the following:

Raise students’ awareness of electronic aides available

Develop students’ ability to make appropriate word choices

Develop students’ range of vocabulary
Develop students’ skills in using key words and phrases.

It will be shown how students will be asked to assemble a small collection of academic texts in the field of their specific interest. These texts will be re-formatted into an appropriate format and paraphernalia will be cut from the reference corpus thus created. By creating frequency lists and lists of the most frequent clusters as found in their respective reference corpus, students will be made aware of relevant terminology. Using the frequency list, they will also create key-word lists in order to make visible the most salient concepts and wording choice currently employed when discussing the topic of their choice. These keywords can then be observed in their typical usage patterns by checking the appropriate concordance lines.

Resulting from the information gathered, students will be able to be more sure-footed in integrating suitable terminology in their own writing.

References and further reading
Pace-Sigge, M. (2015). The function and use of TO and OF in multi-word units. Houndmills,
   Basingstoke: Palgrave Macmillan.

S112 Lynne Flowerdew

Data visualisation for DDL: forging a connection and focusing concentration

The motivation for my exploration of data visualisation techniques for DDL was inspired by my science and engineering doctoral students’ positive reaction to the Michigan Corpus of Upper-Level Student papers (MICUSP) interface, featuring word clouds to display key terms and bar charts to show the sub-components of the corpus. In this seven-minute presentation I make a case for using various visualisation techniques for data display to supplement concordance output in ESP pedagogy. The use of various types of graphic representation can, arguably, forge an important connection with doctoral students and also focus their concentration, with the proviso that the graphic data are meaningfully and purposefully used, as students are familiar with such kinds of data from their disciplinary studies.
In early work in this field Doyle (2012) illustrates, for example, how ‘word trees’ are used to represent a visual concordance and bar charts can be used for showing the distribution of a search term across different genres, particularly useful for ESP pedagogy. However, at the same time, he notes the following: ‘this affordance of the software has not been fully realised in the interfaces available’ (p. 158). While concordance output still remains the main means of displaying corpus data, in recent years ESP teachers have made more use of graphic representations in existing tools, e.g. the Concordance Plot in AntConc. Various examples of visualisation data for ESP pedagogy will be shown drawn from the author’s (2015) own work and that of other ESP practitioners. It is expected that, in the near future, there will be uptake by DDL practitioners of recently developed visualisation tools such as GraphColl, which represents meaning relationships through visual collocation networks (see Brezina et al. 2015).

References

Michigan Corpus of Upper-Level Student Papers (MICUSP). The regents of the University of Michigan.

**S139 Sylwia Twardo**

*Can stylometry be used for improving writing skills? A project*

A brief review of Google Scholar reveals that stylometry may be used, for example, for analysing the variations of styles in literature, determining authorship both of original work and of translation, detecting plagiarism, as well as identification of L1 in L2 texts. The aim of this study is to find if the stylometric tools can be used in essay writing instruction in EFL classes at B2 and C1 levels.

The work will be conducted in two steps. The first one will involve conducting a stylometric analysis of two sub-corpora of student examination essays (B2 and C1 level), which yields clusters with similar sets of significant features as well as a corpus of essays written by native speakers of a similar age and education level. The obtained sets will be analysed manually in order to determine which of them are more similar to those found in essays written by native speakers. The study will be...
conducted with the use of the WebSty Open Web-based System, English version (WebStyEn), using the option: analysis of the grammatical style (Piasecki, M., Walkowiak, T., Eder, M., 2016).

The second step will involve an experiment conducted on groups of B2 and C1 learners in essay writing instruction. The experimental groups will be given access to the WebStyEn Open Web-based System where they will add their essays to the existing corpus of essays so as to determine to which cluster their essay belongs and read the description of the significant features of their essay compared with the set of those found in essays written by native speakers. The participants will be asked to reflect on the results in writing. The control groups will be given traditional instruction and feedback for essay writing. The experiment will be preceded by a pre-test and post-test (in both cases connected with essay writing). All students will fill in a questionnaire with biographical data and questions concerning their motivation and will do a placement test testing their language level (halfway through the course, in both cases for the organisational reasons). The students will also reflect in writing about their attitude to the essay writing instruction (both the control and experimental groups).

The obtained data will be analysed with the use of quantitative and qualitative methods in order to establish if the instruction influenced the use of native-like feature sets by students from the experimental and control groups, whether the level of language and motivation influenced the results and whether there are any similarities or differences in the attitude to the tasks (student reflection).

References

Yuying Hu

Register features revealed by lexical and phrasal profiles in an English logistics corpus

The present research aims at establishing lexical and phrasal profiles of logistics written data which consists of research articles, textbooks and news reporting, and exploring how these salient linguistic features reveal their register features.

The rationale is drawn on two theoretical models. One is meaning generation model (Sinclair, 1996, 2004), which claims that the meaning of a lexical word is the result of co-selection of an invariable core, colligation, collocation, semantic preference and semantic prosody. Another one is register analytical framework outlined by Biber and Conrad (2009). Namely, the register features of a language variety can be revealed by salient linguistic features and communicative purposes these linguistic features have in a certain discourse context.
Underpinned by the above theoretical models, the 3 million logistics corpus is referred against a
general BNC written corpus composed of both academic and commercial data, so that the following
procedures are conducted:

First, a word frequency list is obtained according to the methodology outlined by Carroll, Davies, and
Richman (1971) to see what words are specific to logistics field and what words are commonly shared
by logistics field and other fields (i.e. natural science, social science and humanities).

Second, following Warren's (2010) research method of phraseological variation, top 50 aboutgrams of
the target logistics corpus are established. This procedure helps us not only obtain salient logistics-
specific phrasal profiles, but also compare these discipline- specific data with salient phrasal profiles
observed in the general reference corpus.

Third, a comparative analysis of pragmatic functions of the salient linguistic features in forms of
semantic preference and semantic prosody is conducted between the target logistics corpus and its
reference corpus, to see how salient linguistic features serve their communication purposes within
specialized discipline contexts and general academic contexts.

Findings reveal that there are some differences between language uses of two data groups, and
differences also exist across sub-corpus datasets in the logistics data. These

various characterized linguistic features are not only content-related (Rõmer 2009), but also function-
related (Grabowski 2013).

The findings of the research could be beneficial for teaching practice of English for Specific Purpose
(ESP) regarding vocabulary teaching, writing tutoring as well as optimizing syllabus designs.

References

    Study of Keywords, Lexical Bundles and Phrase Frames in Patient Information Leaflets and
Warren, M. 2010. Identifying aboutgrams in engineering texts. In M. Bondi & M. Scott (Eds.),
Implementing a data-driven learning methodology in academic writing courses for science & technology doctoral students: feedback, results, and perspectives.

Developing and rationalizing academic writing skills is widely seen by language learners as a significant challenge (Fernstern & Reda, 2011). Bringing learners towards autonomy in making informed linguistic choices when writing, especially in a specialized professional context, calls for innovative approaches involving an analytical exposure to the target genre (Bondi, 2001; Weber, 2001). Various empirical studies have highlighted the benefits and limits of corpus-based and corpus-driven methods in ESP/EAP teaching/learning (Boulton, 2010; Charles, 2011; Lee & Swales, 2006; Yoon, 2008).

The present study deals with the specific case of science and technology doctoral students in the process of drafting a research article and seeking methodological and linguistic guidance. Further to needs assessment surveys among this target audience, 24-hour academic writing courses have recently been redesigned at Université Grenoble Alpes, France. A data-driven learning (DDL) approach has been developed, partly based on students’ manipulation of their own subject-specific journal article corpus, in an attempt to go beyond the production/correction-type input requested by a majority of respondents.

Ahead of the first class, each student was required to collect a selection of around ten journal articles representative of their PhD subject, preferably recent and in line with the journals they are likely to publish in. Separately, the trainer put together a corpus of journal articles in one of the specific fields of study involved, i.e. physical sciences, to be used as a reference data source. The main tools presented and used were chosen for their accessibility and user-friendliness for the target population: the AntConc concordancer (Anthony 2017) as the default tool for querying custom-made corpora, and COCA (Corpus of Contemporary American English), with the academic section serving as a reference point for general-purpose queries and methodological background.

An online workspace was set up for students to upload their written productions on a weekly basis. The collaborative nature of this environment made both peer correction and trainer input possible. It also made it easier to produce a corpus of students’ work, to track various types of patterns in learners’ approach to writing, and to include assignments as part of in-class work, such that class materials and activities would keep as close as possible to students’ needs.

This study presents a comparison between classes conducted with a semi-inductive approach, relying mainly on observation, reading and collaborative work, and a DDL approach involving the use of text analysis tools to tackle specific linguistic questions. Four groups of 15 students took part in this work, one being a control group. The analysis of the students’ written output, both from a case study and
from a group-wide point of view, provides insight into the learning process and allows both a qualitative and quantitative characterization of learners’ progress, queries, and adherence to the methodology implemented. The influence of field-specific factors on observed work and learning patterns is also discussed. Feedback provided by the students, both through surveys and dedicated interviews, completes the study through the assessment of students’ perception of various aspects of the course.

References


---

### Conference posters

**S3 Danyang Zhang**

*Mobile-based English dictionaries (MBDs) in Chinese EFL learners’ incidental English vocabulary learning: exploring effectiveness, learners’ use and attitude*

L2 vocabulary acquisition has become a popular focus of great research interest, and one of the essential aspects in ISLA (Schmitt, 2008). However, many scholars like Oxford (1990) refer to vocabulary knowledge as the most sizable and unmanageable aspect in language learning. Unlike syntax and phonology, there are not clear vocabulary rules for learners to follow to develop their vocabulary knowledge (Alqahtani, 2015). When moving the focus to China, Chinese university students attach much importance to English learning especially vocabulary learning. A majority of Chinese university students are accustomed to using rote-learning methods to memorise the spelling words and their Chinese explanations. On the one hand, this type of approach and the model may be beneficial for students’ adaptation to the Chinese exam-oriented education system. On the other hand, only a few words can be directly taught by instruction in the context of the classroom (Nagy & Herman, 1987). Also, students may become bored with mechanically remembering words and feel frustrated about their English vocabulary learning (Chen, 2001). Considering the great importance of English vocabulary acquisition and the difficulties Chinese students usually encounter, further exploring English vocabulary learning is of theoretical and practical significance and becomes the main concern of this research. Profiting from the rapid development of mobile technology, mobile-based dictionary (MBD) which is flexible and convenient, become increasingly popular among Chinese language learners. On the basis of that, my study will predominantly concentrate on MBD, especially that in incidental vocabulary learning.
On the whole, there are three main foci in this project. First, it underscored how MBDs facilitate learners’ English vocabulary learning. Second, because of the significance of learner’s decision in the way of using mobile device, this project filled in the research gap by exploring how learners explore MBDs when reading. Third, as language learners’ attitudes impact their performance (Nyamubi, 2016), what attitudes language learners have with regard to the effectiveness of mobile-based products (Ma, 2015) have been considered.

In this mixed-method research, three methods have been used for collecting data: (1) questionnaires; (2) pre- and post-tests; and (3) interviews, for collecting both the quantitative and qualitative data. To investigate how participants use MBD, the self-report questionnaire and the semi-structured interview have been collectively used. The questionnaire mainly designed to collect self-report data asked a participant to report whether they have looked up the target word, as well as what aspect they have focused on. The semi-structured interview particularly aimed to explore the reasons why more attention has been paid to certain words/aspects. For evaluating the effectiveness of MBD and the differences between the types of dictionaries, a pre-test, an immediate post-test and a delayed post-test have been carried out. Both an attitude questionnaire and another semi-structured interview have been run to determine attitude. The interview investigated the reasons why a student holds a viewpoint or other supplementary opinions.

This paper illustrates some key findings of this research with regard to the effectiveness of MBD in learner’s incidental English vocabulary learning, as well as their uses and attitudes.

**References**


A three-stage corpus-based model for L1 teaching at primary and secondary schools

Despite the fact that applications of corpus linguistics to language teaching began as early as the late eighties and nineties (e.g. Higgins, 1988; Stevens, 1991; Flowerdew, 1993), researchers seem to have neglected the idea of using corpora and corpus-based resources in L1 teaching at primary and secondary school levels, with several exceptions such as the CLLIP project for 8–10 year-olds (Sealey & Thompson, 2004), the web-based Englicious project (developed by the team of Survey of English Usage, UCL) or a data-driven experiment in L1 French (Leray & Tyne, 2016).

In this poster, I would like to suggest a three-stage model for L1 teaching at primary and secondary school level and argue for a different type of corpus data for each stage. Although the proposal is modelled to fit Czech education system with its two compulsory levels of primary school, the recommendations strive to be fairly universal and applicable to other languages as well.

At stage 1, pupils are to get acquainted with corpus materials and methods via texts that are most accessible to them, i.e. popular children’s fiction. The purpose of this stage is to introduce a new look at language, emphasising the role of context and word combinations (collocations).

At stage 2, new corpus data should be introduced, such as different text types (non-fiction and newspaper texts) and registers (non-standard spoken language in addition to the standard written). At this point, students learn about the variability of language and relevant linguistic means for different communication situations.

At the final stage 3, students at secondary schools already know a lot about how language works, so they can focus more on specific types of word combinations, such as terminology (based on academic texts), set phrases and idioms (especially those used in contemporary texts, such as newspapers) and typical grammatical/lexical patterns (also in comparison to foreign languages based on parallel corpora).

In addition to this theoretical model, the poster will suggest sample activities and provide examples of real-life application of the corpus-based approach for each level.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Age group</th>
<th>Objectives</th>
<th>Keywords</th>
<th>Corpus data</th>
</tr>
</thead>
</table>

S66 Lucie Chlumská


<table>
<thead>
<tr>
<th>Stage</th>
<th>Age Range</th>
<th>Objectives</th>
<th>Keywords</th>
<th>Data Sources</th>
</tr>
</thead>
</table>
| 1     | 8–10 (lower primary) | • to show how language works  
• to teach about a word’s behaviour in text  
• to stress the importance of word combinations | context usage combinations | • Children’s literature (fiction)  
• Favourite books’ corpus (Harry Potter etc.) |
| 2     | 11–15 (upper primary) | • to discuss variants in language  
• to stress the differences between text types  
• to distinguish between registers (formal v. informal) | variability style relevance | • teen’s literature (mostly fiction)  
• carefully selected texts from non-fiction, newspapers and spoken registers |
| 3     | 16–19 (secondary) | • to provide a genuine view of contemporary language  
• to discuss terminology in different academic areas  
• to show parallels with other languages | heterogeneity terminology patterns | • general all-purpose corpora (written, spoken and parallel)  
• academic texts from different domains |

Table 1. The three-stage corpus-based model: objectives, keywords and data

References


The Longitudinal Corpus of Chinese Learners of Italian (LoCCLI)

The *Longitudinal Corpus of Chinese Learners of Italian* (LoCCLI) is the first large-scale longitudinal corpus of Italian as a second language (L2). It was started in 2016 and made available via CQPweb in mid-2017 ([https://www.unistrapg.it/cqpweb/](https://www.unistrapg.it/cqpweb/)). The LoCCLI includes written data collected from a group of Chinese learners of Italian, who attended Italian language courses for six months at the University for Foreigners of Perugia (Italy). It is made up of 350 essays written by 175 learners. Each of the learners contributed two essays to the corpus: one essay was written at the beginning of the course, and the other at the end of it. Importantly, a range of proficiency levels is represented in the corpus: through a placement test, learners were assigned to one of three proficiency levels: A1 (n=39), A2 (n=86), and B1 (n=50). The learners came from Mainland China and were between 17 and 33 years of age. On average, they spent 1.7 months in Italy before writing the first essay.

The same 175 students who wrote the first essay also wrote the second essay. The students who only wrote one of the two essays were not included in the corpus. Three comparable topics were offered: 1) My first impression of Italy and Italians, 2) My hobbies: what do I usually do in my free time, 3) My last holidays. The students were instructed not to write on the same topic more than once. Hence, all students chose two of the three topics. Finally, the students were taught by the same group of teachers at the same university. Thus, the corpus creation addressed the issues of homogeneity of topics, teaching style and learning environment (Gilquin, 2015).

The LoCCLI was post-tagged and xml-annotated, and its total size is 97,000 tokens. While the size cannot overall be considered large, the large number of participants assures a variety of learner texts. Moreover, as argued by Gut (2012), “a corpus with rich annotations and a standardised data format despite having a relatively small size offers numerous possibilities of testing previous concepts and claims in L2 acquisition research” (p. 20). The creation of the LoCCLI was mainly aimed at supporting a longitudinal investigation of vocabulary development in learner writing (Siyanova-Chanturia & Spina, in preparation). The aim of the poster is to present the general features of the corpus, together with some descriptive statistics on vocabulary use in a L2.

References


A variationist approach to learner corpus research: development of subject-verb word order in L2 learner Portuguese

This study adopts a variationist approach to corpus linguistics (Deshors & Gries, 2014; Gries, 2016), to investigate second language (L2) development over time. While variation in subject-verb word order has been extensively studied from a sociolinguistic perspective in both Spanish and Portuguese as first languages (L1s) (Cerrón-Palomino, 2017; Brown & Rivas, 2011; Mayoral Hernández, 2014; Spano, 2008), there has been little work on this language variable in learner production. In addition, most studies on language produced by learners of Portuguese focus on errors and not on language variation (Grannier & Carvalho, 2001; Mohr, 2011; Torres, Rodrigues, & Aluíso, 2014; Yokota, 2014). The present study fills these gaps in the research of acquisition of Portuguese as an L2 by investigating how the target language production of a Spanish-speaking learner develops over time based on the variation in subject-verb order use.

The small longitudinal corpus used for this investigation consists of 226 online journal entries written in a period of 19 months by a Spanish-speaking woman living in Brazil (total of 32,358 tokens). Information-structural, weight-related, animacy-related, and semantic factors (Gries, 2016) were coded for 409 instances of expressed syntactic subject. Logistic regression was then run with mixed-effect models in R (lme4 package, Bates et. al, 2017). Preliminary results show a longitudinal shift in the factors that constrain subject-verb word order in the learner’s production of Portuguese, which may be evidence of L2 development.

At the beginning of the learning period, syntactic subjects that are either clauses or lexical noun phrases favour postverbal subject order, while pronouns of all types favour preverbal subject. Subject animacy governs subject-verb word order in the beginning of the learning period as well, with non-human non-animate subjects favouring postverbal subject order. These findings match results for subject-verb word order in Spanish L1, where both subject animacy and subject type are the most important factor groups to govern subject-verb order (Brown & Rivas, 2011).

On the other hand, only verb type constrains subject-verb word order at the end of the learning period: unaccusative verbs favor postverbal subject order, while transitive and unergative verbs favor preverbal subject order. These results in turn match findings in Brazilian Portuguese L1, which show verb type as being the main predicting factor of subject-verb word order (Spano, 2008).
Overall preliminary findings can be seen as evidence to support the hypothesis that the variable rules of Portuguese L2 production by Spanish speakers tend to be similar to Spanish L1 at initial stages of acquisition, but at later stages the L2 learner production approximates Portuguese L1 use. Finally, based on these results, this poster presentation argues for classroom practices (especially those related to language assessment) that are centred on more authentic learner production.

References


A corpus-based study of semantic prosody in academic written English and its pedagogical application

Semantic prosody describes the linguistic phenomenon that some words or expressions are habitually associated with collocates with certain connotations or evaluations. It is important for non-native speakers to know the semantic prosody of a word or a phrase in order to make the correct choices of collocates and avoid producing unintentional irony. However, many L2 learners do not even have the awareness of semantic prosody (McGee, 2012). Although corpus analysis is useful to reveal the hiddenness of semantic prosody, the existing studies have focused on semantic prosodies of a few words or phrases, e.g., *cause, provide and bent on*. The studies are not systematic and many of them have ignored the roles of genre and context on semantic prosody. However, some studies (Hunston, 2007; Louw & Chateau, 2010) have shown that semantic prosody may vary across genres and contexts, e.g., the neutral prosody of *cause* in scientific articles in comparison to its negative prosody in general corpora (e.g., Hunston, 2007). Moreover, the findings have not been applied to pedagogical uses. Few dictionaries or other reference materials have explicit annotations for semantic prosody.

The present study aims to systematically investigate semantic prosody in academic written English and apply the findings to pedagogical uses. To achieve this end, a specialised corpus will be compiled using research articles published in the leading academic journals in the fields of natural sciences and social sciences. Each of the two sub-corpora will be divided into specific components based on the major sections in the research articles. A general corpus will be used as a reference corpus to compare semantic prosody across genres. A number of analyses (wordlists, keyword lists, n-gram analyses) will be used to extract the lexical items with marked prosodies and then the semantic prosody of each item will be identified through the frequencies of its collocates with negative or positive connotations. To explore the roles of genre and context on semantic prosody, comparisons of semantic prosodies will be made (1) between the specialised corpus and the general corpus; (2) between the articles in natural sciences and in social sciences; (3) in different sections of these research articles. A reference material of semantic prosody in academic written English will be compiled based on the findings of the present study.

The present study is expected to have three major outcomes: (1) an operational definition of semantic prosody and a methodology for identifying semantic prosody; (2) the systematic extraction of the words and phrases with marked semantic prosodies in academic written English and comparisons.
across genres and contexts; (3) the pedagogical application of the findings, i.e., the compilation of a reference material of semantic prosody for both teachers and learners. Aiming at systematically exploring this under-researched topic, the present study will make original theoretical and pedagogical contributions to language education.

References


**S93 Barry Kavanagh**

*Can corpora be useful in English Language Teaching in Norway?*

The research is article based, with an article covering each of: a survey of teachers’ familiarity with corpora, a proposed semester-long course in corpus linguistics for teachers, and proposed case studies of how some of the participants in the course subsequently make use of corpora in, or for, the classroom.

I make a distinction between teacher educator, in-service teacher, and pre-service teacher trainee. The aim of the research it to work with in-service teachers, to solicit their views, to introduce them to corpora, and to follow their corpus work into the classroom.

The first article: By the time of the conference, I will be able to present the data, and the article may also possibly have been written by then. This article aims to show the level of awareness of corpora in Norway among teachers of English. Data will be gathered through an online questionnaire. There are previous questionnaires in the literature (e.g. Breyer 2011: 161-2 and Leńko-Szymańska 2014: 268) that have been used to ascertain what respondents know about corpora, with for example the question, ‘Have you ever heard the term corpus and do you know what it is?’ (Lenko-Szymańska 2014: 268). At the time of writing (January 2018), the questionnaire is about to be released.

The second article: The aim is to make a group or groups of teachers familiar with corpora. I would incorporate corpus linguistics into a semester-long *Kompetanse for Kvalitet* (KfK) course for in-service teachers at my institution. Teachers can be introduced to corpora, concordancing software, and
uses for corpora in their teaching (indirect use in preparation, or direct use in the classroom). Establishing which material is the most user-friendly will require me to conduct in the first instance a pedagogical review of software. Teachers’ assignments in corpus linguistics in the course can be collected as data (subject to their permission). There is precedence for the usefulness of such data: Hüttner et al. (2008) and Breyer (2011: 175-185) analysed corpus assignments by teacher trainees.

The third article: In-depth case studies conducted over some months. A small number of volunteers would be found from the in-service teachers who receive the abovementioned introductory corpus education. These volunteer participants would then receive further support in corpora use. The corpus support would involve helping the volunteers to incorporate the use of corpora into their teaching. This means using a corpus or corpora in a way that is relevant to them and their pupils.

References


S106 Miki Hyun Kyung Bong and Masako Tsuzuki

A Study on the uses of adverbs (really & actually) and adverbial particles (up & on) based on the NICT JLE corpus data

Meanings of English words and phrases change through generations diachronically (language change) or sometimes may vary among different communities synchronically (dialects) and seem to diverge between English native speakers and English speakers of other native languages as a second (foreign) language (L2) or as a lingua franca (LF). Really and actually are both polysemous adverbials, sharing some senses. Actually has undergone changes from a verb phrase (VP) adverbial to an epistemic sentential adverbial and further to a discourse marker (Traugott & Dasher, 2002). Actually as a discourse marker expresses what follows will be unexpected; it signals that the speaker wishes to be a cooperative conversationalist (Lenk, 1988). Really has also undergone changes from a VP adverbial to on the one hand an intensifier and on the other hand an epistemic adverbial and further a discourse marker (Bong & Tsuzuki, 2017). Diachronic Changes in meanings or dialects have been studied
thoroughly, while various usages of English as a secondary (foreign) language (L2) or English as a lingua franca (ELF) by Japanese speakers of English has not been studied systematically. To explore Japanese uses of ELF or L2, we have examined thoroughly how *really* and *actually* are used by Japanese in English conversations taken place in Japan, using the NICT JLE corpus (NICT: National Institute of Information and Communications Technology, JLE: Japanese Learners of English). The corpus data are derived from the transcripts of the audio recorded samples (1,281 samples, 1.2 million words, 300 hours in total) of English oral interview test called ACTFL-ALC SST (Standard Speaking Test). We found that Japanese speakers of English as L2 (or ELF) tend to stick to one use each of *really* and *actually*. They use *really* as a response to the addressee’s statement, expressing their surprise or just signaling that they are listening. They use *actually* as an epistemic adverbial or else as a discourse marker with an adversative meaning. In other words, Japanese use *really* and *actually* almost exclusively to help them conduct English conversation in a smooth and cooperative way. We discuss L2 lexical learning strategies and divergence or gap between the senses of English words and phrases depicted in the dictionary (or L1 usages), and the senses of words and phrases in L2 usages (L2 lexicon). In addition, we further discuss pedagogical implications of these findings in teaching English in Japan.

**References**


---

**S126 Rezan Alharbi**

**Acquisition of lexical collocations: a corpus-assisted contrastive analysis and translation approach**

Research from the past 20 years has indicated that much of natural language consists of formulaic sequences or chunks. It has also been suggested that learning vocabulary as discrete items does not necessarily help L2 learners become successful communicators or fluent and accurate language users. Collocations (two-word combinations that frequently co-occur as one form of formulaic sequences) constitute an inherent problem for ESL/EFL learners. Non-congruent collocations (collocations that do not have corresponding L1 equivalents) are proven to be especially difficult to process and to acquire by ESL/EFL learners.
This study examines the effect of three instructional approaches on the passive and active acquisition of non-congruent collocations: 1) the non-corpus-assisted contrastive analysis and translation (CAT) approach, 2) the corpus-assisted CAT approach, and 3) the corpus-assisted non-CAT approach. To fully assess the proposed combined condition (i.e. the corpus-assisted CAT) and its learning outcomes, a control group under no-condition was included for a baseline comparison. Thirty collocations non-congruent with the learners’ L1 (Arabic) were chosen for this study. A careful method to operationalizing collocational congruency was applied in this research given the fact that Arabic is rich in polysemy and has different varieties and forms. The congruency decisions were made based on proficient bilinguals’ responses in two stages.

129 undergraduate EFL learners in a Saudi University participated in the study. The participants were assigned to the three experimental groups and to the control group following a cluster random sampling method. The corpus-assisted CAT group performed (L1/L2 and L2/L1) translation tasks with the help of bilingual English/Arabic corpus data. The non-corpus CAT group was assigned text-based translation tasks and received contrastive analysis of the target collocations and their L1 translation options from the teacher. The non-contrastive group performed multiple-choice/gap-filling tasks with the help of monolingual corpus data, focusing on the target items. Immediately after the intervention stage, the three groups were tested on the retention of the target collocations by two tests: active recall and passive recall. The same tests were administered to the participants three weeks later. The corpus-assisted CAT group significantly outperformed the other two groups on all the tests. These positive results of the proposed approach were accounted for by a synthesis of hypotheses in SLA. The ‘noticing hypothesis’, the ‘involvement load hypothesis’ and the ‘pushed output hypothesis’ accounted for the cognitive processes whereby the learners engage with the linguistic environments as the prelude for learning.

The discussion includes an evaluation of the three instructional conditions in relation to different determinants, dimensions and functions within the hypotheses.

S132 Klara Klimekova, Vishal Bhalla and Aisulu Rakhmetullina

The missing link between learners’ language use and their language learning – Elia

Language learners increasingly use their non-native languages through digital media, e.g., they read articles, watch tutorials or write emails (Sockett, 2014). These new digital practices not only represent a valuable resource for language learning but also reflect the new needs of language learners. Nevertheless, they remain largely unexplored and neglected by the educational practice which could be attributed to their private and heterogeneous nature (Meyers, Erickson & Small, 2013). Despite the obvious challenges, there is a pressing need to create language-learning solutions which would unlock
and maximize the potential of these emerging contexts of language learners and thus contribute to lifelong learning. Thanks to the recent developments in the corpus and computational linguistics enabling to automatically process natural language, the solutions are no longer only desired but also possible.

Adopting the design-based research methodology (Pardo-Ballester & Rodriguez, 2013), this ongoing doctoral project attempts to design, develop and evaluate an intelligent CALL assistant *Elia* which could support English learners in the context of their digital English use and automatically create additional learning materials to enhance their language learning. *Elia* takes the form of two interconnected parts, a browser plugin and a mobile app, both interacting with the learner and any digital text of choice. The main function of the browser plugin is to provide real-time individualized assistance in reading and writing in the target language by responding to the learner’s immediate needs, and also by proactively drawing the learner’s attention to the most useful language constructions given their long-term goals. Additionally, the browser plugin’s task is to monitor the learner’s interactions. Alternatively, the mobile app aims to provide additional learning opportunities in form of various tasks and exercises. These are based on the long-term learning objectives and previous experience of the learner and are constructed to map the incremental learning process to increase efficiency. Similarly to the browser plugin, it continuously tracks the learner’s performance in order to evaluate the progress and adapt the learning content.

On the one hand, *Elia* draws on different ready-made reference corpora and word lists. For the goal of academic vocabulary learning, the Academic Vocabulary List (Gardner & Davies, 2013) and the BAWE corpus (Nesi, 2011) is used to automatically extract collocations, synonyms, lexicogrammatical patterns and example sentences for individual words. These are then used either directly as a reference for assisting learners or indirectly for automatic generation and scoring of exercises (multiple-choice, gap-fills, free association, etc.). On the other hand, *Elia* also creates corpora for each learner separately from the tracked data of the learner’s interactions; one with the authentic digital language content the learner is exposed to and the other one from the language the learner produces. These are used for the assessment of the learner’s proficiency which, in turn, informs the future instruction, e.g., the prioritization of words and their different aspects or the selection of tasks stimulating different learning processes.

**References**


**S141 Shona Whyte**

*An audio-visual corpus of technology-mediated classroom language teaching: creating an open repository for CALL teacher education*

Historically, corpora have often been developed with an eye on practical applications, and as Boulton and Tyne (2014: 301) remind us, “in many cases, these applications were pedagogical in nature.” Cheng (2010) detects a shift in recent years from teaching to learning, with more attention given to tools and training for teachers to support learner use of corpora via data-driven learning. This goal of encouraging greater learner autonomy is mirrored in teacher education in what Farr (2010a: 621) calls “a cocoon of post-transmissive and post-directive approaches” which favour “independent and self-directed learning, and critical and reflective engagement.” A useful tool for teacher education in this respect is offered by teaching corpora, which O’Keefe, McCarthy and Carter (2007: 220) view as a unique application of corpus linguistics, since they focus not on “what we can learn about language use from a corpus” but rather on “what corpora can tell us about our own teaching.”

O’Keeffe and colleagues have used transcriptions from audio-visual teaching corpora to raise language awareness (O’Keeffe & Farr, 2003; O’Keeffe & Walsh 2012) and to support pedagogical development among trainee teachers (Farr 2010a, 2010b), using both discourse analysis and conversation analysis frameworks. Research in computer-assisted language learning (CALL) research has also investigated teacher corpora, using multimodal corpora to explore the semiotic dimensions of online language teaching, such as multimodal interactions via webcam (Cohen & Guichon 2016; Guichon & Wigham 2016; Holt, Tellier & Guichon, 2015). To date, however, little research has considered video corpora in CALL teacher education.

Our research in this area is built on two funded European projects supporting language teacher integration of classroom technologies. A first project collected short video clips of actual classroom practice with interactive technologies in a range of target languages at different age/proficiency levels. These practice examples were tagged for a variety of language, pedagogical, and technological features to create a searchable open repository for teacher education (Whyte, Cutrim Schmid, van Hazebrouck Thompson & Oberhofer, 2014). A follow-up project was designed to address technopedagogical concerns identified in the first corpus (Whyte, 2015), this time adopting a specific
pedagogical approach (task-based language teaching; TBLT), a wider range of technologies (mobile devices and videoconferencing), and longer videos showing edited teaching sequences.

This presentation analyses this second teaching corpus, ITILT 2, constituted by 117 video examples of learning activities prepared by 28 pre- and in-service teachers in 15 schools and universities in 5 European countries. The poster shows the background to the project and an overview of the teaching corpus created. The videos are analysed in comparison with the original corpus in terms of language, pedagogical, and technological features, as well as with respect to the new dimension (TBLT sequences). Secondary data on teachers and learner perspectives provides additional insight on this open learning project and the opportunities for teacher development afforded by this kind of teaching corpus.

References


Exploring lexical and syntactic complexity and sophistication in L2 writing among young writers in an EFL setting

The purpose of this study is to explore ways in which syntactic complexity is manifested differently in two genres of L2 writing among 8th graders in an EFL (English as a foreign language) setting. The motivation for such study is in twofold: 1) majority of research on L2 writing has geared towards university learners, as opposed to secondary school learners (Leki, Cumming, & Silva, 2008). Thus, it leaves a gap with respect to how young writers use different syntactic structures in L2 writing. 2) studies have shown there to be a correlation between syntactic complexity and writing quality (e.g. Crowhurst, 1980). Since different genres require specific syntactic structures that are closely related to its communicative goals (Beer & Nagy, 2011), learners’ writing should be scrutinized from various genres. The current study sought to compare the following two genres: descriptive and argumentative writing. These two genres were chosen since students are most familiar with these types of writing from classroom instruction. The first prompt asked to describe series of pictures to form a story and the second prompt was writing about the person they admire the most. A total of 68 students from three intact classes participated in this study and were given 20 minutes in which to complete each writing prompt. Writing samples were compiled and analyzed using TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity) Results are reported in relation to different proficiency levels (beginner, intermediate, advanced) and comparison of syntactic complexity is made across different proficiency levels. This study is a preliminary study to shed light on how diverse syntactic structures are used across two genres in a secondary school setting. Nevertheless, this study hopes to benefit both practitioners and researchers ways in which young L2 writers write and how classes can be facilitated to cater to the needs of students in secondary school setting.

References


S153 Xiao Wang

*Exploring data-driven learning in academic writing education: the acquisition of hedging devices*

The poster presentation will demonstrate the research design of my PhD investigation into the efficacy of data-driven learning (DDL) for learners to acquire hedging markers in English academic writing. DDL is a new teaching pedagogy that endorses teaching philosophies such as learner autonomy, inductive learning, and student-centeredness and the use of authentic, attested language data for language learning. The objectives of the study are to compare the learning effects of two types of DDL, i.e. the inductive DDL (hands-on, computer-based, “hard” version) and the deductive DDL (more conventional, paper-based, control group-like, “soft” version). The effects of DDLs as a learning tool and a cognitive tool will mainly be measured from two perspectives: 1) learning as a product - learners’ declarative knowledge of the target linguistic items in tests as well as appropriate usage in their written products in planned and unplanned settings; 2) learning as a process - learners’ internal cognitive process.

To fulfill the research objectives, mixed-methods research design will be used to guide the research. For the quantitative research strand, a quasi-experiment design will be applied to quantitatively measure learning performance of 70 research participants. For the qualitative strand, multiple-case study will be used to gain some in-depth qualitative data about the effects of types of DDL on learner-internal cognitive process. Furthermore, a questionnaire supplemented by a set of interviews and will be administered to explore learners’ attitudes towards their DDL experience.

The inductive group will receive a less-controlled, hands-on consultation on their self-compiled corpora with much less guidance. Statistical analyses will be conducted on performance data to examine if there is a significant inter- and intra-group difference before and after treatments. Test results are analysed to see whether different types of DDL effective for the learners’ improvement on pragmatic awareness and appropriate usage. Introspection data are collected through think-aloud protocol to see whether inductive DDL, serves as a better cognitive tool to promote higher-order cognitive strategies, and therefore, gives the learners an advantage in learning target features.

Feedback data will also be collected to explore the learners’ attitudes towards the DDL approach. This research is original in that it proposes to use the innovative pedagogy of DDL to teach L2 pragmatics. Hopefully, the discussion of this study may push the boundaries in three dimensions.
Theoretically, it is expected to contribute in some way to the establishment of a more effective model of DDL instruction. Methodologically, it will seek to explore the measurement of L2 pragmatics acquisition through a triangulation of research methods. Pedagogically, it aims to make a case for integrating corpus into L2 classroom. It also serves as a timely response to calling for more DDL research using self-compiled corpora for apprentice writers in EAP and ESP by Tribble (2015, p.59).

References


A corpus analysis of collocations in Vietnamese EFL learners’ writing

Knowledge of collocations (e.g. to make tea, heavy smoker) is considered important for language learners (Nation, 2013; Nesselhauf, 2003). However, collocations, especially verb-noun collocations, have been found to be difficult for foreign/second language (L2) learners (Laufer & Waldman 2011; Nesselhauf, 2003; Peters, 2016). There have been few empirical studies investigating the use of collocations of L2 learners and even fewer studies could be found on Vietnamese learners of English as a foreign language (EFL). It is, therefore, far from clear whether Vietnamese EFL learners encounter the same problems with collocations as other learners investigated in previous studies. For that reason, this mixed-method study will be conducted to examine the use of English lexical collocations by Vietnamese EFL learners at three levels of proficiency in free written production. A learner corpus comprised of about 50,000 words of letters, argumentative and descriptive essays written by Vietnamese EFL learners will be compiled for the investigation in this study. Lexical collocations will be extracted from this learner corpus using Wordsmith Tools software and checked against the *Oxford Collocation Dictionary for Students of English* by Lea, Crowther and Dignen (2002) to ensure their accuracy. These lexical collocations will then be categorised into different types based on Benson et al. (1986) to explore the most and least common types of collocation produced by the learners. In addition, the learners’ mistakes in producing collocations will be identified to find the most problematic type of collocation for the learners and later compared against their first language to investigate whether cross-linguistic interference, i.e. the negative influence of first language, can be found. Comparisons will also be made among the three levels of proficiency to see whether there exist any differences in the learners’ use of collocations and any mistakes in collocation use remain consistent across the three levels. Finally, 15 learners will be involved in semi-structured interviews so that their perspectives on their own learning and use of collocations will be investigated, providing...
more profound insights into learners’ collocation learning and use. The findings of this study will make both theoretical and pedagogical contributions to the learning and teaching of collocations in Vietnam and other similar EFL contexts.

References


S159 Yolanda Noguera

**Salvage and rescue of submarines: a pilot study**

Maritime English is the entirety of all those means of the English language which, being advisable for communication within the international maritime community, contribute to the safety of navigation and the facilitation of the seaborne business (Trenkner, 2000:77).

Submarine English (SE) can be considered a subgenre within the Maritime English scope and, given the specificity of the field, there have been so far no studies focused on language patterns used in SE contexts. Bearing in mind the difficulties to compile a submarine English corpus, we took as a starting point the non-restricted publications of one subject taught at the Submarine School in Cartagena, Spain: Salvage and Rescue of Submarines. We will use this pilot study to discuss the possibilities of these materials for DDL and examine the most outstanding lexical profiles and language patterns.

Sketch Engine (Kilgarrif et al. 2004) analysed lexical parameters and noted that the five more frequent words in our pilot study are acronyms. Different grammatical relations were examined and found a tendency for these words to occupy subject slots in sentences.

Biber et al. (1999) stated that acronyms, as noun post-modifiers, gave the technical reference or explanation of the previous nouns in academic and news registers. However, in their examples they
generally appear between brackets. On the contrary, in this corpus they are independent content words with different collocational relations with pre-and post- modifiers.

As a second stage, this analysis will be used as the basis for some vocabulary building activities (fill the gaps, multiple choice...). Following Nation and Schmitt (2008) who exposed the need to know about 97% percent of the words in the text in order to read a text properly we could assume that it is impossible to read this register without the knowledge of these acronyms and their collocates.

References


---

**S165 Francesca Poli**

*Teachers’ perceptions of students’ errors: a learner corpus analysis*

The exploration of students’ mistakes falls under the study of error analysis (Corder, 1967), which, despite debate in the literature (cf. Schachter, 1974; Schachter and Celce-Murcia, 1977), is an important part of the investigation into learner language. Teachers’ perception studies have often focused on the recognition of errors by native vs non-native teachers (cf. Sheorey, 1986; Hyland & Anan, 2006 and Rao & Li, 2017), but does the perception of students’ errors match the reality of their mistakes?

To investigate this, we compiled a written learner corpus of Italian learners of English (100,000 tokens) covering four different stages of learning, middle school, high school, undergraduate, postgraduate. No evaluation was made as to the learners’ proficiency. The corpus was manually tagged for errors, the teachers’ perceptions of their students’ errors were sought through a questionnaire and compared against the errors the corpus contained, identified via the tags using AntConc (Anthony, 2014).

The questionnaire was sent to the seven teachers who had provided the texts and it included a range of questions, such as the number of years they had taught, the grammar topics they had covered and the
kind of errors they thought their students make. The present study concentrates on a part of this corpus, i.e. the high school sub-corpus, the largest of the four (30,000 tokens). Four grammatical categories were selected for this study: the prepositions at, on, in, the present perfect tense, the past simple tense and word order. These had been taught by the teachers according to their responses in the questionnaire.

Each text was analysed in terms of number of prepositions used and number of wrong occurrences, as well as number of sentences per text and number of word order errors. The intermediate results on the prepositions and word order show that one teacher had correctly estimated students would make errors with prepositions and word order (e.g. 23.07% of instances of at are wrong, 10.15% of sentences contain word order errors), a second teacher had predicted errors with prepositions and other categories, but not word order, which is present in her students’ texts (16.66% on average for prepositions and 7.26% for word order), and lastly one teacher had not anticipated errors with the prepositions and word order, though the corpus suggests these may be areas of weakness for the students (on average for the prepositions at, on, in, 9.22%, 5.55% and 4.04%, and 16.11% for word order).

From the analysis of the corpus, it seems that the teachers’ perceptions generally correspond to the students’ mistakes when it comes to prepositions – two teachers out of three in the high school sub-corpus indicated these as potential problems – but as far as word order is concerned, they failed to identify it as a problematic area for their students: only one teacher anticipated word order errors in her students’ texts.

References


---

**S171 Xin Xu**

*Data-driven learning for lower-level ESL learners: a quasi-experimental study in Southern China*

The majority of English as Second Language (ESL) learners in contemporary China are rural students who are disadvantaged in terms of educational resources and opportunities. Research into the provision and practice of language teaching in this key area – minimising disadvantage to students – is urgently needed. Previous research results briefly sketched out from successful instantiations of Data-driven learning (DDL) – a learning approach driven by accessing linguistic data through corpus – are highly encouraging. However, traditional DDL approaches require computer to access data, most underprivileged high-schools do not have steady access to computer. Thus, paper-based DDL materials – on-line data print-outs – is more suitable for this study. It aims to evaluate and understand the effectiveness of paper-based Data-driven learning (DDL) materials on learners’ vocabulary acquisition in a typical underprivileged high school in southern China.

This study adopts a Critical Realist (CR) approach because other approaches such as quasi-experimental method only allows researchers to claim one intervention to be successful without learning anything about why it works. CR is in a better position to give a contextual explanation of how DDL works in a specific context. A classroom of 58 ESL learners in a typical high-school in the rural county is selected to be research participants. Quantitative data will be collected from pre-test, post-test and delayed post-test. Qualitative data should be complemented with participants review and interviews to gain insights into the research aim. If the general results were to be positive, it would enable paper-based DDL to reach a wider audience in China and beyond.

**References**


Aleksandra Swatek

Building the Corpus of Online Academic Spoken English (COASE)

The corpus of online academic spoken English (COASE) is being currently developed for the purpose of examining formulaic language used in online video lectures. The corpus consists of orthographic transcripts of over 8000 monologic online video lectures from Khan Academy platform. This platform is one of the most popular free and open education sites targeting students at all levels of education, with over 60 million registered users and many more who use the resources without registration. The corpus was obtained through API (application program interface) of Khan Academy platform, which allows for automatic and fairly easy collection of large collection of data.

The corpus can be used for analysing language used in teaching concepts across various academic levels from elementary school to graduate study. The transcripts are annotated with information about the speaker, discipline, field, and topic of the lecture. These video lectures are short, varying in length between 2-10 minutes. Currently, the corpus contains 8,841,666 words. The key disciplines represented in the corpus are: mathematics (37%), medicine (19.2%), science (12.8%), humanities (10.1%), economics (4.8%) and others. There are 78 speakers in the corpus, but one speaker (Sal Khan) authored 46% of the material. This poses challenges for ensuring balance and representativeness of the corpus.

The aim of this poster is to:

• Introduce the corpus and present preliminary results on the study of formulaic language used for teaching
• Describe the methods used in data collection, specifically harnessing readily available transcripts from MOOC and other online courses
• Discuss issues related to ensuring the corpus is balanced and representative in terms of speakers and disciplines
• Discuss potential for further research
Workshops

W1 Niall Curry and Olivia Goodman

Using the Cambridge learner corpus to develop learning materials

Given the nature of TaLC, the interests of its members and participants, and the constant need to drive for better applications of corpus research to language teaching and learning, this workshop allows Cambridge University Press to delineate the processes involved in the development of our corpus-informed materials. Focusing on the written Cambridge Learner Corpus (CLC) the workshop demonstrates the impact these data can have on materials development while also learning from TaLC members about how this can be done better, through the practical analysis of the data.

Beginning with an introduction to the 30 million word English language learner corpus, the workshop addresses the opportunities within and limitations of the data as well as the functionalities available through the chosen corpus analysis software: Sketch Engine. This is to equip participants with knowledge of the data and tools, so that findings can be successfully interpreted. Next participants are given temporary exclusive access to the complete CLC. The CLC will then be analysed where the practical elements of workshop are structured around the editorial challenges that materials writers, editors and publishers face when engaging with learner corpora for the development of language teaching and learning materials for ELT contexts.

Those attending this workshop will learn about corpus linguistics for materials developments, the CLC, how they can access it and how to get involved if they want to work with CUP on this type of work. They will also get an insight into publishing language learning materials and how corpora are at its core. Participants need to bring laptops.

W2 Olga Vinogradova, Stefania Spina, Luciana Forti, Ivan Torubarov, Nikita Login

Error annotation in learner corpora: tools and applications in English and Italian

In the first part the participants of the workshop will get acquainted with REALEC, the collection of English essays written by Russian university students, with its error classification scheme and the main principles for annotating errors in BRAT. A short video of the text annotation will be demonstrated, followed by interactive exercises in annotation, and a short competition for the best annotation will complete this part. Then the attendants will apply a test-maker – a computer tool for generating testing questions from the errors annotated in the corpus. The participants of the workshop will get a bank of automatically generated questions, edit them, do a test, and analyse its results. The
third stage will involve getting automated feedback on an uploaded text by applying REALECInspector – a tool that compares some features of this text with corresponding features of similar texts in the corpus. The participants will compare text inspection for the texts of two genres and for the texts of different writing proficiency.

In the second part of the workshop, a system for the annotation of Italian collocation errors in learner texts will be presented and discussed. The error annotation is performed on the Longitudinal Corpus of Chinese Learners of Italian (LoCCLI). This part of the workshop will specifically focus on the following key aspects of error annotation, which are particularly challenging in the case of collocations:

- choosing target hypotheses;
- coherently assigning categories to collocation errors;
- interpreting recurring collocation errors.

The interest of the workshop lies mostly in the different perspectives from which the practice of error annotation is approached: in the first part, the annotation is aimed at integrating and improving CALL systems, and the specific uses of a learner corpus for EFL/ESL instructors are in focus in the presentation; in the second part, it is intended to detect and analyse a specific area of difficulty for learners, with the aim of providing useful data to improve our knowledge on learner behaviour. Another benefit of the workshop lies in the focus on two different second languages, which can raise interesting problems and propose new challenges for those interested in error annotation.

The use of laptops is encouraged for this workshop. There are no requirements regarding the operating system, and there is no need to install special software, but everyone will need a modern web browser, preferably Google Chrome. No special requirements as far as the choice of Windows or Mac is concerned.

References
W3 Silvia Molina, Plaza María del Mar Robisco, Martín Verónica Vivancos Cervero, Ana Roldán-Riejos, Paloma Úbeda Mansilla

Phraseological and corpus-based study of scientific and technical discourse: English and Spanish

The main aim of this workshop is to showcase explorations on specialised discourse through corpus-based analysis in English and Spanish. The phraseology of scientific and technical discourse is an area that has received little attention previously, although it has been proved that, in many cases, words do not appear in isolation, on the contrary, they show a tendency to combine and team up (Aguado 2007, Pérez-Llantada 2013, Roldán-Riejos & Molina 2016). The procedure referred to source extraction will be detailed and data collected from online common written academic and professional genres will be detailed. It is assumed that lexical clusters or collocations co-occur with single lexical items in the configuration of scientific and technical discourse. Hence, we focus on how the lexical components of such clusters are arranged according to context. We will consider meaning construction and the ways in which mechanisms like metaphor and metonymy are encapsulated both conceptually and linguistically. Additionally, their function will be also analysed. Cross-linguistic Spanish-English/English-Spanish analyses of discourse are finally conducted. Possible practical applications of this model are considered from three different angles: 1. To design specialised dictionaries and data banks 2. for translation and 3. for learning purposes in the classroom or online.

It is expected that the the phraseology of specialised language and its scientific compilation and analysis in two languages will be relevant for TaLC participants. Those attending this workshop will learn that phraseology has a relevant occurrence in specialised language and also how this phraseology can be compiled and analysed for research and learning purposes.
Participants need to bring either a tablet or a laptop to work with practical examples. Windows software is fine.

References


W4 Laurence Anthony

Data-Driven Learning (DDL) in the Technical Writing Classroom

This workshop will offer and hands-on demonstration and explanation of Data-Driven Learning (DDL) tools and techniques applied in the area of technical writing. The workshop will open with a brief review of research that demonstrates the effectiveness of the Data-Driven Learning (DDL) approach. Next, various corpus data sources and easy-to-use software tools will be introduced that can facilitate the DDL approach with learners of varying levels of familiarity with computers. In this part of the workshop, participants will explore ways to guide learners in the building of custom, discipline-specific corpora using AntCorGen (Anthony, 2017) and AntFileConverter (Anthony, 2017). They will also receive guidance on how to introduce the AntConc (Anthony, 2017) corpus toolkit to learners and help learners discover useful language patterns for use in discipline-specific writing contexts. The workshop will finish with a general discussion on the challenges and limitations of DDL and strategies to overcome the common problems associated with incorporating DDL into a regular technical writing course or program. By the end of the workshop, it is hoped that participants will have new ideas for DDL research, effective techniques for introducing DDL to learners, and useful suggestions for becoming involved in DDL software development projects.

Participants will learn about:
1) the evidence that exists to support the case for introducing Data-Driven Learning (DDL) into the technical writing classroom.
2) corpus data and software tools that facilitate the DDL approach.
3) teaching strategies that can guide learners in the building of discipline-specific corpora.
4) teaching strategies that can help learners discover useful language patterns for use in discipline-specific writing contexts.
5) the challenges and limitations of DDL and strategies to overcome common problems associated with DDL.
6) ways to become involved in DDL software development projects.

Participants are encouraged to bring their own laptops. Before the workshop, indications will be provided as to the software that they’ll need to download and install.

References

W5 Adriana Picoral, Shelley Staples, Ji-young Shin, Aleksandra Swatek

Exploring variation and intertextuality in L2 undergraduate writing in English: Using the Corpus and Repository of Writing Online Platform for research and teaching

The Corpus and Repository of Writing (Crow) is the first online platform to integrate a corpus of texts produced by L2 writers of English in their early undergraduate years with a repository of the pedagogical materials that these students use to create their texts and formulate their arguments. It represents a collaboration across five U.S. institutions and includes over 30 researchers (see writecrow.org for more information). The Crow corpus currently contains over 8 million words from two institutions (Purdue University and University of Arizona) and represents over ten genres of university level scholarly and non-scholarly writing (e.g., rhetorical analyses, literature reviews, literacy narratives, proposals, research papers, genre analyses, tweets, Facebook posts). It also contains rich metadata (e.g., country of origin, TOEFL scores, students’ majors, drafts) that can be selected within the online platform. Individual students’ written work can also be tracked across the different genres they produce. A beta version of our online platform will be available in May 2018, and TALC participants will be the first to interact with the platform outside our research team. Our goal for this workshop match the conference aims: to promote the creation of corpus-informed teaching materials based on Crow, integrating a discussion between practitioners and researchers.
Workshop attendants will engage in corpus research and create corpus-based classroom activities. They will learn how to:

A) Search the original corpus using conventional corpus tools (e.g., concordancer)
B) Refine corpus searches by locating texts in a genre of interest
C) Find assignment sheets or lesson plans related to the same genres of interest in the repository
D) Filter search results through available metadata (e.g., the major of the author at the time of data collection, country of origin, and TOEFL scores)
E) Add selected texts to a working corpus (i.e., a customized collection of texts built directly on the Crow platform by each participant)
F) Use Crow’s built-in tools to code for intertextuality

Participants will also have time to work in pairs or small groups to develop pedagogical materials using Crow for their own classroom context.

After our workshop, participants will be able to: 1) use Crow to explore linguistic and rhetorical features, including those related to intertextuality in the texts of L2 writers of English, 2) discuss how information from these searches can be further developed for research and inform language teaching and, 3) develop classroom activities based on the corpus and repository available in the Crow platform.

Participants need to bring their own devices, There is no need to download or install anything in advance. Everything used in the workshop will be self-contained in the online platform. Participants can use their tablets, and Crow works across different platforms (Windows and Mac).

W6 Vaclav Brezina, Dana Gablasova, Irene Marín Cervantes

Using corpora to teach sociolinguistics: A practical workshop

Teaching sociolinguistics both in the L1 and the L2 contexts presents a challenge. It involves pedagogical considerations about how to best draw students’ attention to the fact of linguistic variation (cf. Meyerhoff 2011; Brezina & Meyerhoff 2014) as well as practical concerns such as finding resources (data, teaching materials etc.) suitable for classroom use. At the same time, sociolinguistic awareness and competence are becoming increasingly important both in L1 (e.g. A level English Language, AQA 2014) and L2 contexts (e.g. Geeslin & Long 2014; Sung 2016). This workshop offers a discussion of the role of sociolinguistics in the classroom as well as practical examples of corpus tools and materials designed to help analyse and teach about language variation. The following topics will be covered:
• Discussing language variation in the classroom
• Presenting and visualising corpus data
• Useful corpus tools
• The Spoken BNC 1994 and 2014 and other corpora of contemporary British English
• Creating teaching materials

The workshop will be of interest to both researchers and practitioners. The participants will be introduced to corpus tools such as #LancsBox and BNC Lab, both recently developed at Lancaster University, which allow efficient exploration of corpora in the classroom. The workshop will offer an introduction to new data analysis and visualisation techniques, which the participants will be able to apply in their specific educational contexts. The workshop will also focus on the development of corpus-based teaching materials – multiple practical examples of effective teaching materials will be provided.

Participants will need to bring their own laptops. The computers need to be able to download and run java programs such as #LancsBox http://corpora.lancs.ac.uk/lanesbox/download.php

References
Author index

Aas

Hege Larsson, 57

Catherine

Abe

Brissaud, 45

Mariko, 70

Cebron

Ackerley

Neva, 77

Katherine, 89

Cervantes

Alghamdi

Irene Marín, 154

Ayman, 72

Cervero

Alharbi

Martín Verónica Vivancos, 151

Rezan , 136

Charles

Alruwaili

Maggie, 34

Awatif , 73

Chen

Anthony

Alvin Cheng-Hsien, 67

Laurence, 152

Chen

Atwell

Meilin, 53

Eric, 38

Chlumská

Bhallal

Lucie, 128

Biel

Chujo

Łucja , 113

Curry

Bong

Kiyomi, 79

Miki Hyun Kyung, 135

Darwish

Boulton

Hosam, 19

Alex, 8

De Jong

Brezina

Franciska , 92

Vaclav, 105, 154

Diemer

Brunner

Stefan, 28

Marie-Louise, 28

Dirdal
<table>
<thead>
<tr>
<th>Name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klimcikova</td>
<td>Login</td>
</tr>
<tr>
<td>Klara, 137</td>
<td>Nikita, 149</td>
</tr>
<tr>
<td>Kobayashi</td>
<td>Loock</td>
</tr>
<tr>
<td>Yuichiro, 70, 79</td>
<td>Rudy, 115</td>
</tr>
<tr>
<td>Kondo</td>
<td>Lozano</td>
</tr>
<tr>
<td>Yusuke, 70</td>
<td>Cristóbal, 30</td>
</tr>
<tr>
<td>Kreyer</td>
<td>Mansilla</td>
</tr>
<tr>
<td>Rolf, 69</td>
<td>Paloma Úbeda, 151</td>
</tr>
<tr>
<td>Kunilovskaya</td>
<td>Marinov</td>
</tr>
<tr>
<td>Maria, 103</td>
<td>Sanja, 51</td>
</tr>
<tr>
<td>Lazic</td>
<td>Mark</td>
</tr>
<tr>
<td>Katarina, 76</td>
<td>Geraldine, 67</td>
</tr>
<tr>
<td>Lee</td>
<td>Mendikoetxea</td>
</tr>
<tr>
<td>Pinshuan, 43</td>
<td>Amaya, 30</td>
</tr>
<tr>
<td>Leńko-Szymańska,</td>
<td>Mickiewicz</td>
</tr>
<tr>
<td>Agniezska, 62</td>
<td>Adam, 18</td>
</tr>
<tr>
<td>Lew</td>
<td>Mizumoto</td>
</tr>
<tr>
<td>Robert, 18</td>
<td>Atsushi, 79</td>
</tr>
<tr>
<td>Lim</td>
<td>Molina</td>
</tr>
<tr>
<td>Dong Ok, 141</td>
<td>Silvia, 151</td>
</tr>
<tr>
<td>Lin</td>
<td>Morgoun</td>
</tr>
<tr>
<td>Huifen, 43</td>
<td>Natalya, 103</td>
</tr>
<tr>
<td>Liou</td>
<td>Murakami</td>
</tr>
<tr>
<td>Hsien-Chin, 25, 26</td>
<td>Akira, 70</td>
</tr>
<tr>
<td>Liu</td>
<td>Nesi</td>
</tr>
<tr>
<td>Szu-Yu, 25</td>
<td>Hilary, 11, 75</td>
</tr>
<tr>
<td>Liu</td>
<td>Nirwan</td>
</tr>
<tr>
<td>Tanjun, 41</td>
<td>Sharma, 18</td>
</tr>
</tbody>
</table>
Noguera
   Yolanda, 144

O’Donoghue
   John, 47

O’Keeffe
   Anne, 7

Oghigian
   Kathryn, 79

Pace-Sigge
   Michael, 119

Park
   Hyeson, 90

Pecorari
   Diane, 84

Pérez-Paredes
   Pascual, 67

Perraud
   Sylvain, 124

Perri
   Francesca, 86

Petrovic
   Maja Milicevic, 76

Picoral
   Adriana, 131

Picoral
   Adriana, 153

Poli
   Francesca, 145

Ponton
   Claude, 45

Prado
   Malila Carvalho De Almeida, 101

Prinz
   Michael, 106

Quinn
   Daniel, 75

Rakhmetullina
   Aisulu, 137

Rees
   Geraint Paul, 18, 64

Roberts
   Jonathan C, 18

Robisco
   Plaza Maria del Mar, 151

Roldán-Riejos
   Ana, 151

Rollinson
   Paul, 30

Rørvik
   Sylvi, 57

Satake
   Yoshiho, 20

Schaeffer-Lacroix
   Eva, 117

Seely
   Jane, 12
<table>
<thead>
<tr>
<th>Name</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seracini Francesca</td>
<td>110</td>
</tr>
<tr>
<td>Shin Ji-Young</td>
<td>108</td>
</tr>
<tr>
<td>Sing Christine</td>
<td>96</td>
</tr>
<tr>
<td>Siyanova-Chanturia Anna</td>
<td>130</td>
</tr>
<tr>
<td>Spina Stefania</td>
<td>130, 149</td>
</tr>
<tr>
<td>Staples Shelley</td>
<td>59, 153</td>
</tr>
<tr>
<td>Sugisaki Kyoko</td>
<td>106</td>
</tr>
<tr>
<td>Swatek Aleksandra</td>
<td>148, 153</td>
</tr>
<tr>
<td>Tono Yukio</td>
<td>81</td>
</tr>
<tr>
<td>Torubarov Ivan</td>
<td>149</td>
</tr>
<tr>
<td>Tosqui-Lucks Patricia</td>
<td>101</td>
</tr>
<tr>
<td>Tsuzuki Masako</td>
<td>135</td>
</tr>
<tr>
<td>Turner Chris</td>
<td>11</td>
</tr>
<tr>
<td>Twardo Sylwia</td>
<td>121</td>
</tr>
<tr>
<td>Tyne Henry</td>
<td>33</td>
</tr>
<tr>
<td>Vincent Benet</td>
<td>75</td>
</tr>
<tr>
<td>Vinogradova Olga</td>
<td>149</td>
</tr>
<tr>
<td>Vu Duy Van</td>
<td>143</td>
</tr>
<tr>
<td>Vyatkina Nina</td>
<td>36</td>
</tr>
<tr>
<td>Wang Xiao</td>
<td>142</td>
</tr>
<tr>
<td>Zhaozhe, 96</td>
<td></td>
</tr>
<tr>
<td>Weisser Martin</td>
<td>55</td>
</tr>
<tr>
<td>Werner Valentin</td>
<td>13</td>
</tr>
<tr>
<td>Whyte John</td>
<td>65</td>
</tr>
<tr>
<td>Wickens Paul</td>
<td>111</td>
</tr>
<tr>
<td>Williams John</td>
<td>65</td>
</tr>
<tr>
<td>Wold Stephanie H. G.</td>
<td>82</td>
</tr>
<tr>
<td>Wolfarth Claire</td>
<td>45</td>
</tr>
<tr>
<td>Wolk</td>
<td></td>
</tr>
</tbody>
</table>
Christoph, 93
Wu
Yi-Ju Ariel, 94
Wunderlich
Claudia, 46
Xu
Xin, 147
Yang
Tzu Wei, 26
Zhang
Danyang, 126