TaLC2020

Abstracts
Information on strands and synchronous discussions ................................................................. 5

Strand [1]: Data-driven learning ........................................................................................................ 5
Strand [2]: Specific/academic purposes, specialised language and specialised corpora .............. 5
Strand [3]: Acquisition, interlinguistic/comparative studies, translation, linguistic approaches ..... 5
Strand [4]: Classroom teaching/learning, teacher training, tools and resources for teachers/learners, planning and programmes ................................................................................. 5
Strand [5]: Writing, discourse, text .................................................................................................. 5
Strand [6]: Posters ............................................................................................................................ 6

Plenary abstracts ............................................................................................................................... 7

Maggie Charles. From take-up to take-off? A longitudinal view of data-driven learning in English for academic purposes .............................................................................................................................. 7
Mark Davies. The expanded one billion word COCA corpus: new features for researchers, teachers, and learners ........................................................................................................................................ 8
François Maniez. Tracking diachronic variation in the oral components of English corpora ............ 11
Pascual Pérez-Paredes. Rethinking learning in Data-driven learning ............................................... 12

Paper abstracts ................................................................................................................................. 14

Katherine Ackerley. ESP corpus construction and analysis to produce syntactically reduced, persuasive texts [5] .............................................................................................................................................. 15
Alex Boulton. KWIC and easy? Students building their own corpus for non-linguistic goals [1] ...... 21
Nigel Bruce and Peter Storey. Big corpora development for disciplinary DDL in an EFL-medium university setting [1] ............................................................................................................... 22
María Luisa Carrión-Pastor. The use of corpora for teaching multimodal metadiscourse in academic English [2] .......................................................................................................................... 23
Meilin Chen and John Flowerdew. An Investigation of PhD Students’ use of corpora in facilitating their writing for publication purposes [5] .................................................................................................................. 25


Peter Crosthwaite and Luciana Luciana. “Corpus missionaries, spreading the word”: Introducing Data-Driven Learning to teacher trainees in South East Asia [1] .......................................................................................................................... 28

Niall Curry. Using contrastive linguistics and learner corpus research to inform the teaching of academic metadiscourse: a corpus-based study of target, first, and learner language corpora [2] 29


Luciana Forti. Data-driven learning effects in the development of L2 phraseological competence: the combined role of semantic transparency and frequency [1] ........................................................................................................ 33

Ana Frankenberg-Garcia, Simone Sarmento, Paula Tavares Pinto and Ana Bocorny. A data-driven learning approach to supporting the dissemination of research from non-English speaking institutions [2] ............................................................................................................................................ 35


Sandra Götz, Christoph Wolk and Katja Jäschke. Fluency across L1s, Task Types and Learning Context Variables: A learner corpus-based perspective [3] ............................................................................................................................................ 38


Yuying Hu and Ziwei Cheng. An Initial Investigation into Discourse Connectives in Written Texts by Chinese and Native ESP Learners [5] ........................................................................................................... 41


Fu-Ying Lin and Kuan-Yu Shen. The Grammar Patterns of The Spoken Academic English (of MOOCs) [2] ......................................................................................................................................................... 51
Nikita Login, Ksenia Pospelova, Ekaterina Afanaseva and Elizaveta Ershova. A testing tool driven by learner corpus data and the platform for its use [4] ................................................................................................ 56
Philippe Millot. Building specialised corpora to conceptualise learners’ specialised needs: The case of a corpus in records and information management [2] .......................................................................................... 57
Mehmet Nacar and Adnan Biçer. A corpus-based study: Complement clauses in written and spoken outputs of Turkish EFL learners [3] ........................................................................................................ 58
Aleksey Novikov. Lexico-grammatical development of L2 Russian learners [3] ........................................................................................................ 60
Maria Oberyukhtina. To what extent is it possible to find some generalizations between the development of auxiliary verbs in English and the stages of language acquisition by L1 (or L2) learners? Can such generalizations serve as diagnostics for language change? [3] ........................................................................................................ 61
John O’Donoghue. An exploratory investigation into the way in which L2 undergraduate writers perceive the nature of academic writing and the cultural factors determining such perceptions [5] ................................................................................................................................. 62
Anne O’Keeffe, Pascual Pérez-Paredes and Geraldine Mark. Aligning usage-based theory and Data-Driven Learning [1] ......................................................................................................................................................... 63
Tristan Purvis. Academic English Learners’ Corpus of Writing Students in Nigeria [5] ........................................................................................................ 68
Tugba Simsek and Cem Can. Pre-service language teachers’ perceptions on teachers’ duties in material development: can a corpus literacy course change their perceptions? [4] ........................................................................................................ 69
Bruna Sommer-Farias, Adriana Picoral, Aleksey Novikov and Shelley Staples. Teachers’ Perceptions of Interactive DDL Using a Multilingual Learner Corpus [1] ........................................................................................................ 69
Shelley Staples. Building DDL materials with teachers of second language writing: Bridging the research/teaching gap [1] ......................................................................................................................................................... 70
Aleksandra Swatek. Direct Hypothetical Reported Speech in Math Instructional Videos: A Corpus-based Study [4]........................................................................................................................................ 71

József Szakos and Ulrike Glavitsch. “Genomes of corpus” and the problem of “genres”: Re-configuring elements of corpus linguistic solutions in teaching languages, taking readylingua as an example [4]................................................................................................................................. 72

Szilvia Szita. Overwhelming, time-consuming, user-unfriendly... and now what? Teacher training for the successful implementation of a corpus-informed methodology [4]............................................................................................................................ 73

James Thomas. Learning language and content from a one-text corpus [4]............................................................................................................................ 74

Biagio Ursi. Corpus-based resources in conversation: Learning with the multimodal concordancer of the FLEURON database [4]........................................................................................................................................... 74

Benet Vincent, Hilary Nesi and Daniel Quinn. Exploring teacher engagement with an open DDL resource [1]........................................................................................................................................... 75

Olga Vinogradova, Anna Viklova, Darya Overnikova, Darya Matyash and Mikhail Paporotsky. “Genre is a useful concept only when used evaluatively not descriptively”: Observations over differences between two genres of essays in learner corpus [5] ............................................................................................................................ 76

Karin Whiteside. Writing development across an undergraduate degree: A contrastive corpus analysis of student writers’ first and third year Essay texts from the disciplines of History and Politics & International Relations (PIR) [5] ............................................................................................................................ 79

Yi-ju Ariel Wu and Chi-Ting Robert Tsai. Exploring the Effects of Data-Driven Learning in Learning Legal Vocabulary of International Law [1]........................................................................................................................................... 81

Poster abstracts .................................................................................................................................... 83

Cathryn Bennett. Preliminary Findings of Corpus Literacy by In-service EFL Practitioners in Ireland [6] ........................................................................................................................................... 83

Aivars Glaznieks, Jennifer-Carmen Frey, Maria Stopfner and Lorenzo Zasina. LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English [6] ........................................................................................................................................... 84

Barbora Hrabalova. Web-based dictionary for learners of Czech: methodology and design [6]..... 85

Elżbieta Kaczmarska and Adrian Jan Zasina. Infrastructure of the Polish Learner Corpus [6] ........ 85

Wooji Park and Seok-Chae Rhee. The Durational Difference between Stressed and Unstressed Vowels as a Cue for English Proficiency of Korean Learners [6]........................................................................................................................................... 86

Juliane Schopf. Authentic spoken data and applied conversation analysis as a resource for the improvement of communicative competence in healthcare [6]........................................................................................................................................... 87

Stefania Spina, Luciana Forti, Filippo Santarelli, Valentino Santucci and Giuliana Grego Bolli. From corpus construction to automatic text difficulty classification for second language learners: developing MALT-IT2 [6]........................................................................................................................................... 89
Information on strands and synchronous discussions

In order to make the synchronous discussion sessions on the 16th and 17th of July as coherent as possible, papers are sorted into 5 “strands” according to common research themes. Strands are indicated with a number following the title (see the list below). There is no order of importance from 1 to 5; the numbering is totally random.

Obviously, many papers could fit quite comfortably into more than one strand, so do not be alarmed or distressed to find yourself wanting to participate in more than one discussion as there are bound to be overlapping concerns. Authors are encouraged to follow as much as possible and join in as often as possible. And don’t forget the online discussion pads (links on conference website) and the comment feature in Vimeo.

Each strand has a chair (or chairs) whose job it is to lead the discussion sessions. The starting point for discussion should be the asynchronous comment and discussion. All authors are encouraged to monitor their paper over the week starting on the 13th. Obviously, the more questions and answers are supplied asynchronously, prior to the live discussion sessions, the more content the chair and discussants have to feed into the live sessions.

Strand [1]: Data-driven learning
    Chaired by Peter Crosthwaite & Benet Vincent

Strand [2]: Specific/academic purposes, specialised language and specialised corpora
    Chaired by Karin Whiteside

Strand [3]: Acquisition, interlinguistic/comparative studies, translation, linguistic approaches
    Chaired by Olga Théophanous

Strand [4]: Classroom teaching/learning, teacher training, tools and resources for teachers/learners, planning and programmes
    Chaired by Rosamond Mitchell

Strand [5]: Writing, discourse, text
    Chaired by John O’Donoghue
Strand [6]: Posters
Chaired by John Williams
Plenary abstracts

Maggie Charles. From take-up to take-off? A longitudinal view of data-driven learning in English for academic purposes

Since its inception by Johns (1991) in his work with graduate students at Birmingham University, data-driven learning (DDL) has had a long and productive association with teaching English for Academic Purposes (EAP). Indeed, in a recent survey Tribble (2015) points out that EAP courses still remain the single most frequent site of DDL use; moreover, he is optimistic about the further adoption of corpus approaches, a view shared by other researchers (see e.g. Boulton and Cobb, 2017; Chen and Flowerdew, 2018). However, despite some encouraging signs of increased DDL take-up, it is pertinent to ask just how extensive and deep-rooted the practice is among teachers and learners. Using data from the UK, this talk presents a longitudinal view of the adoption of DDL in EAP during the period 2009-2019. It focuses first on teachers, drawing on evidence of professional activity from the British Association of Lecturers in EAP (BALEAP), the UK association which promotes itself as ‘the global forum for EAP professionals’. The take-up by learners is shown through research indicating the extent of corpus use at a large UK research university over a period of nine years: 2009-2017 (Charles, 2019). Taking both viewpoints into account, this paper will argue that DDL in EAP is not yet at the point of take-off. For teachers, there are several barriers to the adoption of corpus-based approaches, including lack of adequate exposure to DDL, lack of time to implement new approaches and lack of management support (Chambers, 2019; Davis, 2019). Insufficient time is also a key issue affecting learner take-up, along with the sporadic nature of their writing practice due to the many varied and competing demands of advanced academic study (Charles, 2014, 2017). The talk will conclude by referring to work on the adoption of innovation (Rogers, 2003) and technology in language learning (Bax, 2003, 2011) to suggest some strategies which could contribute to an increase in DDL take-up in the future.

References


Mark Davies. The expanded one billion word COCA corpus: new features for researchers, teachers, and learners

The Corpus of Contemporary American English (COCA) is perhaps the most widely used online corpus. In this presentation, I will discuss the important additions and improvements that were made to the corpus in early 2020. These include:

1. More recent: COCA now includes texts through Dec 2019, including about 50 million words of data from 2018 and 2019.
2. Size: COCA is now about one billion words in size (about double the previous size), which means that it provides much better data for medium and lower-frequency words and constructions.
3. New genres: COCA has been expanded from the existing five genres of spoken, fiction, magazine, newspaper, and academic, to include 370 million additional words of data from blogs, other web pages, and TV and movie subtitles.

As far as changes that directly impact teachers and learners, the following are new:

4. Frequency results now include links (for each matching phrase) to translations, pronunciation, images, videos, and web searches -- helping learners to move beyond the bounds of the corpus itself.
5. Browse and search. Learners can now browse through and search the top 60,000 words – by word form, part of speech, frequency, and even pronunciation, and then there are (for each word in the list) links to information for that word (see #6).
6. Extensive information for each of the top 60,000 words, including pronunciation, images, and videos; translation of the word into 50-60 different languages; detailed frequency and range information; related words; synonyms, hypernyms, and hyponyms; collocates (now grouped by part of speech); related topics (other words that co-occur in the 500,000+ texts); concordance lines; and clusters (the top 2, 3, and 4 word strings for a given word) – all with extensive links between each “word page”.

Hopefully, all of these new features will make COCA even more useful for teachers, learners, and researchers.
Despite the seminal contributions of 19th century European phoneticians to foreign language (L2) education methodology (e.g. Paul Passy and the founders of the *International Phonetic Association*, cf. Durand, 2017; Galazzi, 2002), the ‘pronunciation’ component in L2 education has almost always been overshadowed by its lexical and grammatical counterparts in the fields of Second Language Acquisition (SLA) and L2 educational methods. However, thanks to the work accomplished over the past 10 years in L2 phonetics and phonology, pronunciation has turned – according to some prominent scholars in the field – from being the “Cinderella” of SLA to being the “Belle of the Ball” (Derwing, 2019: 27), a view which must still find its way into the subfield of applied corpus linguistics and data-driven learning (DDL) and teaching. In fact “pronunciation” is not included in the language aspects (vocabulary, lexicogrammar, grammar, discourse) examined in the meta-analysis of corpus use in language learning carried out by Boulton and Cobb (2017: 380); it is absent from the introductory article in a special issue of *Language Learning and Technology* dedicated to corpora in language learning and teaching (Vyatkina & Boulton, 2017); and it seems to be only significantly mentioned once (Meunier, 2019) in a recent volume dedicated to data-driven learning “for the next generation” (Crosthwaite 2019).

Likewise, in the *TaLC2018* book of abstracts, only one paper (out of 173) examines pronunciation issues in an English for Specific Purposes context (Staples 2018), while another one simply mentions the pronunciation dimension in the description of its spoken corpus data collection protocol (Abe, Kondo, Kobayashi, Murakami & Fujiwara, 2018). Therefore, five years after the publication of the *Oxford Handbook of Corpus Phonology* (Durand, Gut & Kristoffersen, 2014) – which includes one chapter on corpus phonology and SLA (Gut 2014) – it seems to be the right time for a plenary talk on the topic of Pronunciation and Teaching and Language Corpora – especially as, in the meantime, international pronunciation-focused conferences such as *Pronunciation in Second Language Learning and Teaching* welcome, as plenary speakers, famous corpus linguists such as Douglas Biber (whose “former life” actually included articles on accent in Somali and contour tones, Biber 2019).

The core of this talk – adding an extra French flavor to the Perpignan-based *TaLC2020* – will rely on the output of the international and collaborative work carried out within the framework of the *PFC* project (phonology of contemporary French, Durand, Laks & Lyche, 2002), especially its education-oriented applied offshoots, namely *PFC-EF* (PFC for French education, Detey, Durand, Laks & Lyche, 2016) and *IPFC* (Interphonology of Contemporary French, Detey, Racine, Kawaguchi & Eychenne, 2016). After a short presentation of the historical foundations of the project and its resulting corpus (https://www.projet-pfc.net/), the talk will tackle three main issues: 1) the relevance of spoken corpora and DDL for L2 pronunciation teaching and learning; 2) the relevance of spoken corpora and DDL for language awareness education both in L1 and L2 at the pronunciation level and beyond; 3) the challenges embedded in the design of a spoken corpus-based L2 pronunciation education programme in the light of contemporary corpus-based and DDL-oriented pedagogy (Boulton & Tyne, 2014, Tyne et al. 2014).

While Biber (2019) tried to convince applied phonologists/phoneticians at *PSLLT2019* that they needed to know grammar to teach pronunciation, this talk will aim at highlighting the fact that corpus linguistics and pronunciation teaching are not so disconnected, especially when it is looked at from a psycholinguistic (hybrid models of speech processing, Nguyen, Wauquier & Tuller, 2009) and a
sociolinguistic (variationist approach of speech, Durand, Laks & Lyche, 2016) perspectives (Detey, 2009, 2010). This, in turn, should stress the importance of pronunciation skills (both in perception and in production) in the (initial) acquisition of the lexicogrammatical component of the L2 at stake, and therefore possibly convince applied corpus linguists that they cannot fully overlook pronunciation to teach grammar...

References


François Maniez. Tracking diachronic variation in the oral components of English corpora

The oral components of major corpora like the British National Corpus and the Corpus of Contemporary American English (Davies, 2010) are increasingly used for teaching English, by means of comparisons between oral and written production (Römer, 2010). With the increasing availability of information technology tools in the classroom, the use of concordances to provide authentic examples of language use by natives has been gaining currency (Tribble & Jones, 1997) and is becoming an established teaching method. However, the question of the ideal size for the oral component in any corpus that claims to accurately reflect language use is rarely raised.

We first describe the makeup of the two above-mentioned major corpora and highlight the differences in their oral components. Using some of the indicators suggested by Biber et al. (1998), we compare the BNC and COCA in terms of frequency for such indicators and we explore the combined use of so-called vague language markers (Quaglio, 2009) and the emerging syntactic structure so not + (ADJ/PRONOUN/V+ING) as markers of oral speech.

We also discuss the use of larger corpora such as the Google Books Corpus and web corpora (Schäfer, 2015) as an alternative to the BNC and COCA for studying recent linguistic change in oral speech, and the extent to which the presence of language samples that actually replicate conversation can be said to reflect oral style, even though such corpora are fundamentally representative of the written form.
It’s 11 years now that Johansson’s (2009: 41) claimed that more systematic studies are needed in order to test the benefits of DDL and to discuss ‘students’ problems with corpus investigation’. The meta-analyses carried out by Boulton and Cobb (2017) and Lee, Warschauer and Lee (2018) have cast robust results on the benefits of DDL. Boulton & Cobb found average effect sizes of 1.50 for pre/posttest designs and 0.95 for control/experimental designs. Lee, Warschauer and Lee (2018) have reported a medium-sized effect of corpus use on vocabulary learning. Overall, these are medium or large effect sizes that support the use of DDL for language learning. The second claim presents researchers with different angles to exploit, and can possibly be studied using different approaches and foci. Chambers (2005: 111) called for further research that looks at ‘the integration of corpora and concordancing in the language-learning environment’, and Pérez-Paredes et al. (2012: 499-500) noted that few studies have examined the interaction of learners with corpora and, as an extension, how learning happens. This study problematized the role of learners as researchers as it was found that only a small percentage of the searches either were effective or showed the minimum levels of sophistication needed to carry out the task at hand.

DDL research has paid considerable attention to measuring the outcomes of learning, but perhaps not so much to what learning and how learning takes place. Lee, Warschauer and Lee (2018) noted that using ‘pre-selected, comprehensible concordance lines appears more effective in supporting their corpus-based activities’ and so does learner-friendly concordancer software specifically designed for L2 learning’, something that had been pointed out in Pérez-Paredes, (2010). An emphasis on technology and tools is manifest. In this talk I will examine one of those angles that need further attention: learning. I will draw on the systematic review of the uses and spread of data-driven learning (DDL) and corpora in language learning and teaching across five major CALL-related journals during the 2011–2015 period in Pérez-Paredes (2019). This review examined 32 research papers published in high rank CALL journals and concluded that the normalization (Bax, 2003) of corpus use in language education has only taken place in a limited number of contexts, mostly in Higher Education, where

References


Pascual Pérez-Paredes. Rethinking learning in Data-driven learning
language teachers and DDL researchers subsume overlapping roles. This finding echoes Boulton’s (2017) call to widen the scope of our research to a more comprehensive spectrum of learning contexts. I will outline a taxonomy of language learning focus, learners’ abilities and processes advocated in the body of research examined that can enhance our understanding of the role played by learning in future data-driven language learning theory or theories.

This talk seeks to contribute to previous work that has tried to bridge the gap between research and practice (Chambers, 2019). I will argue that DDL researchers need to move away from a technology-oriented DDL (Godwin-Jones, 2017) and pursue efforts that widen our understanding of the contributions of DDL to SLA, the contributions of SLA to DDL, and the in-depth analysis of the role of DDL in the broader language learning context, including the use of cognitive strategies (Lee, Warschauer & Lee, 2020).

References


Research on complexity, accuracy, and fluency (CAF) has not generally employed longitudinal designs, which are based on spoken production, except for some studies (e.g., Larsen-Freeman, 2006; Vercellotti, 2017). Accordingly, it is still not clear how CAF in language performance changes during the development of language performance. In series of studies using the Longitudinal Corpus of L2 Spoken English (LOCSE), we claimed to have found evidence (e.g., the number of words between disfluency markers) for tracking learners’ speech fluency improved across time (Abe et. al., 2019). Thus, using the same learner corpus, the present study investigated novice learners’ development of complexity. The LOCSE consists of the data collected for three consecutive years from 2016, with the total of eight data collection points (Abe & Kondo, 2019). The spoken performance was elicited from the same 122 learners. The Telephone Standard Speaking Test (ALC Press, 2016) was employed and learners’ performances were evaluated holistically by multiple certified raters.

Our analyses examined indices of syntactic complexity and the use of subordinates. In order to measure syntactic complexity, learners’ utterances were parsed on the basis of Dependency Grammar. After that, we calculated the index of the graph centralization (Freeman, 1979) of the parsed dependency trees of the utterances. This index was developed to measure structural centrality in networks, but it is also applicable to measure syntactic complexity of a parse tree. As a result of the study, linguistic complexity of novice learners’ utterances developed in three years, and the correlation between the measure of complexity and that of fluency changed across time. Regarding the use of subordinates, learners heavily relied on a very few simple subordinates to fulfil linguistic functions such as reasoning (e.g., because, so) and contrasting (e.g., but). The use of other subordinates was extremely limited, and novice speakers seemed to have difficulty in connecting more than one information unit in a logical manner. In other words, lexical variety of subordinates was negatively correlated with syntactic complexity. Considering these results, learners should be encouraged to use various connectors between multiple sentences, rather than focusing on accuracy in a single sentence.

References


This paper reports on students’ use of an ESP corpus of product descriptions to explore the linguistic features of texts they are required to write for a project-based ESP module, the final output of which is a fictional company website. The participants are 84 Italian university students in their second year of a degree course in ‘Language, literature and cultural communication’. Employment opportunities for students on this course include working in commerce. Indeed, industry in the area (Veneto) is characterised by a high number of micro, small and medium-sized enterprises. Those companies engaged in international trade often require language graduates to write specialised texts in English for their websites and other forms of commercial communication. During the module, the students work in groups to compile corpora of ‘About us’ pages, presentations of staff members and product descriptions. They are then trained to analyse these corpora to identify the linguistic features of these genres for their own writing.

Given the lack of research into the teaching of the language used in spatially constrained (Bruthiaux 2000), syntactically reduced, persuasive texts, this paper focuses on product descriptions. All but one of the groups chose to set up a fictional clothing company, so the product descriptions that the students intended to produce were mainly for items of clothing. On a grammatical level, key features include the omission of redundant items (such as articles and auxiliary verbs), and the use of long noun phrases, imperative verb forms and –ed clauses. As for lexis, the short texts may include the highly specialised vocabulary of product-related attributes, but also non-product-related attributes that reflect the “imagery, experiences, and feelings that consumers have in relation to a product” (Camiciottoli 2018:142). These persuasive elements may indicate glamour, luxury, quality, convenience, affordability, and environmental awareness. The hands-on data-driven learning (DDL) approach (Boulton 2012) adopted aimed at enabling the students to identify these features. The students had previous experience (Ackerley 2017) of hands-on DDL using AntConc (Anthony 2014), but in this module they were introduced to POS tagging.

This paper illustrates how the corpora were exploited for hands-on data-driven learning and then investigates the following research questions:

What kinds of searches did students carry out on their corpora?

How did the use of corpora impact their writing?

To answer question 1, students were asked to note down specific search terms, POS tags and strings of POS tags searched for; the lexical items or phrases retrieved; and the actual phrases used in their product descriptions.

To answer question 2, a pre-/post-test approach was used. At the start of the module, the students were asked to write a product description based on a photograph of an item of clothing. A learner
The corpus of these texts was then compared with a corpus of texts that the students produced for their website, with a focus on the structure of noun phrases.

The findings of the study should contribute to awareness of how the teaching of ESP and, in particular, of English for occupational purposes can be enhanced through data-driven learning.

References


Within the field of Applied linguistics and second language pedagogies, the emphasis has shifted from deductive to inductive learning, where the concept of 'noticing' (Schmidt, 1990) language patterns is promoted. This type of student-centered self-discovery of lexico-grammatical patterns can be greatly aided by corpus linguistics methods, specifically 'Data-Driven Learning (DDL)' (Johns 1986; O'Keeffe et al, 2007). This involves learners investigating real language use through computers. Many corpus-based studies have been carried out in relation to language learning (Boulton & Cobb, 2017), but language corpora have not been integrated into mainstream teaching practices. Teachers who would like to incorporate language corpora into their instruction are often overwhelmed by the task of locating appropriate corpora or designing activities for their students (Reppen, 2010; Zareva, 2017).

This paper will report on a longitudinal experimental study whose aim was to provide useful skills and processes in the use of language corpora as a writing resource in the EFL classroom of 24 pre-intermediate level learners. The evaluation data of the study included pre-, post-, delayed post-test data, and Learner Autonomy Profile form, whose statistical analysis revealed the beneficial impact of computer-based inductive approach of DDL on the learners’ grammar competency and writing skills, independent learning skills, as well as the contribution of cognitive strategies to long-term implicit knowledge. It also included semi-structured interview data, which uncovered the learners’ increased engagement in the learning process, the positive change in their attitudes towards their own learning, and the ways of demonstrating autonomous abilities in working with concordances. These data also brought to light some of the fears and challenges of using DDL. Thus, the findings as to how to enhance the language learning process, maximize the probability for long-term retention of knowledge, foster
learner autonomy, increase motivation and cultural awareness will serve all the participants of this hugely important ELT sector - researchers, language educators and learners.

References


Francisco J. Álvarez Gil and Elena Domínguez Morales. Teaching point of view in English for tourism. A corpus linguistics approach [4]

The use of text compilations for the study of language is gaining ground very rapidly, and this is well represented in the number of texts and manuals for teaching and learning languages, either as L1 or L2. Clear examples are the COBUILD grammar and dictionary (Sinclair, 2003) and the Longman Grammar of Spoken and Written English (Biber et al., 1999). Computer tools are not only very much needed to accomplish the study of these compilations, but they have also become essential in order to show students proof of the existence of patterns and uses in languages as a way of illustration in English for tourism studies. Given a classroom setting, our goal in this presentation is teaching students the use of corpora and how to interrogate them to get examples of perspectivization in real English for tourism according to certain designed criteria. Our intention is to address how large textual compilations of languages can be used in the classroom to develop language skills: writing, speaking, listening, reading and interaction with special attention to the communication of the speaker’s perspective. In short, our objective is to give ideas concerning the corpus linguistics methodology in second language teaching in English for tourism studies. Our hypothesis is that the use of this approach allows students to improve their command of the language by means of certain deductive procedures arising from their use of real English for tourism as shown in textual compilations of texts in the domain of tourism. In this paper, we focus on a compilation of written texts amounting to ca. 1,000,000 words of specialized English for tourism and conclusions will deal with aspects concerning the usability of corpus linguistics to English for specific purposes.

References


Even though advanced L2 learners develop grammatical skills and vocabulary in the target language, their written productions sometimes fail to satisfy the expectations of native readers. Because the organization of the ideas expressed throughout the text differs from one language to another, and due to the linguistic influences of the writers’ L1, learners face challenges that are usually not addressed in the classroom.

In order to better understand these challenges, we have built a corpus of non-academic argumentative essays written by university students on a single subject. The corpus consists of:

Two control groups:
- 10 monolingual Turkish students
- 10 monolingual French students

One learner group:
- 10 Turkish learners of L2 French at B2 level

Through the analysis of the two control groups, we developed hypotheses related to the challenges learners may face in argumentative writing. We checked these hypotheses against the group of learners. We analyzed and compared all texts on a rhetorical as well as linguistic level, and more specifically looked into references to three types of modalities which seem to be most widely used in argumentative texts: the modality markers related to assertiveness, the epistemic modalities, and the deontic modality markers.

Results indicate that on a rhetorical level, the use of paragraphs (usually absent in monolingual Turkish corpus) is a challenge that can be overcome through teaching. However, the ability to generate ideas and develop arguments is a serious hurdle in L2 and results in significantly short texts.

On a linguistic level, it appears that learners are influenced by their L1 grammatical strategies in expressing assertiveness and formulating hypotheses related to the future in L2. Turkish writers also tend to overuse deontic modalities in argumentation. This corpus-based research could lead to new teaching strategies. The corpus itself (especially the monolingual French component) could be used in the classroom to provide an authentic insight into first/second language writing.

References

The growing research interest in Content and Language Integrated Learning (CLIL) has resulted in empirical findings assessing the purported outcomes on learners’ language proficiency and their engagement in the everyday learning. Student motivation in CLIL classrooms has been shown to improve as compared to traditional English as a Foreign Language (EFL) instruction; however, this effect is likely mediated by many factors (Doiz, Lasagabaster & Sierra, 2014). Notably, the role of pedagogy and the design of instructional materials in CLIL on student motivation have not received adequate attention. The present study will examine the effect that readability of textbooks, specifically vocabulary difficulty (referred to here as lexical sophistication), might have on student motivation in secondary school CLIL programmes in Andalucia, Spain. The theoretical framework and methodology draw on research in motivation (Dörnyei and Clément, 2001; Gardner, 2007; Dörnyei, 2009), lexical threshold theories of readability (Nation, 2006; Schmitt, 2008), and Krashen’s theory of Comprehensible Input (1989). We hypothesized that if texts were difficult, learner motivation might be adversely affected.

Ninety-three students in first year and thirty students in their third year of compulsory secondary education, enrolled in Spanish-English CLIL programmes, completed the vocabulary test and a questionnaire on motivation. We operationalized the construct of readability by determining the gap between students’ lexical proficiency and the lexical sophistication in course texts, as follows. First, a test of receptive lexical proficiency provided a score which we have aligned to a CEFR-based scale (EnglishProfile, 2019). Then, we created word lists - based on our own modelling of these test results - representing words students were likely to know or not know (modelled after Cobb, 2007); all Spanish-English cognates were presumed as ‘known’. Finally, we determined percent coverage of words ‘likely to be known’ in course texts (AntWordProfiler, n.d.). Finally, these results were compared to students’ motivation towards CLIL, as determined by a Likert-scale survey.

The lexical coverage analyses revealed that vocabulary difficulty of instructional materials for both grades is beyond the students’ expected language competence, since the number of known words in class texts lies well below the 95% coverage recommended for comprehensible reading. Considerable variation was found within both groups in terms of measures of their receptive vocabulary knowledge. Regression analyses revealed cognate awareness to be the major predictor of the vocabulary test scores. Finally, students generally demonstrated a positive attitude towards CLIL classes, and no significant correlations were observed between the motivation factors and participants’ receptive lexical knowledge.

The presentation will discuss some major methodological challenges for CLIL programs, and will suggest a thorough examination of the difficulty of course materials, with respect to their lexical sophistication. These findings provide practical implications for publishers to develop materials that address students’ linguistic and cognitive needs, which would simultaneously maintain increased levels of motivation.
Raffaella Bottini. Lexical complexity and L2 English speech: the effect of test takers’ individual properties in the Trinity Lancaster Corpus [3]

This paper presents an investigation into the effect that learners’ individual properties, such as age, L1, and proficiency level, have on lexical complexity in the Trinity Lancaster Corpus (TLC). The TLC is a 4.3-million-word learner corpus of spoken L2 English. It is based on the Graded Examination in Spoken English (GESE), developed and administered by Trinity College London, a large international examination board. Studies on learner language have shown that vocabulary knowledge is one of the best predictors of language use and overall proficiency (e.g. Milton, 2013; Nation, 2013). Different measures of vocabulary knowledge have been proposed in the field and lexical complexity plays a key role among them (e.g. Kim, Crossley & Kyle, 2018; Lu, 2012; Tidball & Treffers-Dallers, 2007). However, little is known about different aspects of lexical complexity in spoken L2 production and how these relate to learners’ individual characteristics; also, there is no general agreement about which of the many existing complexity measures to use. This study uses indices of lexical complexity specifically selected to analyse spoken language. It examines three proficiency levels (B1, B2, and C of the CEFR), looking into lexical complexity and its relationship with test takers’ age and L1. Preliminary results show significant correlations between lexical complexity and proficiency. The analysis of the effect of test takers’ age on lexical complexity highlights a more complex scenario, with significant correlations, but a small effect size. Implications for SLA and language assessment, as well as directions for further research, will be discussed.
References


Alex Boulton. KWIC and easy? Students building their own corpus for non-linguistic goals [1]

Corpora have long been used for pedagogical purposes, with learners, teachers and other L2 users exploring language data in a process commonly called data-driven learning or DDL (Johns 1990). However, a recent systematic trawl reveals that of 384 empirical DDL publications 1989-2019, only about 4% involve students participating in the corpus building process. This paper looks beyond DDL to examine how Master’s students can reconcile a corpus approach with their own needs and interests in literature, cultural studies, linguistics and pedagogy. This distance course required them to build their own corpus for analysis with AntConc to answer their own questions, submitting the results in the form of a short research paper. Over 5 years (2013-18), 122 texts usable for present purposes were submitted and are presented here with an eye on the corpora and tools as well as the range of topics covered. The main part of the talk then concentrates on the final “personal feedback” (PF) section of their papers. These sections were themselves compiled into a corpus of 30k tokens so that linear reading can be complemented by a corpus analysis of frequent words, expressions and concepts. The main finding is that this first contact with corpus linguistics is initially daunting for the vast majority of students; however, as they proceed (often heuristically), they largely come to appreciate the power of the approach, and the autonomy and individualisation it affords. This is illustrated by comparing the first sentences of their PF with the rest of the text, and by analysing words that typically follow intensifiers, where ‘positive’ words such as *interesting, useful or easy*, for example, outnumber negative items such as *difficult, hard or lost* nearly two to one. The conclusion is that corpus tools and techniques can be hijacked for multiple purposes; while not easy for everyone, perseverance (where circumstances allow and needs justify it) can bring major rewards to a sizeable minority at least, even in a distance learning programme.
Noah Bradley. Learner parallel concordancing in a new context: Perceptions among university French immersion students [1]

Data-Driven Learning (DDL) is an increasingly important facet of second language learning with flourishing applications, one of which is learner parallel concordancing (LPC). LPC involves the use of bilingual corpora by language learners in order to elicit comparisons between the learner’s L1 and L2. This application has been explored since before the turn of the century; nevertheless, only a handful of studies have been conducted on the subject, and LPC has not been extensively adopted by second language teachers. Furthermore, despite the abundance of studies carried out on the French immersion program in Canada, none have involved LPC—including at the university level in the post-French immersion context. Using a sequential mixed methods study design, the researcher implemented a quantitative questionnaire and qualitative interviews to answer the research question: namely, to capture the perceptions that post-French immersion students (n = 16) hold towards using an online parallel concordancer called Linguee for improving their writing in French. The study also provides educators with a model for how to implement corpus pedagogy in the second language classroom. The results demonstrate an overall positive perception among students regarding the concordancer, although some students took an overall negative view, and several students experienced certain difficulties with the concordancer. Nevertheless, these difficulties ultimately lead the students who were interviewed in the study (n = 3) to express various types of metalinguistic reflection and language awareness. Through these reflective processes, the interviewed students demonstrated their ability to verify concordancer searches with other language learning websites, as well as assemble various online resources into their own customized learning system. The interviewed students thus concluded that concordancing is not the ultimate, exclusive resource for language learning; rather, concordancing is one of many online tools that can be mutually cross-referenced to create a clearer understanding of how the French language functions. Further research should thus focus on the relationship between LPC and other types of language learning aids, as well as how students formulate their own personalized language learning toolkits.

Nigel Bruce and Peter Storey. Big corpora development for disciplinary DDL in an EFL-medium university setting [1]

With increasing cuts to the funding of direct EAP teaching at tertiary level, Data-driven Learning (DDL) as a vehicle for tertiary writing support has become an economic necessity. The main problem for concordancer developers has always been getting past copyright to obtain sufficient quantity and quality of textual data, particularly in the academic context. A related problem has been generating a streamlined concordancer that exclusively targets the learner-writer, with the sacrifices in complexity – and research funding prospects - that entails. In this paper, we report on two online writing support resources: the prototype test site LegalEasy, a 26m. word single-genre database of judicial case reports, now freely available to the public online, and PsychEasy, a 2 bn. word database, tagged for IMRAD* sub-genre, of published articles in 39 Psychology journals covering the past decade, obtained through a trial agreement with a major publisher. We follow in the footsteps of work by Peter Crosthwaite (2019a, 2019b) and Ana Frankenberg-Garcia (2019) and their associates which has sought to integrate concordancing and lexicography within the student writing process, to produce resources
that are transparent and intuitive for learner-writers. Our target users are Chinese NS university students, studying in a language (English) that has a different script and major grammatical differences to Chinese. Hence, while we are pursuing the visualisation options deployed by Crosthwaite and Frankenberger-Garcia, our search results continue to feature linear concordanced displays of lexicogrammatical frequency patterns in our corpora. We believe in imitation as the sincerest form of academic flattery, in a context where the students’ focus is firmly set on articulating their substantive ideas and where they are subjected to as little metalanguage as possible.

This presentation begins with a brief display of unique features of our prototype writing support website, LegalEasy [at legaleasy.hksyu.edu **]. A key feature is that concordanced results are displayed in descending frequency, with one example per pattern found. Users are offered an initial focal choice between lexical POS combinations: a permutation of verb, noun and adjective [rarely adverbs]. Lexical searches tend to be over 90% nominal [e.g. legal concepts], so users will be offered three display options: Adj + N, N + N, or Vb + N; reasons for these decisions will be given. We then describe how we handle students’ writing and how LegalEasy fits into that process; the long-term aim is for students to become autonomous users of LegalEasy during the rest of their legal studies. Finally, we offer results of our analysis of student usage and of student evaluation of the online writing support offered by the LegalEasy program. Instruments include automatic capture of search data to examine search focus and patterns, and surveys that cover a profiling of users’ experiences and preferences regarding academic writing support, including comparison with alternative language support resources at our university.

* IMRAD: the classic rhetorical framework for research output: Introduction – Methods – Results and Discussion.

** Access via username: guest@writeeasy.org; password: passthrough; the Tutorial shows how the website works.

References


María Luisa Carrió-Pastor. The use of corpora for teaching multimodal metadiscourse in academic English [2]

Nowadays, due to technology and the free access to information, English learners have different ways of accessing to academic papers, which include many visual elements that guide readers and illustrate ideas and research results. The elements that guide readers in textual academic discourse have been
analysed by several researchers (Hyland and Tse, 2004; Hyland, 2005; Mur-Dueñas, 2011; Carrió-Pastor, 2014, 2016; Jiang and Hyland, 2016), but visual metadiscourse has received less attention (Kumpf, 2000). Additionally, the analysis of multimodal discourse in English for specific purposes has received till now little attention. To fill this gap, this paper focuses on the multimodal analysis of academic papers from the specific fields of linguistics, engineering and medicine. The paper pays attention to the identification of interactive metadiscourse devices (Mur-Dueñas 2011) and visual metadiscourse elements (Kumpf, 2000). The hypothesis of this paper is that writers with different academic backgrounds but the same linguistic background express in academic English differently in academic papers. Thus, the main objectives are, first, to study the interactive metadiscourse devices in the three specific fields of academic discourse; second, to identify the visual metadiscourse elements used in the three specific fields of knowledge, and finally, to classify the multimodal elements used to cohere the different parts of the paper with the aim of guiding the reader textually as well as visually. To this end, sixty academic papers were compiled from journals included in Journal Citation Reports (2016 edition). Twenty belonged to the specific field of linguistics, twenty to engineering and the last twenty to medicine. The corpus was analysed with the help of specific tools and also manually to check the data extracted. The results showed that there are in fact differences in the way academic writers use textual and visual metadiscourse, the study revealed that engineering writers prefer the use of visual metadiscourse meanwhile linguistic researchers prefer textual metadiscourse. Also, textual and visual patterns associated to the different specific fields of knowledge were identified. Finally, conclusions were drawn that answered the objectives of the study. This paper shows some of the results of the research project FFI2016-77941-P (Ministerio de Economía y Competitividad, Spain).

References


Due to the rapid development of globalization and the knowledge economy over the last several decades, universities worldwide have been striving for more research outputs with greater impacts in international journals in order to be more visible in the international academic community. Such pressure to publish internationally, mostly in English, falls on not only university academic staff but also research postgraduate students. As a result of such pressure, some EAP practitioners have experimented with the use of data-driven learning (Johns, 1991) to help graduate students in their research writing (e.g. Bianchi & Pazzaglia, 2007; Charles, 2007, 2014; Cortes, 2007, 2014; Diani, 2012; Flowerdew, 2009, 2015; Lee & Swales, 2006). The use of corpora in research writing can indeed be highly beneficial for research postgraduate students, given that they can access or easily create discipline-specific corpora that contain real-life published writing samples from their own research fields (Charles, 2015, 2018; Chen & Flowerdew, 2018). Except for notable studies by Charles (2014) and Yoon (2008), however, little research has yet been done on whether, why, or how research students may use corpora in their actual research writing in the longer term after being introduced to the approach.

Accordingly, this study investigates corpus use by PhD students for research writing during one year after attending a corpus-based research writing workshop. In the workshop, the students were taught to use the BNCWeb online corpus and to build and search in their own personal corpora using AntFileConverter (Anthony, 2016a) and AntConc (Anthony, 2016b) respectively. The study is mainly based on qualitative data from semi-structured interviews with 13 PhD students from five different universities who attended the workshop and analyses their corpus use histories. The interviews were either small-group or individual interviews, depending on the interviewees’ preferences, and were conducted in the language chosen by them (English, Cantonese, or Putonghua).

Although all students fully acknowledged the benefits of corpora for their writing, many had difficulties forming a long-term corpus habit (Charles, 2014), one major reason being that they did not write much or at all during long periods of their study, due to time being taken up with experiments or data
collection for their research. Another finding was that, while many of the participants made use of wildcards and other complex search techniques rather than simple single-word searches, some found it difficult to remember all the search techniques taught in the workshop. Furthermore, the students’ preferences for corpus tools were greatly affected by the complexity of the interface or the amount of time required to carry out searches. For this reason, although students were taught to make their own corpus, this approach was less favoured, even though many participants acknowledged that having a personal corpus is highly beneficial. Of course, this finding, as with others, may have been due to the inadequacy of the workshop and not the corpus tools.

As regards the specific purposes the participants made use of corpus tools for in their research writing after the workshop, the interviews revealed that they used corpora for five major purposes: 1) to ascertain whether an expression or a sentence they produced was native like; 2) to explore alternative expressions by checking the co-texts of certain key words or phrases; 3) to discover patterns that were unknown to them, very often collocates or colligates of certain words/phrases; 4) to find useful example sentences or sentence templates for different rhetorical purposes; 5) to identify new technical terms specific to their research fields by using their personal corpora with AntConc.

Findings from this study suggest that corpora could become an important resource for PhD students to improve their research writing. Although the benefits of corpora are evident to students, they do need prolonged training to develop long-term habits. Given that many PhD students, especially those in Science, do not write in English regularly, in order to ensure a long-lasting effect of the teaching, short-term tutorials or workshops that are organised repeatedly throughout the year would be more beneficial than longer one-off single workshops.

References


Researchers have successfully described the linguistic characteristics of academic language use from a variety of perspectives (Biber et al., 2002; Biber, 2006; Csomay, 2005), including the use of academic vocabulary (Coxhead, 2016). While the majority of the studies focus on academic vocabulary examining written texts that students are exposed to or discuss how to teach such vocabulary, relatively few studies explore how students actually use these kinds of words in their scholarly papers. Most recently, scholars describe the lexical characteristics of non-native student writing to explore developmental patterns, for example, in the use lexical bundles (Yan & Staples, 2017) or in the use of academic words (Csomay & Prades, 2018).

The present study investigates a specific group of second/foreign language (L2) learners’ use of academic vocabulary in their writing assignments. These students attending an American STEM degree program offered in Tbilisi, Georgia (Eurasia) produced a variety of naturally occurring classroom-assignment papers in English in their General Education courses during a period of one academic year. Over a thousand student papers were compiled into a corpus originating from lower- and upper-level composition classes, History, Philosophy, and Political Science. Using Gardner and Davis’ (2014) Academic Vocabulary List (AVL) generated from the Corpus of Contemporary American English (COCA), the percentage of academic vocabulary was counted in the final draft of each student paper. The students were neither exposed to this academic vocabulary list during their studies nor were they explicitly taught how to use academic words in their writing assignments.

Results indicate a) that students use incrementally more academic vocabulary as they progress in their coursework; b) that academic vocabulary is used differently across the text-types students produced; and c) that students use words differently in the two levels of instruction. Pedagogical as well as programmatic implications are discussed.
Ever since Römer (2009, p.84) outlined the need for corpus enthusiasts to “spread the word” about the applications of corpora for education as a kind of “missionary”, there has been a sharp increase in the frequency of studies on the direct use of corpora in the language classroom in the form of Data-driven learning (DDL). However, this increase has been largely restricted to technologically-rich Western and East Asian countries (Boulton & Cobb, 2017), and this situation is reflective of a “digital divide” (Lozano & Izquierdo, 2019) between relatively ICT-poor countries and their richer neighbours. As a result, very little is currently known about teacher trainees’ or educators’ acceptance of or preparation for DDL in this region. However, if DDL is to gain a foothold in mainstream education, more work has to be done to win over the “hearts and minds” (Crosthwaite, 2019) of teacher trainees in adopting this approach in their own language classrooms.

This paper reports on initial attempts to implement DDL pedagogical practice within the curriculum of a large pre- and in-service teaching training institute in Indonesia, as part of a larger ongoing project to determine potential enablers and barriers to DDL within teacher education in the S.E. Asian region including partners in Vietnam and Singapore. Over 50 participants took part in a 6-hour online DDL training course (Improving Writing Through Corpora, Crosthwaite, 2020) to learn the basics of corpus query techniques and corpus-assisted second language error correction in SKELL/SketchEngine. Face-to-face workshops were then held to provide instruction in self-built corpora for specific teaching contexts, and tips on modifying existing detailed lesson plans to include activities involving corpus consultation. Participants then developed a sample lesson plan involving DDL according to the educational and cultural needs of their target students, before completing a post-training questionnaire modified from similar research involving DDL for teacher trainees in Hong Kong (Chen, Flowerdew & Anthony, 2019) with accompanying interviews to ascertain trainees’ and teacher educators’ individual and collective reactions to DDL.

In our talk, we outline the findings of interview and questionnaire surveys, while presenting some of the sample DDL lesson plans teacher trainees had produced. Participants’ responses were generally favourable to DDL, although comments related to workload in planning lessons involving corpora, a lack of ICT infrastructure / high speed broadband in certain regions, overreliance on translation tools for language learning, and difficulties conceiving how and where to implement corpora into current materials and curricula were featured as potential barriers to the implementation of DDL in Indonesian schools. We close the talk by outlining potential solutions to the problems raised by the teacher trainees, and discussing whether these Indonesian findings may be replicated in other S.E. Asian contexts.

References


Niall Curry. Using contrastive linguistics and learner corpus research to inform the teaching of academic metadiscourse: a corpus-based study of target, first, and learner language corpora [2]

It is well established that non-Anglophone users of academic language have been disadvantaged when writing in English, owing to insufficient command of academic style (Bocanegra-Valle, 2014). The linguistic disadvantage has resulted in the development of a large teaching industry, which aims to equip learners of academic English with rhetorical, interactive, and metadiscoursal competencies in academic writing (Hyland & Hamp-Lyons, 2002). Historically, one such approach to informing the teaching of academic English has been to conduct contrastive studies. However, their weakness as the sole means of predicting errors and challenges is well documented, and Gilquin (2008) sees a value in the study, in synchrony, of target, first, and learner language in order to better identify language areas of pedagogical interest. This paper adopts such an approach with an explicit focus on academic metadiscourse. Using linguistics and medicine corpora in English from METOOL (Carrió-Pastor, 2019) and French from KIAP (Fløttum et al., 2006), this study identifies the presence and use of metadiscourse in academic writing in English and French. This is done in order to identify differences in metadiscourse use by expert users across discipline and language. Next, using academic learner corpora, such as the Cambridge Learner Corpus, first language French speakers’ academic English is studied for the presence and use of metadiscourse. The findings of the study of the target, first, and learner language are compared in order to identify and better understand metadiscourse use by French speakers and indicate areas of metadiscourse that require explicit teaching in academic language teaching contexts. The results of this study not only add to contrastive and learner corpus research on academic metadiscourse, but can also be used to better understand potential language transfer, better inform language teaching materials, and raise language teachers’ awareness of issues surrounding metadiscourse use across disciplines and first language.

References


In the field of English for Specific Purposes (ESP), as in any other type of interlanguage, phraseology contributes significantly to successful academic writing (Biber & Barbieri, 2007). For particular learner varieties, such as Romanian English, few studies have examined formulaic sequences (Hyland, 2008), mainly focusing on lexico-grammatical patterns (Chitez, 2012, 2014). The proposed paper investigates the use of phraseology in Romanian students’ academic papers, written during their ESP courses, by adopting a double contrastive perspective: first, we contrast texts produced in two different disciplines (Literature Studies and Information Technology), and, second, we compare the academic phraseology in learner language with native speaker phraseology. The aim is to find out whether the use of academic formulaic expressions differs according to the discipline and the extent to which students integrate expert academic phrases into their writing.

For the analysis, we have compiled two corpora (ESP-LIT and ESP-IT), each consisting of 40 texts representing a discipline specific didactic genre, e.g. essay. As reference, we used the Academic Phrasebank (Davis & Morley, 2018), “a valuable online resource containing lists of key phrases serving specific functions in research papers, providing support in language and organisational norms” (Stojković et al, 2017, p. 146), which has been developed using the corpus methodology as well. We counted the frequency of academic phrases in the two corpora and compared the ESP learners’ phrase lists with the list provided by the Academic Phrasebank. Consequently, in order to validate our results, we correlated the percentages (expert and non-expert phrases) with the AW (Academic Word List) results extracted, for each text, with the help of the online text analysis tool Text Inspector (Bax, 2015). The findings showed that students in the Literature Studies predominantly use other types of academic phrases than students of Information Technology. At the same time, a slight correlation could be observed between the degree of expert phrase use and AW ranking, plus additional parameters, in Text Inspector, thus contributing to discussions on the relationship between n-gram use in learner language and writing proficiency (e.g. Garner et al, 2019). We propose the inclusion of our results into a pedagogical model that requires the combination of group-adapted digital resources for academic writing support.

The study is part of the research project ROGER (Academic genres at the crossroads of tradition and internationalization), in progress at the West University of Timisoara, Romania, which aims, among others, at creating an online linguistic-resource platform hosting a bilingual Romanian-English comparable corpus of student genres, useful to teachers and researchers for the development of digital analyses and learning tools.

References


Nihan Erdemir and Hacer Hande Uysal. Corpus-Based Rhetorical Move Analysis of the Research Article Results [2]

The research articles (RAs) written in English are a major means of distributing academic knowledge internationally; therefore, it is of great significance for researchers to understand the structure and conventions of this genre described by the members of a discourse community. Despite a number of rhetorical move analysis studies in various disciplines, the research in Educational Sciences is very limited, and also, there is still comparatively little research on the results sections compared to other sections. Thus, the present study aims to explore the discourse conventions of the results sections in L1 English RAs by means of identifying the communicative move and step segments and examine text structure in the sense of frequency, sequencing, patterning, and cycling. The study adopted a corpus-based approach to discourse analysis because the systematic organizational patterns of variation at the discourse level in the corpus was analyzed. First, in the pilot study move analysis based on Swales’ (1990) tradition in genre analysis was conducted in the comparable corpora compiled with 12 published RAs by implementing the framework of Cotos et al. (2015). In accordance with the findings of the pilot study, a new rhetorical model was constructed for the field of Educational Sciences by adopting the frameworks of Cotos et al. (2015) and Morena and Swales (2018), and this new model was consulted in the main study. Move analysis was conducted with 40 published RAs to identify rhetorical moves and steps. The inter-rater coding reliability was measured by Cohen’s kappa (k) statistic. Different from the previous models, the findings of the current study might first contribute to the understanding of the distinctive features characterizing the results and discussion sections in Educational Sciences. First, the results sections embody the move describing the methodological procedures (M2) as essential in addition to reporting and commenting on the results move (M3) while the discussion sections lack methodological procedures. Second, establishing the study move (M4), which is more identified with the nature of the discussion sections, was less frequently noticed in the results sections. Moreover, in the sense of frequency use and sequencing of moves, the findings suggest that L1 English RAs utilized rhetorically complicated and lengthy text structures, and at the same time, L1 English RAs had a better command of using a wide range of steps at high levels. In the sense of cyclicity, a high degree of move pattern cycling was observed in the corpora. Lastly, the move
analysis results suggest a new rhetorical model, which is identified with four moves and seventeen steps in the field of Educational Sciences. The findings are expected to provide insights to novice writers about the accepted conventions of English language regarding the organization of research article results, and also to provide suggestions to practitioners, material designers, and course book writers to address EAP learners’ lack of knowledge of the rhetorical organization in academic writing.

Katerina Florou. Comparison of two Translator Learner Corpora: The case of the Italian opinion verbs [3]

Most Italian teachers seem to agree upon the fact that verbs of expressing opinion e.g. “credo”, “penso”, “considero” etc. have always been an entrap. Due to their strict syntax, students of Italian language tend to avoid using them. In defining what the difficulty is the Academia della Crusca’s grammatical rule gives some answers: “The verbs that express […] an opinion support the subjunctive” and for this reason learners prefer using other verbs or not to express opinion at all rather than use subjunctive, which is considered the most demanding conjunction. Even though this is the tactic that teachers have noticed in the classroom of the Italian language, is it indeed what happens in case of advanced students of Italian language or translator trainees? And is it different from what happens in other language classrooms, in English for instance, with verbs such as “think” or “consider”?

To explore this hypothesis in Italian and (subsequently) in English language we examined the data from the Multilingual Translator Learner Corpus (MTLC) (Florou, 2019). The translations used were collected from students of the Department of Italian Language and Literature and the Department of English Language and Literature at the National and Kapodistrian University of Athens in Greece. The corpus compiled used Greek as a source language and contains approximately 75 essays in Greek and their translations in Italian, and an equal number of essays and translations for the English sub corpus as well.

Analyzing the MTLC, beginning from the Greek texts, we found a notable amount of verbs that express opinion in both sub corpora (in the Greek sub corpus which has been translated in Italian, but also in the Greek sub corpus which has been translated in English), probably as a result of the narrative genre of the texts and the style of the topics. The findings reveal a vast use of “opinion verbs” in all three languages, but this use did not always indicate personal opinion. In many cases we saw the significance of thought or of belief, in other cases they were used verbs that do not support subjunctive (especially at the Italian sub corpus, e.g. “second me”) and, finally, we noticed a considerable number of zero translation/omissions (See tables 1 & 2 below):
Unlike the original intuition, the results coincide in both Translator Learner Corpora as the number of zero translations of the opinion verbs are up to a notable presentence. As it is shown the translator trainees are particularly hesitant on using opinion verbs in translation. However, this note due to the difficulties that this choice may have, as we first assumed. It is rather possible to avoid using them because of their luck on confidence on expressing a personal opinion in the foreign language or because of certain mechanism they have developed during foreign language learning. Thus, there is need for further research to cross examine the above allegations.

1 Translation of: Reggono il congiuntivo i verbi che esprimono “una volizione (ordine, preghiera, permesso), un’aspettativa (desiderio, timore, sospetto), un’opinione o una persuasione”, tra cui: accettare, amare, aspettare, assicurarsi, attendere, augurare, chiedere, credere, curarsi, desiderare, disporre, domandare, dubitare.

References


Academia della Crusca, https://accademiadellacrusca.it/it/consulenza/uso-del-congiuntivo/104

Luciana Forti. Data-driven learning effects in the development of L2 phraseological competence: the combined role of semantic transparency and frequency [1]

Data-driven learning (DDL) effects have been investigated in relation to a wide number of variables. In particular, the kind of learning aims involved in DDL pedagogical treatments have been looked at in terms of lemma, category and phrase (Mizumoto & Chujo, 2015), language aspect (i.e. vocabulary, lexicogrammar, grammar, discourse) (Boulton & Cobb, 2017) and dimension of knowledge (Lee,
Warshauer & Lee, 2018). However, learning aims in DDL seem to have received little attention in terms of their linguistic and psycholinguistic properties.

Despite this, a growing body of research sheds light on the complexity of learning aims such as collocations, for instance, in relation to their identification, processing and use (Gablasova, Brezina, & McEnery, 2017; Henriksen, 2016). In particular, semantic transparency has been investigated as a likely predictor in the development of phraseological competence (Wang, 2016; Gyllstad & Wolter, 2016), as has frequency (Durrant, 2014). This study seeks to evaluate DDL effects from the perspective of these two collocation properties combined.

A set of 32 semantically transparent and semantically opaque verb + noun collocations was identified by calculating the inter-rater reliability coefficient (Hayes & Krippendorff, 2007) based on the evaluations of 13 raters. All the raters had passed at least one Linguistics exam at MA level. The classification criteria for the raters to refer to were drawn from those proposed in Howarth (1996, 1998). The identified collocations were then matched with corresponding frequency data based on a reference corpus (Lyding et al., 2014; Spina, 2014).

The collocation set was then used to develop both DDL and non-DDL pedagogical materials, which were integrated in eight 1-hour lessons taught over 8 weeks. It was also used to build a phraseological competence test to collect data at 4-week intervals. The participants in the study were 123 Chinese students of Italian, divided into 8 classes and enrolled in a 10-month long Italian language course at the University for Foreigners of Perugia, Italy. The study was based on a between-groups design, with 4 classes addressing the set learning aims with corpus-based materials, while the other 4 classes addressed the same learning aims by using non-corpus-based materials.

The collected data is analysed through mixed-effects modelling in order to build a predictive model of DDL effects, integrating learning patterns over time and the combined properties of semantic transparency and frequency. The results are interpreted in light of the most recent empirical evidence that is available for L2 collocation learning and DDL effectiveness.

References


Ana Frankenberg-Garcia, Simone Sarmento, Paula Tavares Pinto and Ana Bocorny. A data-driven learning approach to supporting the dissemination of research from non-English speaking institutions [2]

Many researchers across the world do not have the opportunity to work or study in an English-medium institution, where they could develop their knowledge of academic English naturally. These researchers would benefit from dedicated academic English writing support, to help them disseminate their work more widely through high-ranking international publications.

In Brazil, the tutors teaching academic English have little or no experience of high-level research writing, let alone of the specialist subjects of the researchers who seek their assistance. Therefore, they too need support.

This paper reports on how we addressed this challenge by helping researchers and English tutors at the same time, through a series academic English writing workshops that encouraged researchers and English tutors to work collaboratively with corpus tools and resources.

The workshops were held twice at the Federal University of Rio Grande do Sul (UFRGS, Porto Alegre), and another two times at São Paulo State University (UNESP, Sao Jose do Rio Preto), in April and June 2019. A total of 125 applicants participated, although demand for the workshops was more than twice the number of places we were able to offer. The participants included 72 researchers from a wide range of areas (e.g., Astronomy, Biology, Computer Science, Engineering, Politics, etc.) and at different points in their academic careers (from postgraduate research students to full professors), and 53 English tutors with different levels of teaching experience. By pairing up researchers and tutors, we aimed to encourage them to learn from each other. Researchers would benefit from having an English tutor sitting next to them to improve language awareness and ask questions, while English teachers would gain experience with research writing in fields they were unfamiliar with.

The workshops began with an overview of the specific vocabulary and textual organization of journal abstracts and articles. This was followed by an introduction to a selection of corpus tools and resources aimed at helping researchers improve their academic English writing autonomously, and equipping English tutors with materials they could access in order to teach academic English. The participants received training in: (1) consulting a corpus-based dictionary for language production; (2) structuring abstracts and papers; (3) looking up frequently used academic phrases and how to adapt them to their
own writing; (4) understanding collocation and expanding their collocation repertoire; (5) using general corpora; and (6) compiling and consulting specialized corpora in the researchers’ own fields.

Feedback collected via anonymous end-of-workshop questionnaires was very encouraging. The researchers were particularly happy to be able to use the workshop materials to enhance their own writing, and to have the just-in-time support of an English tutor sitting next to them. The tutors appreciated helping the researchers solve real problems, and being able to consult corpus tools and resources when they did not know the answer. However, there were also lessons to be learned. We hope that by sharing with TaLC delegates our experience of what went well and what could be improved in the Brazilian workshops, similar workshops can be successfully implemented in other non-English speaking countries.

Luciano Franco Silva, Paula Tavares Pinto and Luana Nazi-Laranja. A ted talk corpus for teaching lexical bundles for Brazilians EAP students [2]

The teaching of multi-word units, especially regarding EAP contexts, is an increasingly important field in teaching language and corpora studies (Biber & Barbieri, 2007). Taking into account that these sequences of words have been studied under many different terms, we chose to follow Biber et.al (1991) definition of multi-words, which the authors name as “lexical bundles”. Lexical bundles are identified using a frequency-driven approach, which means they are the most frequently occurring sequences of words. Although this definition may seem too simple at first, Biber (2006) argues that lexical bundles have two surprising characteristics. First, most of them are not idiomatic in meaning, thus, bundles like “do you want to” are transparent from the individual words. Second, lexical bundles do not usually represent a complete structural unit. Instead, they are used to bridge two structural units whether they are a clause or a phrase. Also, Salazar (2014) suggests that in order to avoid individual speaker/writer idiosyncrasies, a recurring bundle must occur in multiple texts within a register to be considered a lexical bundle. Prior studies in our research group (Silva, 2018, Pinto, 2019, Nazi-Laranja, 2019) have pointed out the importance of teaching lexical bundles to EAP students and researchers. In the first study EAP students were taught 4-word-bundles as a means to improve listening skills during academic talks; the second study showed that senior researchers in Brazil tend to overuse the same lexical bundles in their published papers and could benefit from masterclasses dealing with further bundles; finally, the third study was conducted using 4-wordbundles to improve reading skills in a high school classroom. Thus, this study is an attempt to show how a TED Talks corpus compiled from 60 talk transcripts of Engineering may be useful in improving and developing fluency of EAP Brazilian Graduate Students’ in English at the Sao Paulo State University (UNESP). On the basis of that, our study predominantly concentrates on identifying and creating a multi-word list of the most common four-word bundles in the aforementioned corpus and then proposing data-driven learning (DDL) activities based on the Language-focused learning strand derived from Nation (2007). Taking this into account, our activities are based on the study of Kennedy and Miceli (2001) who proposed a four-step approach for DDL activities as a way to train the students and promote learning by examples and experience, such steps are: 1) formulating the questions; 2) devising a search strategy; 3) observing the example and selecting relevant ones; 4) drawing conclusions. This study is organized in the following way. First, we selected the 20 most common four-word bundles in our TED Talks corpus of Engineering. Second, we separated those bundles according to their function (stance expression, discourse organizers, and referential expressions). Third, we prepared DDL activities aiming to improve
the use of English of EAP Brazilian students. We have found evidences that bundles helped students remember and understand texts better than teaching individual words. However, we understand that more research on this topic needs to be undertaken before the association between lexical bundles and English learning is more clearly understood.

References


Pinto, P. T. P. Academic lexical Bundles in Brazilian research papers. Conference Abralin 50, Maceió, Brazil, 2019.


Intensive instruction is known for its positive impact on various aspects of English learning, especially in school contexts (see Ligthbown, 2014 for an extensive review). The goal of these programs is to have the students to speak in a way that enables them to be understood by native English speakers. Thus far, however, no studies have explored in a speech corpus the nature of the students’ progression in intensive settings using a tri-dimensional model (Derwing & Munro, 2013) targeting the notions of perceived fluency, comprehensibility and accentedness from the perspective of different raters such as native speakers, English L2 teachers and native speakers (students) of the same age.

This study set out to investigate the development of grade 6 (age M = 11.4) L2 English of French students’ (n = 52) fluency (flow of language output), comprehensibility (perceived difficulty of understanding) and accentedness (perceived degree of accent) when enrolled in an intensive English program (10-month) in Quebec, Canada. The speech corpus is based on the spontaneous speech of learners’ during the first month (Time 1) and last month (Time 2) of the intensive program using a picture-narrative task based on The Suitcase Story (Derwing et al., 2004). Time 1 and Time 2 speech samples (M = 32 sec) were then randomized and rated on each of the three dimensions by native speakers (n = 5) and expert ESL teachers (n=5) using a 7-point Likert scale. Oral production was also
assessed by students of the same age (n=19) using questions based on the perception of the production.

Native speakers and expert ESL teachers’ ratings revealed that learners’ fluency, comprehensibility and accentedness improved significantly during the program (ps < .001), with effect sizes being greatest for comprehensibility development (expert, d = 1.79; native, d = 1.69). The analysis of grade 6 native speakers’ perceptions of grade 6 ESL learners’ speech production reveals that fluency and vocabulary seem to be important factors affecting the perception of speech production.

These results provide new empirical evidence that L2 learners’ comprehensibility, fluency and accentedness can improve significantly over the school year in intensive settings from the perspective of different raters, but it also sheds light on factors that may influence children’s perception of L2 speech. Theoretical implications for these findings are discussed.

References


Sandra Götz, Christoph Wolk and Katja Jäschke. Fluency across L1s, Task Types and Learning Context Variables: A learner corpus-based perspective [3]

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

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

References


Translation and language acquisition have always been closely linked except during the period that followed the grammar-translation method in the 1980s. At that time, translation was considered as the worst tool for language learning because of the grammar-translation method’s mechanical drills. Since then, translation has made a come-back in the foreign language classroom and is now used in different ways, but mainly following pedagogical translation methods rather than translation studies (TS) methods. TS borrows from many disciplines including corpus linguistics, which entered the domain in the 1990s namely with Mona Baker. Baker started applying corpus linguistics’ methods to descriptive translation studies and defined different types of multilingual corpora as well as their usage. This paper presents a study that links translation, foreign language writing, and corpus studies. It has been conducted to test if the introduction of translation strategies in a French composition class impacted students’ writing in French or in other words to see if it induced writing that was more natural. There were three classes used with a total of twenty-four participants: a pilot class in which no translation strategies were introduced, a class where strategies were briefly presented, and an experimental class in which translation strategies were explicitly exposed during class sessions and were used and worked on. Students’ compositions were marked for errors and natural instances by two native speakers: the natural instances, including those that were not completely validated by the native speakers were then checked in a French concordancer “Concordancier-corpus français” created from the newspaper Le Monde’s articles. When some instances could not be verified in the concordancer, they were entered in Google.fr to see how many hits would come back. Even though the status of some occurrences remained the sole decision of the native speakers, as they had no matches in the concordancer or online, the overall results showed that students’ writing improved—or became more natural in the experimental class after the introduction and use of translation strategies.

Yuying Hu and Ziwei Cheng. An Initial Investigation into Discourse Connectives in Written Texts by Chinese and Native ESP Learners [5]

This corpus-based study explores the use of discourse connectives (DCs) in the theses written by Chinese MBA writers and their American counterparts. There are two components of the corpora under investigation. One is the specialized corpus consists of 30 MBA theses produced by Chinese MBA
students. Another component is chosen from The Michigan Corpus of Upper-Level Student Papers (MICUSP), which is functioned as the reference corpus of the Chinese MBA students’ writing data. Drawing on the Liu’s (2008) categorizing framework of DCs, features of the overall frequency of DCs used in two corpora is firstly outlined. Secondly, frequency features of top 10 DCs used in both corpora is examined to look at what are the frequently used DCs in two corpora. Then, the overall distributing features of semantic categories of all DCs in two corpora is investigated, followed by an in-depth examination of three frequently used semantic categories to find out similarities and differences of the DCs used in two corpora. Findings from the quantitative and qualitative analyses are summarized as follows. Overall, learners from both corpora tend to employ a set of limited DCs, possibly due to the narrow topics included in the disciplinary writings and the need of elaboration, reasoning and clarification in the specialized discourse. Comparatively speaking, Chinese MBA learners are cling to a small group of Teddy Bear DCs to organize ideas and establish logical thinking patterns to impress their readers. In terms of the overall features of semantic distributions, learners in two corpora share the uses of three semantic categories. They are types of conditional causal, linking adverbials and dismissal. Case studies of the most frequent use of semantic categories reveal three implications. First, compared with their American counterparts, Chinese MBA learners display a rather weaker performance in employing hedges to make claims. Secondly, compared with their American students’ skillful strategy to remove the readers’ reading obstacles due to abstract and complex concepts in academic writings, Chinese MBA students seem ignore the writing principle of reader-friendly. Last, American students display a stylistically appropriate feature of employing nominalization in their writings, whereas the Chinese MBA students perform poorly in applying nominal expressions in academic writings. Implications from this comparative corpus investigation into the use of DCs by learners and their native counterparts could be as follows. Pedagogically, findings from this study would be useful for ESP teaching field in terms of writing, reading, syllabus design as well as materials writing. Methodologically, this research could shed light on the specialized learner corpus, which remains marginal in the field of corpus research. Lastly, from the perspective of intercultural study, findings of this research would also be insightful for ESP field, particularly the future development of Chinese ESP field.


Producing error-free texts is challenging for the majority of L3/Ln learners. Research has shown that teachers tend to correct all errors at all levels while continually seeking more effective ways to support their learners in writing (Arntzen, Håkansson, Hjelde, & Keßler, 2019). On the other hand, very few textbooks or materials used in communicative settings usually take into consideration the learner’s background in terms of languages previously learned. Recent research has although suggested that better awareness of the associations between the languages learners know could help them and their teachers to make more conscious, appropriate and effective connections among the languages in the classroom context (Orcasitas-Vicandi, 2019). However, the study of CLI to an L3/Ln potentially originating from related and unrelated previously learned languages to support teachers’ decisions in terms of pedagogy and feedback in writing remains an underresearched area, especially in the acquisition of minority languages such as Catalan.
By exploring three different learning paths of L1 French learners of L3/Ln Catalan, this study set out to investigate the potential CLI of related and unrelated languages on written production errors. The analysis was conducted using different perspectives: the linguistic aspects of errors (such as orthographic, morphologic, syntactic and semantic) the modification types (overinclusion, omission, misordering, misselection and blends) and the intralinguistic causes and interlinguistic influences (James, 2008; Corder, 1971; Ellis, 2008).

The corpus is based on texts written by L1 French intermediate (B1) adult learners of Ln Catalan (French-Spanish-Catalan (n=9), French-English-Catalan (n=9), French-Spanish-Roman languages-English-Catalan (n=23)) taken from a standardized exam of Catalan. Students had to write a letter to a friend describing his/her habits and environment. The analysis was conducted using the qualitative analysis software Nvivo. Texts were thoroughly read, and nodes were created in the process of generating a general and specific analysis for each language path.

Preliminary results show that the nature of written errors varies for all three paths. Pedagogical implications for supporting L1 French learners at this level are presented and discussed.

References


There are indications that the use of corpus methods does not occur often in English language teaching in Norwegian schools (Cardona, Didriksen & Gjesdal, 2014; Kavanagh, in preparation), which would not be a situation unique to Norway (Mukherjee, 2004; Callies, 2019). This is despite the variety of beneficial effects of using corpora in language teaching (Lin & Lee, 2015; Boulton & Cobb, 2017). It is necessary to find out more about what lies behind this lack of use, if the benefits are ever to be realized in the school environment. Past research has obtained insights from language teacher educators (Breyer, 2011), teachers in higher education (Lin & Lee, 2015), and pre-service teachers (Breyer, 2011). Somewhat missing has been the in-service teacher perspective, where in-service teachers play the key role.

This qualitative study aims to cast some light on this. Forty-five in-service English teachers received corpus training from the author, and the data consists of follow-up interviews conducted with four of
them. The research question is: What do in-service English teachers find useful about corpus methods and what do they find challenging?

There are three findings in the study that are different to previous research and perhaps speak to the unique perspective of in-service teachers. First, informants did not perceive most of the areas of usefulness highlighted elsewhere (Ebrahimi & Faghih, 2016: 128; Leńko-Szymańska, 2017: 233). Second, teacher workload has been considered a challenge to corpus use (Breyer, 2011: 150; Lin & Lee, 2015: 270-1) but these informants did not mention it. Third, informants talked about language learners in a way not mentioned in the previous research, in terms of them learning a language without being interested in language.

Conclusions: informants can see some use for corpus linguistics, so corpora can have some role in teaching; known challenges to corpus use are confirmed here by in-service teachers; and the challenges must be overcome in real-life schools and there is little use in linguists being evangelical about corpora in teaching until these challenges are met.

References


Kavanagh, B. In preparation. Bridging the gap from the other side: the uses English teachers have found for corpus methods.


CAT is a Corpus of Russian Academic Texts that consists of recently published scientific articles enriched with metadata, morphological and syntactic annotation (see Kopotev et al., 2019). The CAT is used as a reference corpus to automatically evaluate a novice student’s paper against the academic standard. We provide a web service that helps students seeking to improve their writing skills, by getting automatic feedback on the text. As a first step of analysis, the system provides a general analysis of the novice text; the second step involves a fine-grain analysis that analyses three broad areas, found challenging for learners of Russian; 1) lexical knowledge (e.g. unattested lexemes), 2) grammatical knowledge (e.g. overuse of the Genitive case), and 3) collocational knowledge (e.g. well-formed, but unattested collocations). Our project is still a work in progress. The current report is focused on general analysis, where along with the standard measures, we also apply those that focus on academic features of text.

Readability represents how easy it is to read a text. The Flesch Reading Ease and the Flesch-Kincaid Grade level are used, both adopted for the Russian language (Oboroneva, 2006) and specifically for academic texts (Solovyev et al., 2017).

Lexical density is a proportion of content words in a given text. It is determined as the number of tokens such as nouns, verbs, and so forth to the total number of tokens (Biber, 2007: 97–98).

Syntactic words rate is a ratio of syntactic POSs (e.g., then, moreover, so far) in the text (Bach, 1974). The idea behind this metric is to define how meaningful, or meaningless, is a student text if to compare it to an average academic usage as it is reflected in the CAT.

Frequent and infrequent words rate is a feature that represents how common is a token both in standard Russian and in a domain in which a student is majoring. The relative frequency of the tokens is compared with the χ2 test against overall and domain frequencies.

Long or short word rates are two inverse ratios, which show the tokens, which are significantly longer, or shorter than their median length in CAT. The idea is that novice students tend to overuse frequent, thus often shorter, words instead of longer ‘bookish’ synonyms.

Lexical complexity of the text is calculated as a number of rare tokens to the overall corpus size. The less frequent a token is, the higher is its complexity (Zeno et al., 1995).

The working model, current capabilities, and the plans for improvement of the tool will be discussed during the presentation.

References


The paper discusses prospects of applying corpus linguistics approaches to formation of lexico-phraseological skills in linguistics students. As it is widely known, corpus linguistics is a whole methodology that can be applied to many linguistic problems.

The studies of the last few decades have shown that to date in modern corpus linguistics there have been formed two ways to using corpora in linguistic research: one is based on a corpus (corpus-based approach), and the other is driven by corpora (corpus-driven approach). The involvement of corpora in the study of idioms and phraseological units is also crucial because the corpus allows setting frequency indicators for the use of idioms, which can be extremely important and helpful not only for the lexicographer, who compiles a dictionary entry based on the degree of phraseological units use, but also for a teacher of a foreign language, and linguistics students (see, e.g. Geluso & Yamaguchi, 2014).

Since pioneering works by Mindt (1987,1996) revealing that German EFL textbooks contain ‘a kind of school English which does not seem to exist outside the foreign language classroom’ a number of studies comparing textbook corpora with “authentic” native corpora have emerged. The authenticity of frequencies, functions and co-occurrences of grammatical and lexical features in EFL textbooks has been questioned by researchers.

So, the first research question we sought to answer was if all the idioms included in teaching materials recommended for Linguistics students at Peter the Great St. Petersburg Polytechnic University can be found in the Corpus of Contemporary American English (COCA).

If yes, what is their frequency and in which modifications are they fixed in the corpus?

Totally 535 phraseologisms and idioms were selected for the verification from the General English textbook Language Leader Upper-Intermediate, the Manual for Translation Practice, and Illustrated Russian and English Proverbs and Sayings manual.

We have shown that the COCA is relevant for this kind of research as of the 535 phraseologisms and idioms 262 were found in their canonical form and 71 – in modified forms. The paper contains the list
of the most frequent phraseologisms, the list of the phraseologisms from the Manual for Translation Practice not found in the COCA, and the list of typical modifications of phraseologisms. These results could be considered as the first step to a more complete study of the variation of these phraseological units, to further systematization of knowledge about the relationship between the fixedness of lexemes and the degree of their transformation.

Analysis of phraseologisms revealed the following typical modifications:

1) expanding the composition of a phraseologism by introducing an additional lexical component;
2) reduction, or reduction in composition, of phraseological units;
3) a change in grammatical form;
4) lexical variation of the components of phraseological units.

The corpus COCA can be used as a valuable resource in the study of phraseological units, as well as a source of authentic examples for the development of various kinds of exercises and corpus tasks of a new type for linguistics students. Tasks can concern both a simple check of the presence of a certain phraseological unit in the corpus and, therefore, its usability in modern English, and a more subtle analysis aimed at identifying the frequency of phraseological units use, the context of their use, changing the frequency over time, the presence of recorded transformations of phraseological units, which should lead to an increase in linguistic cultural, and computer competence of students. Methods of corpus linguistics provide significant assistance in the teaching process and should be used to create effective manuals for teaching foreign language phraseology. This was shown during the experimental teaching first year linguistics students and evidenced by the results of entrance, final and delayed tests as well as the survey results.

References


Whilst the value of corpora in foreign language teaching has long been understood, researched and documented (e.g., O’Keeffe, McCarthy & Carter, 2007; Boulton & Cobb, 2017), the “large gap between the wealth of applied corpus-linguistic research and the teaching practice” (Mukherjee, 2004: 247) is still a reality. In particular, in school EFL contexts, the “need to convince practising teachers to use corpora and concordances in the classroom” (Römer, 2006: 129) has yet to have been met. To close this gap, a number of researchers have pointed to the centrality of teacher training (e.g., Mukherjee, 2004; Hüttnner, Smit & Mehlmauer-Larcher, 2009). Following Römer’s (2010) call to focus more on language teachers’ needs and inspired by previous such endeavours (cf. Breyer, 2009; Hüttnner, Smit & Mehlmauer-Larcher, 2009; Leńko-Szymańska, 2014), a project-based seminar for M.Ed. students
A training to become English teachers (at primary, secondary and vocational school level) in Lower Saxony (Germany) was developed and tested. The institutional constraints were: 13 weekly sessions of 90 minutes with ca. 30 students. The aim was to convince teacher trainees of the value of using corpora in their teaching and to empower them to do so autonomously using a range of tools and methods. In the pre-intervention questionnaire, two-thirds of the students agreed or strongly agreed with the statement: “I have no idea what corpus linguistics is about”. Only one participant reported already using corpora. All students responded negatively to the statement: “I know how to use corpora to prepare classroom materials”. The first few sessions consisted of a hands-on introduction to the basic principles of corpus linguistics, including debunking some normative linguistic myths (cf. “surprise-the-teacher” modules suggested by Mukherjee, 2004: 245) using available corpora and various offline and online tools. Following a Flipped Classroom (e.g., Reidsema et al., 2017) and Just-in-Time Teaching (e.g., Simkins & Maier, 2010) approach, students “consumed” theoretical input (in the form of journal articles, book chapters and videos) in their own time and answered questions testing their understanding and asking them about their interests. These answers informed the content of future sessions. The advantages and limitations of DDL were introduced and students explored the alignment of the DDL approach with current SLA principles. Students were encouraged to think about the difficulties that non-native ELF teachers typically face and, using a problem-solving approach, gathered ideas to develop appropriate corpus-based materials and tasks to help solve some of these issues. In lieu of an end-of-term examination or seminar paper, course assessment consisted in the contribution of a co-authored chapter to a “Practical Guide to Using Corpora for English as a Foreign Language Teachers”, to be published online. The chapters submitted ranged from using the BNC1994 to help non-native teachers correct non-idiomatic collocations in their students’ writing, to building specialised corpora to create ESP tasks for vocational schools, and using a corpus of TED talks to teach public speaking skills at secondary level. This paper presents the outline and intended outcomes of the seminar. To evaluate its effectiveness, the results of the pre-intervention and post-intervention questionnaires are compared and students’ course evaluation and self-reflection statements are analysed. Furthermore, the chapter contributions are examined to identify the weaknesses and limitations, as well as potential extensions of this project-based approach to teaching corpus linguistics for language education in pre-service teacher training.

References


While it can be claimed that large reference corpora and smaller genre-specific collections have not quite entered the mainstream language education yet, parallel resources are used even more scarcely in the FL classroom. In addition to the usual obstacles including limited availability of good-quality parallel resources and technical and cognitive challenges related to handling software, an important concern has been the use of L1 and translation in the process of language learning, which in recent decades has been discouraged by the current communicative methodology. However, translation not only can be a valuable instructional technique (Cook, 2006; Tsagari & Flores, 2013) but it is also a key target skill in language acquisition for learners willing to choose it as a future profession. Unfortunately, so far few studies have addressed the issue of effectiveness of using parallel corpora in language and translation training as well as students’ perceptions of the usefulness of this resource (cf. Zanettin, Bernardini & Stewart, 2014).

This presentation will report on a project which had exactly these objectives. It will give an account of several types of parallel-corpus-based activities used in a semester-long course offered to 21 Polish students of Swedish who were training to become professional translators. The students worked with a parallel corpus of Swedish contemporary literary texts and their published Polish translations consisting of approximately 5 million tokens (26 books). In the first part of the course the students’ task was manual verification and correction of the automatic alignment of the corpus. In the second part, the students worked on their own corpus-based projects involving explorations of specific grammatical, lexical and discoursal points in Swedish and the ways of rendering them in Polish by professional translators. In the final part of the semester students had a chance to present their results in class and listen to the presentations prepared by their classmates. In addition to these corpus-based activities, the students were engaged in a variety of other non-corpus-based language and translation task during the course.

A rich set of data were collected in the study. In the first class the students took a diagnostic test of their reading skills in Swedish. They also completed a short questionnaire focusing on their knowledge of and prior experience with corpus-based resources as well as their reading habits related to foreign language literature. Next, their projects were collected and scrutinized with a focus on the specific corpus data students had searched for and the types of analyses they had performed. Two class observation sessions were conducted when the student presented their projects. Finally, a two-part
questionnaire was administered to the students in the last class eliciting their perceptions of and attitudes to the verification of the aligned texts and the language projects based on the parallel corpus.

The results indicate that not all the activities were equally appreciated by the students. They did not enjoy reading aligned texts and did not consider it a useful exercise. On the other hand, they valued working on their own corpus-based project, in which they explored specific language and translation points. Interestingly, they did not recognize the value of listening to the presentations given by their peers.

References

Ting-Shiu Lin. An error analysis of Chinese modal verbs of possibility in French learners’ interlanguage and its pedagogical implications [3]

It is well known that modality is difficult for language learners to master: most modals are polysemous, and one modal meaning may be expressed by several words. Within the field of teaching Chinese as a foreign language, many studies have investigated learners’ errors involving usage of modals (Niu, 2012; Wang, 2018; Xiong, 2013; etc.). However, few of these works concern errors produced by native speakers of French.

This paper examines French learners’ errors related to two synonymous modals: néng ‘can’ and kěyǐ ‘can, may’, both of them usually translated as pouvoir in French. Data under analysis come from texts produced by 58 second-year and 21 third-year undergraduate students learning Chinese in a university in Paris. 195 sentences in which néng or kěyǐ were misused, wrongly omitted or replaced by an inappropriate word were found from these texts.

The types of errors most frequently found in our data are as follows. First, learners often confused kěyǐ with kēnéng ‘maybe’ (1). Kēnéng is an epistemic adverb expressing the speaker’s judgment on the probability of the occurrence of a situation, while kěyǐ can only be used in non-epistemic contexts (Sparvoli, 2017: 41).

(1) a. Wǒmen kēnéng/kěyǐ yǒng lìzì shuōmíng. we KENENG/KEYI use example illustrate ‘We can illustrate this point with some examples.’
   b. Wǒmen kēyǐ/kēnéng huán-shēng yánzhòng jíbìng. we KEYI/KENENG catch-up serious disease ‘We may get serious diseases.’

Second, learners sometimes omitted néng in sentences where its presence is necessary (2). In fact, the appearance of néng is much less frequent than that of kēyǐ in our data.
The misuse of the negative form of *kěyǐ* or *néng* is the third most common type of errors found in our data. Some learners were not aware of the fact that the negative form of *kěyǐ* has only the interpretation of prohibition (Alleton, 1984; Sparvoli, 2015) and used it to indicate the impossibility of the realization of a situation (3-a). The negative form of *néng* can be used to express the impossibility to carry out an action (3-a) but are rarely used to indicate the impossibility to achieve a result (cf. Lu, 2004), in which case a potential infix should be adopted.

Based on our findings, we put forward some suggestions that may help to improve the teaching of the modal verbs *kěyǐ* and *néng* in Chinese class.

References


Spoken Academic English (SAE) has made itself a unique variant of Academic English. That is, it shares some features of academic prose in general, often assumed as a variant of Written Academic English (Biber, 1988; Hyland, 2009) and overlaps with the features of varieties included by broader definitions of SAE (Simpson, 2000; Swales & Burke, 2003:2). And the increasing popularity and significance of Massive Open Online Courses (MOOCs) in higher education may give rise to an identifiable subgenre within SAE, referred to as “EMP” (English for MOOC purposes) (Anthony, 2015). Although vocabulary is of critical importance to language learners, one dimension of vocabulary knowledge—grammar patterns, defined as “all the words and structures which are regularly associated with the word and which contribute to its meaning” (Hunston & Francis, 1999: 36)—has received little attention in research. Grammar patterns are developed based on the finding that lexis and grammar should be considered together, but it has long been common to treat grammar and vocabulary separately (Coxhead, 2008). Given the discrepancy between the importance of grammar patterns and scarcity of empirical studies in the research of SAE, this project was designed, following Hunston and Francis’s 1999 book Pattern Grammar. In this paper, the MOOC corpus with over 11 million tokens from 124 courses divided into 5 domains was first collected, compiled, annotated and tagged via Spacy, a library in the programming language Python, into machine-readable texts. A python script was later created to extract the grammar patterns based on the content in the book Collins COBUILD Grammar Patterns 1: Verbs and 2: Nouns and Adjectives. After extracting the grammar patterns of the MOOC corpus, the approach of Collostructional Analysis—specifically, distinctive collexeme analysis (see Stefanowitsch & Gries, 2009), a collocational method for investigating associations between words and alternating grammatical construction-- is undertaken to retrieve the most significant and least ones in MOOC corpus and each 5 separate categories, with the help of the package collostruction (Flach, 2017) in the programming language R. The results show that there are 220 grammar patterns in MOOC corpus in total, of which Noun patterns has 55 ones; Adjective patterns, 52; Verb patterns, 113. And the first 5 patterns with high occurrences are \textit{V}, \textit{ADJ n}, \textit{the N}, \textit{n N}, and \textit{det ADJ n}. And in terms of disciplinary characteristics, such domains as Arts-Humanity, Comp-math and Physic-science, prefer N patterns as their main attracted ones, for example, \textit{the N} (e.g. the world, the story, the number, the function), \textit{N of n} (e.g. point of view, lot of time, example of this, kind of thing), or \textit{num N} (e.g. two functions, two lines, one example), or \textit{n N} (e.g. gradient descent, carbon dioxide, voltage source); disciplines like Life-science and Social-science prefer V patterns as their attracted ones, for instance, \textit{V to-inf} (e.g. go to, want to, have to, begin to..), \textit{V prep} (e.g. look at, come to, happen to, talk to, think about), or \textit{V inf} (e.g. make do it, let zoom in).

References


The distribution of post-verbal subjects in Italian is governed by lexicon-syntax (with unaccusative verbs; Belletti, 1988) and syntax-discourse (information-focus on the subject; Belletti, 2004) interface conditions. In L2, phenomena related to the former interface are more easily attained (Sorace, 2011). Accordingly, the performance of advanced L2-learners of Italian with VS differs depending on whether the subject follows an unaccusative verb or is focus-marked (Belletti et al., 2007; Lozano, 2006 on L2-Spanish). However, it remains unexplored how VS emerges at lower levels of proficiency and whether the divide between lexicon-syntax and syntax-discourse interface phenomena is reflected at the developmental level (cf. Bettoni et al., 2009; Nuzzo, 2015 for a similar perspective).

We conducted a study on the LIPS Corpus, that collects transcriptions of language assessment exams of learners with varied L1 and proficiency (A1-C2) (Vedovelli, 2006). After dividing the transcriptions in units (based on the occurrence of a finite verb) and isolating VS-structures (Table 1), we classified them following a multi-layered annotation of verbs and subjects. We considered verb-type, distinguishing between transitives, unergatives and unaccusatives and “piacere (like)-type” verbs. The latter are ThemaNOM-ExperiencerDAT verbs exhibiting VS as the unmarked word order. We also consider semantic (agentivity, Bambini & Torregrossa, 2010) and information-structure features of the subject. To avoid arbitrariness in coding topic and focus, we relied on the inventory of topic- and focus-associated functions reported in Riester and Baumann (2013). The former includes givenness, while the latter includes contrast with a set of alternatives and occurrence with focus-operators.

We ran an ordinal logistic regression, in order to understand which syntactic and semantic features on the subject and verb of a VS-construction allow us to predict the proficiency level of the L2-speaker who produced this structure. In other words, we considered the proficiency levels (from B1 to C2 – A1 and A2 were excluded due to the few occurrences of VS-structures; see Table 1) as the ordinal dependent variables of the model, while type of verb, agentivity, givenness of the subject, contrastivity of the subject and length of the DP as independent variables. We observe that verbs of the “piacere-type” ($\beta = -0.82$, SE = .33, $t = -2.48$, $p = .013$) – whose relative frequency decreases from B1 to C2 – and transitive verbs ($\beta = 1.15$, SE = .31, $t = 3.74$, $p < .001$) – whose relative frequency steadily increases – are significant predictors of the learners’ proficiency level, while this is not the case for unaccusative verbs ($\beta = 0.34$, SE = .24, $t = 1.45$, $p = .148$). We also find an effect of agentivity ($\beta = 0.47$, SE = .23, $t =
2.05, \( p = .041 \)) and contrastivity (\( \beta = 0.38, SE = .17, t = 2.22, p = .026 \)) on the subject, showing that at higher levels of proficiency, subjects tend to be more ‘contrastive’ and less ‘agentive’. Finally, there is an effect of DP-length (\( \beta = 1.10, SE = .49, t = 2.24, p = .025 \)), indicating that postverbal subjects tend to be longer at higher levels. Figure 1 shows that the higher is the proficiency, the greater is the probability to observe a contrastive subject in association with unergatives and transitives.

We also counted the number of VS-structures produced with incorrect subject-verb agreement. This analysis shows that there is a decreasing number of errors across levels with both “piacere-type” verbs (30% of errors in A1-A2; 18.2% in B1-B2 and 8.7% in C1-C2) and all other verb classes (25.2% in A1-A2, 19.2% in B1-B2 and 12.8% in C1-C2).

Overall, our results comply with the hypothesis of a developmental trajectory from lexicon-syntax to syntax-discourse interface phenomena. The formulaic use of VS with “piacere-type” verbs corresponds to the majority of VS-structures produced at the levels A1 and A2 and seems to function as “pivot” for a syntactic schema that is later extended to other verb types. Crucially, the error analysis shows that at these levels, postverbal constituents are not necessarily assigned subject-status. The absence of any difference in the amount of VS-structures with unaccusative verbs across proficiency levels suggests that VS as an unmarked word order (with unaccusative verbs) is readily attained in L2 Italian. On the contrary, the increasing extension of VS to transitives as well as the parallel increasing occurrence of contrastively focussed postverbal subjects denotes learners’ developing sensitivity to the informational properties of the subject constituent. Finally, the error analysis as well as the patterns observed with the agentivity of the subject show that postverbal constituents are gradually assigned subject-status, which is a necessary condition for discourse-driven subject-verb inversion. At intermediate stages of proficiency, the identification of “subjecthood” is facilitated by the prototypical feature ‘agentivity’, while at the highest level of proficiency learners recognize postverbal constituents as subjects regardless of their semantic prototypicality.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. exams</td>
<td>19</td>
<td>26</td>
<td>60</td>
<td>61</td>
<td>60</td>
<td>63</td>
</tr>
<tr>
<td>N. units</td>
<td>252</td>
<td>569</td>
<td>1665</td>
<td>2444</td>
<td>2255</td>
<td>2784</td>
</tr>
<tr>
<td>N. VS</td>
<td>26 (10,3%)</td>
<td>40 (7%)</td>
<td>134 (8%)</td>
<td>155 (6,3%)</td>
<td>150 (6,6%)</td>
<td>148 (5,3%)</td>
</tr>
</tbody>
</table>

Table 1: Total number of units and VS structures (per level)
**Figure 1:** Distribution of probabilities related to the occurrence of a contrastive subject (contrastività=1, in the Figure) in association to different verb classes (copular, unaccusative, unergatives, piacere-type and transitive, respectively), across proficiency levels.

**References**


The Chinese resultative constructions (see 1 below) is formed as verb + result XP (the latter designates state or location), with the result XP referring to change of state or location of an NP caused by an action denoted by the verb (Shi, 2008).

(1) ta  zou-lei-le
    he    walk-tired-ASP’
‘He walked and as a result he was tired.’

Chinese resultatives are probably one of the most challenging types of constructions for L2 Chinese learners (Yu, 2003). Previous studies have shown that learners of Chinese have difficulties in proficiently using Chinese resultative constructions (e.g. Yuan & Zhao, 2010). Corpus use has been argued to offer an effective teaching method in language learning. However, there is a gap of applying corpus directly to learn languages other than English (Vyatkina & Boulton, 2017). This study, therefore, focuses on investigating the effect of corpus use in learning Chinese resultative constructions by learners of Chinese.

Six learners of Chinese studied in a Chinese university participated in this study. The Chinese proficiency level of these learners was B2 to C1 based on the results of the Chinese Proficiency Test, Hanyu Shuiping Kaoshi (HSK). This study employed a pre-test and-post-test research design and a think-aloud protocol. Learners used the prepared paper-based materials drew from Lancaster Corpus for Mandarin Chinese (LCMC), which is a written corpus with POS tagging and contains one-million words from 15 genres (McEnery & Xiao, 2004).

Results indicated that learners made significant improvements in a sentence-combining test, although no significant changes were found in a gap-filling test. Moreover, learners employed different cognitive and metacognitive strategies at different stages of learning procedure when using concordance-based materials. This study contributes to the empirical study of direct corpus use targeting at a different language, Chinese. Moreover, it helps us understand the effectiveness of corpus use for teaching resultative constructions in Chinese and suggests that the incorporation of corpora into language learning could enhance Chinese learning. It also complements previous studies by using think-aloud to track learners’ actual use of corpus and reflect learners’ strategies in the process of using corpus materials.

References


Back in 1988, Heaton (Heaton, 1988: 37) stated mistake processing as one of the feedback forms. Alongside with that, “Only the scoring of a test can be described as objective” since “such tests can be scored mechanically” (ibid., 25). Returning to the methodological roots of pedagogy, we are glad to admit that our current research project meets both of the above-mentioned principles; to be more precise, our Learner Corpora Lab has been focused on automatic essay-/test-based test generation.

The students’ progress in learning grammar of L2 is often evaluated with help of generic tests. This way, the Language Instructor inevitably fails to track the learning progress, neither at the group, nor at the individual level. However, a solution exists; namely, it has to do with deployment of the learners’ written answers in order to establish the exercise and test frameworks. Moreover, nowadays, the learner progress can be automatically traced within each grammatical domain.

At the current stage, we have used authentic essays from the Russian Error Annotated Learners of English’ Corpus (http://realec.org). We have compared the student’s performance in the format of grammar tests to and with the ratio of respective mistake categories in the essays. One of our spotlights is investigation of Tense Choice precision when completing grammatical versus academic writing related tasks.

Our automated toolbox can be extremely useful both for experimental and educational purposes, especially in line with the principles of data-driven learning. A combination of the tool for extracting erroneous contexts from the corpus with a visually-adapted user interface, indeed, is bound to become a great contribution for performance tracking. The faculty would be able to easily create task sets using only relevant mistake categories, without any confusion with other constituents. Last but not least, our project has been designed not only to provide automated test generation, but also to investigate whether the grammatical mistake ratio in essays is lower than that in tests. Shall that be proven, a new, corpus-based approach to mistake analysis would be implementable.

References

Philippe Millot. Building specialised corpora to conceptualise learners’ specialised needs: The case of a corpus in records and information management [2]

Over the last few decades, French universities have been going through many changes such as the increasing professionalisation of university diplomas which, in turn, has led to the equally increasing specialisation of learner’s needs in English. In this context, specialised corpora have played a growing function such as providing pedagogic material or informing teaching practices regarding what to teach to students who specialise in areas of knowledge which might be largely unknown to language teachers (Gotti & Giannoni, 2014). When one looks at the very large amount of corpus studies in business and professional discourse analysis, one immediately realises that most of these specialised corpora actually concentrate on specific genres such as emails, meetings, corporate reports or mission statements, and the corpora which cover the various discursive realities belonging to whole professional domains are rather scarce. The scarcity of such types of corpora remains an obstacle to the design of English syllabi targeting learners specialising in what may sometimes be considered as highly specialised domains.

In this paper, we present a methodological and experimental framework for building corpora in professional or occupational domains through the case of a specialised corpus in records and information management. The framework is based on what we consider as three fundamental categories in professional discourse (Millot, 2019), namely “theory” – texts in which professional concepts and theories are created and discussed, “mediation” – texts mediating theoretical or highly specialised knowledge to the professional community, and “practice” – texts which directly contribute to the realisation of professional actions. This framework was used to build a specialised corpus in the domain of records and information management containing 1.4 million words. The corpus is then used for two main purposes. Firstly, it provides a tailor-made database from which authentic texts may be retrieved and used for teaching purposes. Secondly, the corpus provides an entry point for the identification and the characterisation of the “specializedness” of records and information management thanks to corpus tools such as WMatrix (Rayson, 2008) and AntConc (Anthony, 2014). We will present how the specialised features were retrieved from the corpus and organised into three semantic dimensions (i.e. “organisational”, “technical” and “promotional”). We will finally show how these dimensions may be used for structuring linguistic knowledge for the learners who specialise in this domain.

References


Millot, P. (2019). Retrieving the specialised substance from a corpus of professional discourse in the field of Records and Information Management. ASp, 76, 49–70.

Learner corpora can reflect pedagogical characteristics of EFL learners in their second language acquisition and learning processes. Identifying mistakes and differences between corpora of native and non-native speakers can provide empirical analysis of lexico-grammatical pattern. While employing lexico-grammatical patterns, EFL learners are likely to use complement clauses erroneously in their spoken and written discourse (Wulff and Gries, 2004). In this regard, this study aims to investigate the frequency of complement clauses, that is, that-clauses, wh-clauses, ing-clauses and to-infinitive clauses, suggested by Biber et al. (1999) in written and spoken outputs of native speakers and Turkish EFL learners. As a written corpus, Turkish International Corpus of Learner English (TICLE) and as a spoken corpus, Louvain International Database of Spoken English Interlanguage-Turkish (LINDSEI-TR) are used. TICLE, consisting of 196900 words, contains written products of students in ELT Departments of Cukurova, Mersin and Mustafa Kemal Universities in Turkey (Can, 2009; Kilimci and Can, 2009). LINDSEI-TR B-turn, consisting of 63968 words, contains spoken products of students in Cukurova University, Turkey (Kilimci, 2014). The spoken reference corpus is LOCNEC (De Cock, 2004) and the written reference corpus is LOCNESS (Granger, 1998). This study also employs Sketch Engine (https://www.sketchengine.eu/) (Kilgarriff et al. 2014) in order to compile and tag the corpora. In order to determine whether there is a statistically significant difference between corpora, log-likelihood value is used.

It can be seen that, in their written outputs, Turkish EFL learners tend to overuse certain patterns. For instance, “start + to-clause” and “want + to-clause”, which are parts of “verb + to-clause” pattern, are overused in TICLE. Log-likelihood values are +51,35 and +65,97 respectively. “think + that-clause” is also overused in written corpus and log-likelihood value is +114,78 when compared to native speakers. “like + ing-clause” is also overused and log-likelihood value is +07,42. However, in TICLE, there are some underused patterns such as “begin + ing-clause” with -14,16 log-likelihood value. Moreover, in spoken corpus, EFL learners tend to overuse “want + to-clause” pattern with +141,59 log-likelihood value. Nevertheless, certain patterns, such as “keep (on) + ing-clause”, are underused by Turkish EFL learners. Log-likelihood value is -7,77 for that pattern.

It seems that there are also some misused patterns in written and spoken corpora of Turkish EFL learners. For example, in spoken corpus, it can be observed that “stop + ing-clause” pattern is used inaccurately when concordance lines are examined. They may use both ing- and to-clause at the same time in their speech. EFL learners use these words: “...stop to comparing...” in their spoken discourse. It can be understood that, unlike native speakers, EFL learners use both “to-” and “ing-clause” together after the verb “stop”.

These findings indicate that Turkish EFL learners and teachers need to focus on the use of complement clause in speech and writings. This study can help learners, teachers and researchers to raise awareness of complement clauses in EFL setting.

References

Chikako Nishigaki and Kathryn Oghigian. DDL for Young Learners: Development of a Pedagogical Corpus and Search Tool, and Implementation in Japanese Elementary and Secondary Classrooms [1]

In Boulton and Cobb’s 2017 DDL meta-analysis, they found research with high school learners to be limited and, unsurprisingly, no studies were identified at the elementary level. This finding has been echoed by Crosthwaite in his 2019 review of DDL research studies. Those of us who use DDL in second language learning know that complex language in authentic corpora and complicated tools are obstacles for lower-proficiency level learners. Beginning in 2015, Chujo et al. created SCoRE, which is a sentence corpus and tool specifically geared toward remedial English learners at the university level, and they have reported successes with its effectiveness. SCoRE concordance lines are short, simple, complete sentences using level-appropriate vocabulary and familiar contexts and are accompanied by the L1 in a parallel corpus. As SCoRE co-researchers, we have taken this one step further by creating a corpus and tool specifically for young learners.

By implementing the SCoRE rationale as outlined in the 2015 paper by Chujo, et al., we have been able to show that using paper-based DDL in elementary and high school L2 classes with tailor-made worksheets can be effective (Author et al., 2019). Results indicated that students using paper-based DDL showed stronger retention of grammar knowledge and a higher increase in noticing grammatical rules compared with a traditional teacher-centered approach. This use of pre-selected, complete sentences for paper-based DDL enables learners to access the many benefits of DDL. Current preliminary data also shows that Japanese translations placed beside English sentences seem to provide a clear benefit to students because the differences in the discrete language rules between the L1 (Japanese) and L2 (English) in language structure are readily visible. This pedagogically needs-based
The approach is especially effective for activating young learners’ inductive learning of language, since their L2 experience and knowledge is limited.

Encouraged by these findings, we have continued to develop and revise two new corpora and search tools for elementary students (“e-DDL”) and high school students (“h-DDL”). The corpus is based on a 24-million-word source corpus created from introductory-level English material such as graded readers, government authorized school textbooks, children’s stories, and news and similar sources, and is comprised of complete sentences using level-appropriate vocabulary. Tasks and selected concordance lines are paper-based to isolate targeted language goals. To make the concordance lines more relatable to students, e-DDL and h-DDL have cut-out paper-craft characters who act as guides. Many of the concordance lines describe these characters and their school and home lives. Moreover, the concordance lines are accompanied by pronunciation samples. In this paper, we will present the rationale for e-DDL and h-DDL, discuss their development, and present students’ evaluation of this material.

Aleksey Novikov. Lexico-grammatical development of L2 Russian learners [3]

Corpus-based investigations have been shown to be a robust way to investigate learners’ L2 complexity development (Biber, Gray & Staples, 2016; Vyatkina, 2012, 2013; Lu, 2010, 2011). Several studies have shown a trend in increasing subordination with increasing program level (Norris & Ortega, 2009; Kisselev & Alsufieva, 2017), especially at lower levels. Another strand of research emphasizes the importance of phrasal complexity, such as attributive adjectives and noun-noun sequences (Biber, Gray & Staples, 2016). However, the investigation of complexity needs to be expanded in terms of lower levels of proficiency, and other languages than English. Furthermore, morphological complexity is a fairly new construct in complexity studies, and an objective metric is needed (Brezina & Pallotti, 2019). To address these three levels of complexity, this study analyzes a cross-sectional corpus of written assignments in L2 Russian (over 400 texts, over 50,000 words) at lower levels over the period of four semesters in terms of: 1) clausal: subordinate adverbial clauses (e.g. temporal, causative and conditional adverbial clauses) and coordinate clauses (e.g. у/and, но/but, а/and); 2) phrasal: attributive adjectives and noun-noun sequences (post-modifying nouns); 3) morphological: “a diversity of inflectional types of a given word class” (Brezina & Pallotti, 2019: 100), specifically nouns and adjectives in six different cases. The preliminary results show a similar trend with previous studies in terms of clausal complexity, namely that subordination is generally increasing with program level. However, if you look at different subordinate clauses individually, the only clauses that do increase are temporal (e.g. когда/when) and conditional (e.g. если/if) with the causative clauses (e.g. потому что/because) staying fairly stable. These results are interesting in the light of Staples and Reppen (2016) who found that L2 writers tended to use more causative adverbial clauses whereas L1 writers used more conditional adverbial clauses. With regards to phrasal complexity, attributive adjectives display an increase in a range of cases they are used in (grammar) and their lexical realizations (lexis). This finding regarding the interplay between grammar and lexis corroborates with other studies, such as Staples & Reppen (2016) and Brezina & Pallotti (2019) in that what constitutes complexity is not just the frequency of grammatical features but also their lexical realizations. Noun-noun sequences are mostly either non-existent at lower levels or are used as fixed chunks, such as часов вечера/o’clock in the evening. Noun phrase complexity in Russian is closely related to nominal morphology, i.e. a noun phrase consisting of two nouns will result in the second post-modifying noun in genitive case. As post-
modifying nouns increase at higher levels, that affects the use of nouns in genitive case, thus increasing morphological complexity. Overall, nouns in genitive and dative cases increase considerably with program level. The findings of this study lend themselves to making more explicit what learners can achieve at different levels of instruction and can be used in assessment rubrics and incorporated into the curriculum as learning outcomes.

Maria Obeyukhtina. To what extent is it possible to find some generalizations between the development of auxiliary verbs in English and the stages of language acquisition by L1 (or L2) learners? Can such generalizations serve as diagnostics for language change? [3]

This paper discusses how similar and/or different the historical development of the English language and the process of its acquisition can be. Nowadays, auxiliary verbs play an important role in the system of English grammar; thus, it might be of interest to find out which factors affected their development and explore to what extent it is possible to find some parallels between the stages of language development and acquisition data from L1 learners. This paper compares the data extracted from several books on language acquisition (Pinker, Lebeaux, & Frost, 1987; Newport, Gleitman & Gleitman, 1977) so that it can become more evident whether there are some regularities in the mistakes that children make. Another important question is if these incorrect sentences follow any syntactic patterns attested in the earlier periods of the English language history and if the use of these patterns helps achieve any pragmatic purposes in the speech produced by adults. As far as the preliminary findings are concerned, the most common and most significant issues children (and possibly L2 learners) face are as follows:

1) the problem of doubling/omitting auxiliaries (Radford, 1996, Guasti, 2017);
2) the problem of excessive use and overgeneralization of auxiliaries;
3) choosing an auxiliary verb for passive constructions and, in connection with that, the overuse of “get” (Pinker, Lebeaux & Frost, 1987);
4) the issue of word order in children’s speech and the role of auxiliaries ( Culicover, 2008).

It also makes sense to find out if such generalizations can serve as diagnostics for future language change (Lopez-Couso, 2011).

Several examples showing both successful acquisition of auxiliary verbs and some developmental errors are provided, too. It is important to emphasize that there is always a certain percentage of information correctly acquired by children: should it not be the case, they would have never acquired the forms properly as their acquisition skills are believed to heavily depend on the input from either their caretakers or their peers (or both) ( Aitchinson, 2001).

It might be also of interest to see how some other Germanic languages use auxiliary verbs and how small children perceive and acquire these forms in those languages. Therefore, in the discussion section some data of modern Dutch will be provided in order to show which words serve as similar auxiliaries and some typical errors that native speakers of Dutch might make.
As the first steps of the analysis show, historical development of the analytical forms and their acquisition by children are not exactly analogous but some similarities provide basis for further investigation.

References


Internet resources

Etymologiebank. [http://etymologiebank.nl/](http://etymologiebank.nl/)

Etymologyonline. [http://www.etymonline.com](http://www.etymonline.com)

John O’Donoghue. An exploratory investigation into the way in which L2 undergraduate writers perceive the nature of academic writing and the cultural factors determining such perceptions [5]

In academic writing research on the performance of L2 students, it is commonly observed that L2 writers perform differently to L1 writers, especially regarding hedging and epistemic modality. They tend to overstate their position, be overconfident and too direct. Rather than viewing writing as a dialogue they see it as a monologue, overemphasizing their own point of view and leaving little room for the reader to disagree. The solution to this problem is often a prescriptive one, i.e. if these L2 deviant writers are exposed to sufficient L1 writing then they will slowly imitate this standard and approach the desired target level. This approach seems to assume that the writing is taking place within an English-speaking context; this is, however, often not the case. Consequently, the proposed solution fails to take the cultural background of such L2 writers into account and also ignores the institutional setting in which this writing takes place. Questionnaires conducted with several undergraduate student classes of an international business programme in Germany revealed perceptions about academic writing that valued directness, confidence and objectivity. Could it be that
this L2 writing performance is not linguistic deficient but culturally determined and consciously constructed? If this is the case, then instructors need to engage with these novice writers in a discussion which makes these students aware of such cultural factors and their validity in an international context. This debate leads us to issues concerning the ownership of the English language in academia and the rules that do or do not apply when using English as an Academic Lingua Franca.

Anne O’Keeffe, Pascual Pérez-Paredes and Geraldine Mark. Aligning usage-based theory and Data-Driven Learning [1]

A usage-based (UB) model of language learning has gained momentum in recent years (Ellis, 2012). Crucially, from an SLA perspective, the UB model holds that acquisition happens through experience of and exposure to language. In both first and second language acquisition, the model holds that there is an associative process of tallying the probabilities of occurrence of form-function mappings (based on an individual’s accruing experience of language). Structural regularities (i.e. phonology, morphology, syntax and discourse), “emerge from learners’ lifetime analysis of the distributional characteristics of the language input” (Ellis, 2012: 196). Within this model, it is recognised that there are three major factors in the process of language acquisition: 1) frequency (the more times a construction is experienced and eventually used, the better); 2) recency (the more recently we experience a construction, the stronger our memory of it) 3) context (the frequent experience of a construction in a given context triggers an association and mental categorisation). In sum, it is held that frequency of exposure to language and experience of it triggers cognitive processes of categorisation which are seen as core in the process of acquiring a language system and its rules. This paper will argue that for proponents of DDL, the UB model is a very attractive theory to explore both theoretically and empirically. The paper will discuss the synergies between the UB notions of 1) frequency and the importance DDL places on observing form, frequency and patterning; 2) recency in terms of alignments with carefully curated DDL materials; and 3) context and the DDL principle of promoting learning through inductive cognitive processing of language in meaningful contexts. The paper also proposes some insights from the UB model that might challenge and inform DDL, for example, possible opportunities for DDL in the acceleration of learning at lower levels.


Formulaic language has established itself as an integral part of second language learning and use. It has spawned major investigatory efforts in both research and practice across various fields of inquiry such as corpus linguistics, psycholinguistics, and language education (e.g., for a recent overview, see Siyanova-Chanturia & Pellicer-Sanchez, 2019). In the past three decades, a variety of formulaic units have been identified and researched (e.g., lexical bundles, binomials, collocations, congrams, collgrams, phrasal verbs, etc.), and linked to fluent and natural production of both spoken and written language. Despite these efforts, relatively few studies have thus far endeavored to investigate the longitudinal development of the productive knowledge of formulaic sequences in second language (L2)
learner writing. As a result, despite the strong evidence base that the study of phraseology has established in the past few decades, our understanding of how the learner phrasicon develops over time is still limited.

To address this gap, the present study set out to investigate the developmental trajectory of L2 phraseological knowledge by looking at the use of verb-noun collocations in a large-scale longitudinal corpus of essays written by Chinese learners of Italian. For this purpose, we examined three key dimensions of phraseological knowledge, each assessed by a specific measure: exclusivity (Log Dice), phrasal frequency (logarithmic frequency), and phrasal diversity (RTTR). Following this, we conducted three separate LME regression analyses to verify if, and to what degree, the above-mentioned measures varied as a function of time, in combination with other related factors (e.g., proficiency level, prior language exposure) that might influence the production of word combinations in a second language. The results of our analysis showed that L2 phraseological development is a multifaceted process, in which the multiplicity of interacting elements can profoundly affect learner performance over time. Importantly, our results point to the conclusion that time is but one part of this process and that language proficiency and other individual-level learner differences can influence the overall trajectory of L2 collocation development.

Michał B. Paradowski and Elżbieta Pawlas. Communication breakdowns in intercultural communication: Causes, prevention, and remedial strategies [3]

Communication breakdowns have deservedly been attracting the interest of researchers, as they constitute important factors influencing the process of linguistic interaction and language acquisition. Not only do they affect the process of communication per se, but also have other, often serious, consequences. Particular interest should be accorded to the process of achieving—and failing to achieve—understanding when English is spoken as a vehicular language.

We will present the results of the first comprehensive analysis of the complete conversations subcomponent of the Vienna-Oxford International Corpus of English (VOICE), focusing on the i) possible causes of communication breakdowns, and ii) strategies employed by speakers in order to both prevent and overcome such failures. We categorise and show the distribution of the sources of 122 detected breakdowns as well as the compensatory strategies employed by interlocutors to successfully avert and solve communication problems.

The VOICE contains transcripts representing naturally-occurring face-to-face ELF interactions, whose participants come from different cultural and linguistic backgrounds. For the purpose of this study we selected all speech events tagged as ‘conversation’ (the analyses hence do not consider other speech event types, such as seminar discussions or interviews). After the selection, the reduced corpus comprised 36 speech events (158,071 words), corresponding to approx. 15 hours of spoken interactions. The speakers come from different, mostly European, countries, have different L1s and occupations. Their ages vary from 17 to more than 50. The relations between them are fairly symmetrical. The conversations belong to different thematic domains: 21 of them are tagged as ‘leisure’, 7 as ‘professional research/science’, 4 as ‘educational’, 3 as ‘professional organizational’ and 1 as ‘professional business’.
The entire material was first analysed in search of characteristic features and communication breakdowns. These were then analysed again in detail with regard to what caused the failures and how they were resolved, or at least how the speakers tried to resolve them. The list of identified causes covered unintelligible speech, simultaneous talk, overlap, pause, lack of topic shift signalling, lack of explicitness, wrong anaphoric or deictic reference reconstruction, faulty semantic reconstruction, code-switching, lack of shared cultural/world knowledge, misinterpretation of proper names, lack of shared lexical knowledge, wrong use of an existing word, wrong word order/tenses, and wrong/unfulfilled listener presupposition. Similar causes and similar strategies were then grouped together and tallied. Finally, the remaining data were again scrutinised in search of preventative strategies. These included enhancing explicitness, paraphrase, repetition, metadiscursive devices, completion of earlier utterance, dividing utterance into smaller parts, requesting assistance from other interlocutors, translating code-switches into English, and code-switch into language other than English.

The paper concludes with pedagogical recommendations.

References


Pascual Pérez-Paredes and Joyce Lim. Near-native learner essay writing: a developmental perspective using a large learner corpus [5]

In the last two decades, register research has made great contribution in both Applied linguistics as well as corpus methods in increasing our understanding of language use. However, there is a need to better understand how L2 learners approach different written registers at different stages of their learning trajectories. Given the fact that most learner language research has explored the written register (1) in relatively small corpora (Aarts & Granger, 1998; Paquot & Granger, 2012) and (2) has used schooling year or age as proxies for language competence (Meunier, 2015) rather than attested performance levels (Green, 2010). Our research, however, will be unique in examining the category of
“essays” in the largest written learner corpus available and in which the learner data is tagged for performance level (Common European Framework of Reference).

In this study, we examine the parts-of-speech (POS) sequences that shape the transition from upper immediate (B2) to near-native proficiency (C2) in a corpus of essays (n=21,780) from the Cambridge Learner Corpus. We looked at the frequency and distribution of 4-slot POS sequences in the essays contributed by English language learners between 1992 and 2011. Using a usage-based perspective, we found evidence of how language competence increases by means of adjusting the frequency and distribution of the most frequent abstract constructions analyzed in this study. C2 essay writing is characterized by (1) the use of a group of core sequences that had been already acquired at the B2 level and remain stable in essays during the transition towards near-native-like proficiency and (2) emerging sequences characterized by post-modification slots (e.g. DE JJ NN IN/ JJ NN IN DT) in the NP including prepositional phrases and non-finite clauses. A case study examined the sequence NN IN DT NN using grammar pattern approach (Hunston & Francis, 2000), and we discovered that within the same POS sequence, there is a transition towards more variety of complex constructions and a wider group of meanings that can be categorized under the same complex construction (N of N) in C2 writing. This research offers new evidence to our understanding of the development of different levels of constructional abstraction in learner essay writing in EFL contexts.

References


Adriana Picoral. Copula constructions in a multilingual learner corpus as evidence of L3 development [3]

Although the field of corpus linguistics has extensively addressed the effect of L1 on the target language (Golden, Jarvis & Tenfjord, 2017; Staples & Reppen, 2016), it has not addressed multilingual cross-linguistic influences to the same extent. The field of third language (L3) acquisition has for about a decade investigated the complex processes involved in the acquisition of an additional non-native language when two or more languages have been previously acquired (Alonso & Rothman, 2016; Cabrelli Amaro, 2017; Child, 2014; De Angelis, 2007). Previous research on L3 acquisition has shown that the source language for cross-linguistic influence can be either an L1, an L2, or both (De Angelis, 2007; Rothman, 2014; Slabakova & García Mayo, 2017; Bardel & Falk, 2007; Flynn et al., 2004). Some of these studies hypothesize that either typological similarities between languages previously acquired
and the target language (Rothman, 2010), or language status (L1 vs. L2) of previous acquired languages (Bardel & Falk, 2007) determine the source of cross-linguistic influence. However, more evidence is needed to confirm some of the assumptions present in these L3 acquisition models. The current study takes a corpus-based approach to the investigation of cross-linguistic influence on lexico-grammatical patterns in L3-Portuguese produced by three groups of English-Spanish bilinguals: 1) L1-English L2-Spanish, 2) L1-Spanish L2-English, and 3) L1-Spanish/English (i.e., heritage speakers of Spanish). The L3-Portuguese learner corpus used in this study comprises language produced by all three of these groups (for a total of 255 participants, 2,075 texts, and 536,168 tokens) divided into three L3 levels. In addition, six baseline corpora representing both previously acquired languages (i.e., Spanish and English) and the target language (i.e., Portuguese) are also included in this study. These are used to establish what language patterns are transferred from the L1s and L2s, and also at what level of proficiency L3 production approximates the language patterns found in the target language baseline.

The lexico-grammatical structure investigated is copula (e.g, *ser, estar, be*) due to its ubiquity in language production and language teaching materials. Multivariate quantitative analyses (i.e., word embeddings and logistic regression) of copula constructions take into account the context of each copula token. Results show that L3-Portuguese development takes place for all three groups of Spanish-English bilinguals. However, cross-linguistic influence from Spanish and English into L3-Portuguese production is not the same across these groups, varying in degree depending on the lexico-grammatical pattern. Spanish status (i.e., L1, L2, or heritage) seems to affect which copula constructions are transferred from Spanish (i.e., property-by-property transfer). In addition to form-frame-concept connections based on cognate words, this study offers evidence of connections based on selective attention (Ellis, 2006). Pedagogical implications for L3-Portuguese instruction include the need for highlighting certain lexico-grammatical structures, e.g. in Portuguese we say *meu amigo está ali na esquina* (i.e., my friend is around the corner) when the referent is mobile, and *meu apartamento é ali na esquina* (i.e., my apartment is around the corner) when the referent is stationary.

References


Tristan Purvis. Academic English Learners’ Corpus of Writing Students in Nigeria [5]

This paper presents a novel corpus comprising 701 essays written by 194 African students (primarily Nigerians) at a private university located in north-eastern Nigeria and discusses in broad terms patterns of errors and/or non-standard forms. The analysis also focuses on assessing the likely influence of the Nigerian cultural/linguistic context. The university from which data were drawn (American University of Nigeria) attracts a mixed population of students from various social backgrounds and geographical locations—predominantly different regions of Nigeria, but also from different countries in Africa (Cameroon, Ghana, Rwanda, South Africa, and Uganda) as well as the African diaspora abroad. Likewise, the faculty members represent a mixture of national origins and linguistic backgrounds (North and South America, Europe, Asia, and various countries in Africa). Though promoting a so-called American style of education, the language standard is not clearly set and tends to represent a mixture of American, British, and Nigerian standards. Throughout the course of four semesters, electronic submissions of essays from students in a first-year writing class were marked for a variety of errors and non-standard forms (as judged from academic writing standards) in the categories of capitalization, diction, punctuation, spelling, sentence structure, and word form, as well as problems with content, organization, and formatting. Several aspects of linguistic background are hypothesized as potential sources of some of these common errors and perceived non-standard forms: one’s mother tongue (if other than English) and associated “interlanguage,” Nigerian Pidgin English (a popular non-standard variety of English), and competing standard varieties to which students are exposed (Nigerian, British, American) as well as differing tendencies in various spoken versus written genres. This corpus promises to benefit both pedagogy and learning of academic writing, especially for West African student populations. Error patterns for many of the students in this corpus are certainly attributable to second language acquisition, while on the other hand, regardless of language background, a variety of errors are typical of college students requiring training in the rules of written English. Of particular note in this corpus, however, are patterns that are attested in Nigerian English, including professional settings by educated speakers. With this in mind, writing instruction should promote awareness of variation in standards of English (both modal and dialectal)—acknowledging the value of different varieties of English while training students in the internationally recognized formal, written standards.
Tugba Simsek and Cem Can. Pre-service language teachers’ perceptions on teachers’ duties in material development: can a corpus literacy course change their perceptions? [4]

Representing the authentic language and teaching the most frequent language items, corpus-informed materials’ use has been supported by many researchers (Meunier & Reppen, 2015; Römer, 2005). Since its importance has already been accepted in the field of language teaching, foreign language teachers, as the practitioners in the classroom, should be informed about corpus-informed materials and how to develop them. In line with this need, this qualitative study introduces corpus-informed materials and how to develop these materials to future language teachers and aims at seeking an answer to the question whether a corpus literacy course affects pre-service language teachers’ perspectives on the role of language teacher in material evaluation, adaptation, and development.

The course was a 16-week corpus literacy course taught as an elective course at an English Language Teaching (ELT) department. In the first part, the basics of corpus linguistics were introduced, and in the second part corpus-informed materials and their development were taught to pre-service language teachers. The participants were 20 pre-service language teachers, and their thoughts on teacher role in material development were gathered through a pre-course survey, semi-structured interviews, and focus group discussions. The collected data were analysed through computer content analysis. The pre-questionnaire results showed that before the course, 45% of the statements students used were rather general such as “a teacher should develop effective materials”. However, the results of semi-structured interviews and focus group discussions indicated that the answers of the participants became significantly diversified after the corpus literacy course. The participants stated that the teachers’ pedagogical characteristics (24.48%), material content selection (22.38%), and need analysis (16.08%) are highly important in material development, and teachers’ role as material producer (18.88%), material evaluator (11.89%), and material adapter (4.90%) has also been emphasised by the pre-service language teachers. The answers of the participants revealed that after the corpus literacy course, authenticity and frequency were seen as important components of material content selection; and the pre-service language teachers also stressed that while evaluating, adapting, and producing materials, teachers should make use of corpora to present real-life language to the learners.

References


Bruna Sommer-Farias, Adriana Picoral, Aleksey Novikov and Shelley Staples. Teachers’ Perceptions of Interactive DDL Using a Multilingual Learner Corpus [1]
Learner corpora have untapped potential for language teaching by providing instructors with large amounts of representative learner data (Granger, 2002) to create data-driven learning (DDL) materials (Johns, 1991; Gilquin & Granger, 2010; Flowerdew, 2015). Previous research on teachers’ perceptions of DDL have reported positive attitudes towards incorporating DDL in language classes despite concerns, such as time constraints and lack of confidence in the teachers’ knowledge of corpora and technical skills (Breyer, 2009; Chen, Flowerdew & Anthony, 2018; Farr, 2008). However, few DDL studies incorporate learner corpora and the existent studies largely focus on English as a Foreign or Second Language. These facts therefore signal the need for more studies involving teacher training and perceptions on how to design and implement corpus-based materials for less commonly taught languages (LCTLs) (Spina, 2017). This study addresses both gaps by analyzing data from workshops using a multilingual learner corpus for L2 Russian and Portuguese and interactive DDL (iDDL). The iDDL is a resource developed by one of the authors to facilitate incorporation of concordance lines into existing learning management systems (e.g. Canvas, D2L, Moodle).

In this study, we examined in-service teachers’ perceptions about the implementation of both iDDL ready-made and the materials they created. Data were gathered over the course of a year from instructors in a Portuguese foreign language program (N=7) and in a Russian and Slavic studies program (N=5) in Southwest U.S. during professional development workshops on how to use iDDL. All materials used in the workshop were created using a multilingual academic learner corpus of written and spoken assignments. The corpus is available with authenticated access and is composed of 700 texts produced by 255 students of Portuguese (169,382 words) and 350 texts produced by 100 students of Russian (42,682 words). We also analyzed field notes taken during the workshops, instructors’ survey responses after each workshop, and the iDDL materials created by the participants, along with their rationale for implementing these in their classrooms.

Preliminary findings show that instructors found the first workshop effective and envisioned using the ready-made activities more often than adapting them or creating their own. More detailed results will discuss whether and how they used the corpus to create their own materials, and what issues they faced regarding designing and implementing activities in their classrooms. Results will be discussed in light of needs for future training, technical and pedagogical challenges reported by the participants, and how these challenges may be overcome.

Shelley Staples. Building DDL materials with teachers of second language writing: Bridging the research/teaching gap [1]

While DDL has been established as an effective teaching method for second language writing (SLW) courses (Boulton & Cobb, 2017; Charles, 2011, 2014; Cotos, 2014; Shin et al., 2018), most studies testing the effectiveness of DDL materials use a teacher/researcher (or a team of experts) as the creator and deliverer of the materials. The question remains whether and how DDL can be implemented by SLW teachers more broadly, particularly when teachers are not extensively trained in using corpus methods. We also know that while many instructors express interest in teaching with corpora, they often lack the time and resources to develop materials on their own, even when they have been trained in using DDL in the classroom (Chen et al., 2019; Schmidt, 2020). This paper introduces an additional avenue for bridging the research/teaching gap, by discussing the process by which a writing program administrator worked with SLW instructors to develop and implement
materials. Data for the project includes 1) focus group discussions with five writing instructors about their perceptions of student needs and feedback on DDL materials developed for an undergraduate second language writing course; 2) observations of how those materials were then implemented in classrooms; 3) survey data from teachers and students about the materials; and 4) a final focus group with the writing instructors to provide additional qualitative feedback on the materials and their implementation. In addition, student drafts before and after the implementation of the materials (and a comparison with a control group) are examined as another measure of the effectiveness of the course materials. The presentation will discuss affordances and challenges of this approach to implementing DDL, and aspects of the materials and their implementation that led to more or less effective instruction. Preliminary results (on focus group discussions) indicate that there is a delicate balance between creating materials that appear innovative ("look like a corpus", in the language of the instructors) and materials that move beyond frequencies and concordance lines to examine discourse-based language use. In addition, integration with an existing curriculum provides important constraints as well as new challenges for instructors’ implementation. Implications for other DDL projects that attempt to integrate program-level changes with instructors whose background in corpus linguistics is limited will be discussed.


Spoken academic discourse has been at the center of academic language research for over 20 years (Benson, 1989; Flowerdew, 1994; Johns, 1981; Richards, 1983; Waggoner, 1984). The centrality of the spoken discourse in education is not surprising, as the history of education is inextricably linked to oral tradition (Jones, 2007; Ong, 1982). Oral instruction took a new form in the 21st century with the rise of easily accessible and free video platforms (i.e. Youtube): video tutorials. One extremely popular resource originating in YouTube is Khan Academy - a platform for online learning of school subjects via videos, interactive exercises and text-based material. Despite its popularity, there is a dearth of research on discipline-specific instruction in this video modality. Very few studies have focused on the language used in made-for-web resources in comparison to traditional, face-to-face instruction.

With the aim of filling this gap, this presentation will report on one section of results of a larger register analysis (Biber & Conrad, 2009) study. I compiled two corpora: transcripts of math instructional videos from Khan Academy and face-to-face math lectures from MIT OpenCourseWare. Both corpora have comparable size (1.4 mln words) and are centered on similar fields within mathematics, but different in terms of audience levels and number of speakers. The Khan Academy corpus comprises transcripts of bite-sized instructional videos, while the MIT corpus includes much longer transcripts from traditional lectures. I conducted an analysis of how one interactional spoken feature - direct hypothetical reported speech (DHRS) (i.e. ‘you could say, “hey, wait, I still have a parenthesis here, why don’t I do it first?”’) - is realized linguistically in both corpora, as well as how it functions based on the audience needs.

I will present short linguistic and functional analysis of direct hypothetical reported speech (Koester & Handford, 2018) in my corpora. The results show that hypothetical reported speech is used ten times more often in the short instructional videos than in the traditional lectures as represented by my corpora. The difference is discussed in terms of speaker preference, audience needs, level of
education, and interactivity dimension for both corpora. These results are also contextualized in terms of how DHRS is used by the online instructor to represent three types of knowledge - procedural, conceptual, and procedural flexibility knowledge (Rittle-Johnson, 2017) - to model mathematical thinking processes essential for successful performance. I will conclude with a discussion on the pedagogical uses of corpora based on transcripts of video tutorials for CLIL teachers.

József Szakos and Ulrike Glavitsch. “Genomes of corpus” and the problem of “genres”: Re-configuring elements of corpus linguistic solutions in teaching languages, taking readylingua as an example [4]

Lots of technology, software and teaching materials have changed over the past decades. There are constant elements, the “corpus genomes” surfacing in new variations during the experimentation of corpus-based instruction. Even if we devise speech corpora for learners, it just does not catch on. We still have not made corpus the main-stream of language teaching.

Our talk investigates whether the reason is in the genre limitations of (speech) corpus materials or in the “genomes”, the basic elements and methods of speech and written corpora, as they evolve in classroom, online teaching and self-instruction.

The “basic genome” is understood as the multimodal synchronization of text, vocabulary and speech (sound corpus) that has the plasticity and flexibility in interactive manipulation by the learners. The changeable and changing parts are the growing synchronic speech corpora and the way the learner can practice the chunks of speech at variable degrees of resolution, so as to increase his mastery of language. Is it that the methods fall out of the scope of interest of learners or the genres that such materials can cover?

After introducing SpeechIndexer and readylingua through current examples of news and literary corpora used by the subscribers of the website, we intend to show how speech materials can be modified to fit the learning habits of a new generation. But it seems, learners do not really want to use such materials. Similar procedures of synchronized text and sound are only available on TED talks online, and even if our solutions offer a much more detailed and learner-friendly data than TED talks, it does not seem to impress the majority of learners.

The semi-automatic preparation of authentic materials is realized with great care according to readylingua’s mission statement “Languages in Swiss top quality”. We pay attention to both good audio and text quality as well as a precise alignment between the two. Transcripts of news are manually created since radio stations licence and provide the audio recordings only. The alignment of audio and text is supported by an automatic pause detection such that linking between speech units and corresponding text can be easily made by hand. An automatic alignment component currently developed for various languages is a new milestone in preparing more corpus material in less time. In the future, students will be able to select from a large variety of news broadcasts, classical and modern audio books, poems, speeches, etc. to learn languages as comprehensively as possible.

The question is still open: is it the “corpus genomes/genes” (methods) or the “genres” of materials offered which fail to resonate with learners? Or is it the learner? We hope to receive a fruitful feedback to this through our presentation.
Szilvia Szita. Overwhelming, time-consuming, user-unfriendly... and now what?
Teacher training for the successful implementation of a corpus-informed methodology

This contribution will present the programme of a teacher training on the use of corpora and the implementation of an effective corpus-informed methodology in the classroom. This 30-hour long course has been offered at the University of Pécs every year since 2018. 48 (native and non-native) language teachers for Hungarian have attended it so far. The speaker has designed the training and she is one of the two trainers. A particularity of the training was that the majority of the participants were not only language teachers but also expert language learners. This allowed them to evaluate the usefulness of the tools from the point of view of the learners, as well. Based on the participants’ feedback, it will be argued that trainers should begin such courses by explicitly evoking techniques that the participants already use in their lessons before gradually introducing them to new ones. The benefits of the proposed innovations are accepted more easily when they are linked to existing practices and presented step by step. The presentation will show examples of this approach. The presentation will also explore the following questions: How can corpus tools and, even more importantly, the underlying reflection on language increase the effectiveness of teaching? What kind of corpus findings are relevant for teachers and learners? What are meaningful activities for the observation and practice of linguistic items? What could more general guidelines for language teaching look like based on what we know about language from corpus research? Why should all teachers be able to build corpora? A number of methods and activities were used to familiarise practitioners with corpora and corpus analysis tools as well as corpus-informed activities, from which the contribution will offer a selection. The most relevant comments of the participants will also be presented so that others who wish to set up similar trainings can benefit from it. As feedback on this training has been particularly positive, I would like to suggest it as a best practice model for a training for language teachers on how work with corpus analysis tools and create one’s own corpora for teaching purposes.

References
James Thomas. Learning language and content from a one-text corpus [4]

See Versatext workshop presentations.

Biagio Ursi. Corpus-based resources in conversation: Learning with the multimodal concordancer of the FLEURON database [4]

In the last fifty years, corpora have had increasingly important implications for foreign language teaching and learning (Boulton & Tyne, 2014). As for French, today, corpora of naturally occurring interactions (Ravazzolo et al., 2015) are useful resources for the development of conversational competence, which deals with contextually situated forms and sequentially interconnected utterances of talk-in-interaction (Kecskes et al., 2018).

In this paper, I explore the FLEURON database (https://fleuron.atilf.fr/), a collection of audio and video recordings documenting interactions representative of students’ life in France. These recordings are subtitled; their transcripts are searchable through specific tools. Within FLEURON, a major focus is put on its multimedia concordancer. This tool allows users to search for words, to visualize targeted occurrences in their context. Students may watch the corresponding video-extract, aligned with the verbal production in which the targeted occurrence appears. In doing so, they familiarize themselves with the uses of spoken French, acquire sociolinguistic awareness and develop learning autonomy (André & Ciekanski, 2018).

In FLE classes, teachers support students in the mobilization of digital platforms for learning purposes. They introduce the elements of the interface and show how to handle specific tools. Thanks to the concordancer, learners observe forms in context and scrutinize their distributional regularities, according to the principles of data-driven learning (Johns, 1991). Thus, a process of linguistic discovery that leads to the identification of patterns and rules of use takes place. The mobilization of this digital resource is subject to both mediation by the teacher and analytical examination by students. Learners formulate their hypotheses about the linguistic system, like researchers do.

A sequential and multimodal analysis (Mondada, 2018) of teacher-student interactions is proposed to highlight emerging practices in the learning process with the digital tools of the FLEURON database. The focus will be on the identification of specific items in the concordancer, as well as on the mobilization of verbal and gestural resources that allow learners to point out recurrent and also unexpected forms of spoken French, in their natural environment. The generalizations and problematic issues identified by learners arise through specific address practices and one-to-one participation configurations (Goodwin & Goodwin, 2004). According to qualitative analyses, verbal hints as well as pointing gestures, gaze directions and body postures turn out to be primordial resources for assessing the effectiveness of data-driven language methods.
References


Benet Vincent, Hilary Nesi and Daniel Quinn. Exploring teacher engagement with an open DDL resource [1]

BAWE Quicklinks is a DDL initiative which aims to help teachers give their students more efficient feedback on writing issues. This is done by providing access to custom-made corpus outputs from the open-access British Academic Written English (BAWE) corpus in Sketch Engine. The outputs are based on problems recurrently found in the writing of university students which affect communicative force or intelligibility. These might be unidiomatic collocation selections, for example, or inappropriate lexico-grammatical choices. For each issue, we use the corpus query language in Sketch Engine to create outputs from BAWE which are designed to lead students towards more appropriate alternative wordings. Outputs include concordances, collocation lists, word sketches, or the ‘Frequency: KWIC word forms’ function, which lists realisations of a pattern retrieved by a query in order of frequency. The idea is that on spotting an issue, a teacher can then select a relevant link and insert it into the student’s work. A database of around 100 ‘quicklinks’ – hyperlinks to the outputs – has now been created as an open educational resource.

One of the issues in much DDL research is that it involves practitioners who are already committed to the DDL approach. If we want to spread the word about the benefits of DDL it is important to investigate the responses of teachers who have not already invested in the approach, and have little time to do so. This study involves a three stage process. First, we work with teachers to select the hyperlinks from our database that they consider to be most useful. The next stage involves making these links accessible to the teachers while they are marking assignments online by creating sets of the links they have selected. In the final stage we collect feedback from these teachers on the extent
to which they have used the links, and their evaluation of their success. Our presentation will report on this results of the study, reflecting on the reactions of teachers to our initiative.

Olga Vinogradova, Anna Viklova, Darya Overnikova, Darya Matyash and Mikhail Paporotsky. “Genre is a useful concept only when used evaluatively not descriptively”: Observations over differences between two genres of essays in learner corpus [5]

Following the research path in Alexopoulou et al. (2017), we set up the task to evaluate the effects of the task type/genre on the accuracy of learner writing. The research material included examination essays answering the questions of two types – Task 1: Describe the graphical material presented in the task – 2,414 texts; and Task 2: Provide arguments to prove one of the two positions outlined in the task – 2,345 texts. The essays were written by learners of English with Russian L1 in their 2nd-year English examination. The learner corpus with those essays has two layers of annotation – automated POS tagging and expert manual annotation of errors. The characteristics of the two tasks follow the parameters suggested by Skehan (2003) and Robinson and Gilabert, (2007), which are given in Table 1.

Table 1. Overview of task characteristics

<table>
<thead>
<tr>
<th></th>
<th>Graph Description</th>
<th>Argumentative Essay</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LAC model</strong></td>
<td>Medium vocabulary load</td>
<td>High vocabulary load</td>
</tr>
<tr>
<td></td>
<td>Medium structure</td>
<td>Low structure</td>
</tr>
<tr>
<td></td>
<td>High clarity</td>
<td>Medium clarity</td>
</tr>
<tr>
<td><strong>Cognition hypothesis</strong></td>
<td>Medium different elements</td>
<td>Many different elements</td>
</tr>
<tr>
<td></td>
<td>There and then/In the past and in future</td>
<td>Here and now</td>
</tr>
<tr>
<td></td>
<td>Reasoning obligatory</td>
<td>Reasoning obligatory</td>
</tr>
<tr>
<td></td>
<td>Perspective taking likely</td>
<td>Perspective taking obligatory</td>
</tr>
</tbody>
</table>

According to Skehan’s criteria, in Task 1 much of the lexis needed is provided in the prompt, so learners have to demonstrate a medium vocabulary load, while in Task 2 much more is to come from the learner than from the prompt. The limitations of the time allotted for Task 1 (20 minutes) and of the length of the answer (about 150 words) account for the fact that the structure of the essay rigidly requires comparisons and descriptions of changes in this task, while in argumentative essay writers can choose the strategy with more freedom. Parameters in Robinson’s Cognition Hypothesis for Task 1 show a medium number of elements to be discussed either in the past periods of the research, or the changes and trends from the past to the present or future, while in argumentative essays authors have to talk about many elements related to their argumentation, and mainly the discussion is restricted to here.
and now, occasionally giving an example from the past. Both tasks have to involve reasoning, while “perspective taking” is a must in giving reasons for thinking differently from the author’s standpoint, while it may be expected less in describing the graphical material in Task 1 from the writer’s perspective. The opinion that these characterizations capture some core properties of the tasks in question which are bound to produce impact on language use has been given in Alexopoulou et al. (2017) and serves as the basis for presenting these parameters here.

To follow the evaluative direction in the title, we bring in three sets of statistical data accompanied by the graphical representation. Initially we looked at the difference in numbers of errors annotated manually, which we extracted with Python codes. The average length of the essays in these two genres is 1.5 times different (181 vs 282 words, 156%). The initial hypothesis that the numbers of different errors would reflect this 1.5 times difference was soon found unrealistic, as many specific features of task types accounted for more errors attested for one or the other of the two types. To make the analysis more meaningful, and the presentation of the data more concise, we clusterised about a hundred error labels used in the corpus to identify learner errors into 8 classes, which you can see in Table 2. The classes were grouped on the grounds of errors occurring in similar situations in learner writing. The data reflecting error frequencies may seem controversial, so in the talk we try to account for the differences within and between the classes. In order to make the data comparable, we calculated the normalized indices for 1000 words, and compared them rather than the raw total occurrences.

Table 2. Distribution of error tags in student essays of two tasks.

<table>
<thead>
<tr>
<th></th>
<th>Task1</th>
<th>Task2</th>
<th>Task1</th>
<th>Task2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determiners and quantifiers</td>
<td>6,580</td>
<td>7,081</td>
<td>15.05</td>
<td>10.70</td>
</tr>
<tr>
<td>Prepositions - Independent</td>
<td>2,416</td>
<td>2,112</td>
<td>5.53</td>
<td>3.19</td>
</tr>
<tr>
<td>Prepositions - lexical</td>
<td>340</td>
<td>641</td>
<td>0.78</td>
<td>0.97</td>
</tr>
<tr>
<td>Verbs - choice of tense</td>
<td>3,575</td>
<td>1,504</td>
<td>8.18</td>
<td>2.27</td>
</tr>
<tr>
<td>Verb pattern</td>
<td>97</td>
<td>710</td>
<td>0.22</td>
<td>1.07</td>
</tr>
<tr>
<td>Verbs - all other errors</td>
<td>953</td>
<td>1,615</td>
<td>2.18</td>
<td>2.44</td>
</tr>
<tr>
<td>Nouns - grammar errors</td>
<td>1,881</td>
<td>2,323</td>
<td>4.30</td>
<td>3.51</td>
</tr>
<tr>
<td>COMPARISONS</td>
<td>1,363</td>
<td>149</td>
<td>3.12</td>
<td>0.23</td>
</tr>
</tbody>
</table>
The lines without highlighting show “standard” distribution – i.e. the normalized number of errors in Task 2 almost equal to that of Task 1 (in the range from 80% to 120%). Blue lines show lower percentage or errors in Task 2 than the “standard” (below 80%), and the darker blue, the lower the level; the green lines, on the contrary, show the percentage of errors in Task 2 exceeding the expected “standard” above 120%.

Some deviations from the “standard” can be explained by the specific features of the tasks presented in Table 1. Comparative constructions are explicitly required in Task 1, for example. However, for some other features the only possible reason for the discrepancy between Task 1 and Task 2 can be found in the phenomenon of interference with the native language – a factor mentioned in many papers from SLA research. Such is the case with frequent appearance of present tense forms (Present Simple or Present Perfect) in the context with the definition of time in past in the written production of Russian L1 students, even those with a very high level of proficiency. The plausible explanation is that such uses are commonly found in Russian academic discourse reflecting past events or phenomena. Compared to English, a very different convention for the choice of tense can be observed in Russian, as illustrated by the following statistics: about 10% of 10 most frequent verbs used in academic texts in the National Corpus of the Russian language are used in the present tense form in the contexts with clear references to the past events. Both past and present forms of verbs can be used for describing past actions, sometimes even in one and the same sentence. To prove it, we carried out the experiment across a database of academic papers in sociology written in Russian, from which we extracted descriptions of data from past years presented in the papers with some graphical material. Both manual and automated extraction of sentences talking about past events or phenomena was performed, and only those sentences were included in the research dataset that include the time in the past stated or implied from the nearest context. As a result, out of 221 verbs only 130 were forms of the Russian past tense, while 91 were forms of the Russian present tense, participles with the omitted present tense auxiliary verb among them. So, it is no wonder that learners even at an advanced level of English proficiency demonstrate a huge deviation from consistent use of Past tense when they look at past events while describing some academic research. Correspondingly, in Task 1 essays, the level of erroneous uses of present tenses instead of past tenses is much higher than that in
Task 2 essays, where authors talk predominantly about here and now. At the same time all other types of errors in using verbs are within standard distribution between the two task types.

Similar cases are to be considered for the errors in the use of prepositions, which can be compared with research over the same area in EFCAMDAT in Alexopoulou et al. (2017), and in the area much less often covered in research papers – misuses of punctuation by learners of English. The opposite situation was attested in the errors in Relative clauses, with the much higher level of them in argumentative essays.

The final point in this research is to identify the correlation between errors specific for either task type and the linguistic properties of learner language – overuse/underuse of certain patterns, as was suggested by Meurers and Dickinson (2017).

After considering all classes of errors in their distribution of Task1/Task2 levels, we can say that in preparation for the exam, English instructors can benefit a lot from the task-based insights of the presented corpus analysis, and knowledge of specific errors prevalent in student essays of either type may lead to better instruction adjusted to the needs of each task.

References


Karin Whiteside. Writing development across an undergraduate degree: A contrastive corpus analysis of student writers’ first and third year Essay texts from the disciplines of History and Politics & International Relations (PIR) [5]

In recent years increasing research attention has been devoted to undergraduate student academic writing. A growing number of studies involve linguistic analysis of the texts produced at this level, the BAWE Corpus project (Nesi & Gardner, 2012) being a notable example. Some of these studies have looked at the development of students’ writing as they progress through their degree programmes. Woodward-Kron (2009) undertakes a close Systemic Functional Linguistics-informed analysis of texts written by six undergraduate writers in Education, charting how developments lexicogrammatically link to growth in knowledge of discipline, and Binchy (2013) analyses Philosophy module coursework assignments from a single cohort of students at six points across their undergraduate degree programme, finding no discernible pattern of development in the use and distribution of linguistic features. In Gardner, Nesi and Biber’s (2019) multidimensional analysis of the holdings of the BAWE
corpus, level of study is one of the variables, alongside academic disciplines, disciplinary groups (e.g. Arts & Humanities), and genre families.

It is intended that the following study contributes to this body of research by contrastively analysing the writing of individual students across two disciplines and two stages of their undergraduate study (first and final year). A corpus has been created consisting of first-year and third-year texts produced by nine undergraduate students at one UK HE institution, four undertaking degrees in History and five undertaking degrees in Politics and International Relations. The third-year texts in this corpus are from a larger corpus of successful 3rd-year writing in the two target disciplines across two UK HE institutions, originally used to contrastively analyse the semantic patterning of grammatical keywords (following Groom, 2007) across discipline and institution (Whiteside, 2016; Whiteside & Wharton, 2019). Details of the current corpus are as follows:

<table>
<thead>
<tr>
<th></th>
<th>History sub-corpus</th>
<th>Politics &amp; International Relations sub-corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(four student writers)</td>
<td>(five student writers)</td>
</tr>
<tr>
<td>1st year undergraduate</td>
<td>62,101 words (26 texts)</td>
<td>80,695 words (38 texts)</td>
</tr>
<tr>
<td>3rd year undergraduate</td>
<td>57,278 words (19 texts)</td>
<td>76,693 words (28 texts)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>119,379 words</td>
<td>157,388 words</td>
</tr>
</tbody>
</table>

The paper will report on progress made in two strands of planned analysis of the corpus:

1. Building on previous work (Whiteside, 2016; Whiteside & Wharton, 2019), analysis of the proportional distribution of ‘semantic sequences’ (Groom, 2007; Hunston, 2008) around grammatical keywords (e.g. as, that, and this) across discipline and level.

2. Analysis of Appraisal resources (Martin & White, 2005) used within the introduction and conclusion paragraphs of each text across discipline and level.

It will also draw some preliminary conclusions arising from the results regarding implications for the teaching of academic writing at undergraduate level.

References


Yi-ju Ariel Wu and Chi-Ting Robert Tsai. Exploring the Effects of Data-Driven Learning in Learning Legal Vocabulary of International Law [1]

As English becomes the important lingua franca for global transaction and government negotiations, legal English becomes one important branch in ESP (Northcott, 2013). Specifically, legal vocabulary has received the most scholarly and practitioners’ attention because of its close tie with legal reading, known for its complexity and technicality, and high level of difficulty. Because of the promising results shown in many empirical studies of DDL in vocabulary learning, many legal scholars (e.g., Breeze, 2017; Marín, 2014) and DDL researchers (e.g., Boulton, 2011) have called for the use of DDL in legal vocabulary learning. Data- Nevertheless, the understanding of using DDL in legal vocabulary is still insufficient, as only few empirical study has explicitly examined the effect (Yunus & Awab, 2014). Moreover, although studies have shown that DDL use bring learning gains to vocabulary learning, research results are still inconclusive, as there are contradictory results found in various aspects, including whether DDL bring learning gains to both production and comprehension of vocabulary, whether the learning effects of DDL in vocabulary can last until a few weeks after the treatment, and whether learners can infer the vocabulary meaning through DDL without assistance from dictionary. As a team-teaching project of legal professors and one ESP instructor, the suggested implementation by ESP legal researchers (e.g., Hafner & Candlin, 2007), this study aims at bridging gaps identified above by examining how 80 International Law students from social science departments learn 30 legal vocabularies in two different learning conditions. The first class, DIC group (traditional, deductive, English definition), will be provided English definition of the 30 target legal vocabulary which also displays the collocation/colligation pattern of it (the preposition of the target item such as acquiesce “to”) and conventional matching practices to strength the form-meaning connection. The second class, CONC group (paper-based DDL materials) is required to analyze three concordance lines adapted from three specialized legal corpora with teachers’ editing. Participants’ evaluation toward different learning conditions, their pattern inducing and meaning inferring process, and how the process and evaluation are related to their learning outcome were investigated. Data were obtained through (1) learners’ performances: pretest, post-test and delayed post-test and (2) learners’ attitude toward the instruction: 3 questionnaires from 80 students and interviews from 15 students out of two classes respectively. A mixed-method approach including quantitative statistics and qualitative analysis based on grounded theory is used to evaluate learners’ performance, attitude and use of concordance lines. The results showed that CONC group outperformed DIC group in both posttest and delayed posttest, and the difference in performance between two groups were even sharper in the delayed posttest.
Also, learners’ evaluation toward the instruction is higher in the CONC group. It is concluded that the concordance lines provided more context for learners to induce the meaning of legal vocabulary as ordinary vocabulary carrying specialized meanings in legal context (e.g., Nation, 2008; Camiciottoli, 2007). Also, the concordance lines providing more social contexts regarding how legal vocabulary served as a medium for building identity and membership for International law students.
Cathryn Bennett. Preliminary Findings of Corpus Literacy by In-service EFL Practitioners in Ireland [6]

It has long been argued by researchers in the area of Applied linguistics (Johns, 1991; Mukherjee, 2006; Boulton, 2009; O’Keeffe, 2012; Zareva, 2017) that the use of authentic language data in the language learning classroom can have a positive impact on learning outcomes. However, the use of authentic data, in the form of language corpora in the classroom, has not yet made it into mainstream language teaching practice nearly thirty years after its birth. A lack of resources (Naismith, 2017) and training (Farr, 2008) have been suggested for this hesitancy on the part of English as a Foreign Language (EFL) practitioners. As international students coming to study in Ireland reach a new height (Power, 2019), the issue of providing corpus training to teachers is critical.

This poster seeks to understand this phenomenon in reporting preliminary results from the first phase of my PhD in which a quantitative questionnaire was designed and disseminated to EFL practitioners in English Language Teaching Organizations in Dublin, Ireland. The aim of the questionnaire was two-fold: to ascertain EFL practitioners’ current awareness of corpora as a learning tool and whether they use it in their classrooms. By extension, it identifies perceived existing barriers towards uptake by teachers.

Phase two of my PhD will involve corpus literacy training of EFL practitioners in a newly designed prototype teacher education course.

References


This poster presentation introduces a recently created corpus of longitudinal learner data, called LEONIDE. The corpus contains 2,543 texts from 163 pupils, who participated in the project “One school, many languages” conducted in eight schools in the officially multilingual Italian province of South Tyrol / Alto Adige (Zanasi & Stopfner, 2018). The aim of the project was to document the development of the pupils’ plurilingual linguistic and communicative skills by collecting oral and written language samples in Italian, German and English, in order to obtain a global view of their individual linguistic repertoire.

LEONIDE is a collection of all the written texts of the pupils during the course of the project, the overall size of the corpus amounts to ca. 240,000 tokens. The texts were collected over the span of 3 consecutive years (2015-2018) in public middle schools (i.e. lower secondary school, grade 6 to grade 8). The pupils were 11 years old at the beginning of the data collection and 13 years old at the end. In each grade, two written texts were collected that differ with respect to genre: the first text was elicited using a picture story re-telling task; the second text is an opinion text on different aspects related to the pupils’ life and public discourse. For each genre and each grade, the corpus provides texts in the three languages German, Italian and English. In order to reflect the school system of the Province of South Tyrol / Alto Adige, about half of the texts was collected in four schools in which German is the main language of teaching and Italian is taught as L2. The other half of the texts was collected in four schools in which Italian is the main language of teaching and German is taught as L2. In all schools, English is taught as L3 (i.e. as a foreign language at school). Subdivided by language, the corpus contains 850 Italian, 849 German and 844 English texts. Furthermore, a series of relevant person-related data was collected for each learner, providing information about e.g. age, gender, first language(s), language assessment scores for each of the three languages.

As the corpus documents the development of plurilingual competences of individual learners, it allows for contrastive longitudinal research on the development of young learners’ writing skills in different languages, considering also person-related metadata. Moreover, the corpus is a valuable resource for language teachers in order to create and improve their teaching material and language courses as the large amount of authentic and longitudinal data reflects the sequencing of language skills over three consecutive years in three languages. The corpus will be available for corpus queries via an ANNIS search interface and as download for academic purposes (ACA-BY-NC-NORED 1.0) on the Eurac Research Clarin Centre by the end of 2020.

References


The number of learners of Czech as a second or foreign language has been growing quickly in recent years. This increase in interest gives us an opportunity to develop a new type of resource for both teaching and learning. One of the key materials is a learner’s dictionary which can help students to improve their vocabulary knowledge. The aim of this poster is to present the methodology and design used for the set of sample entries (500 headwords - single units and multiple word units) of a new web-based dictionary for learners of Czech. The development of such a reference tool is based on the hypothesis that learners should be exposed to what is frequent in the language and on the assumption that implementing corpus research can optimize the acquisition of the Czech lexicon. The frequency analysis is used for detecting the most frequent word forms for each headword. Czech, as an inflectional language, is characterized by a large number of word forms, especially for nouns and verbs, which do not have equal representation in language production. Each form is motivated by its grammatical pattern and collocation. These patterns and collocations are culled from a representative corpus of Czech, where the three most frequent word forms for each headword are identified and used later in entry description and example sentences. In addition to that, each entry contains common learners’ errors or misuse and hyperlinks to external linguistic online tools. The design of the dictionary reflects all information gathered from corpora from a pedagogical perspective, e.g. entries can be presented in alphabetical order but also according to the grammatical pattern or type of mistake. All of these stated features would be hard to cover in a paper dictionary but can be integrated into a web-based reference tool.

Elżbieta Kaczmarska and Adrian Jan Zasina. Infrastructure of the Polish Learner Corpus [6]

In recent years, learner corpora have become increasingly popular as a source for analysing L2 learners’ language (Gilquin, Granger, & Paquot, 2007: 322–323). However, in comparison with national corpora, which are available for many languages, we find relatively few learner corpora for other languages than English (Štindlová, 2013: 62–65). In particular, a learner corpus is still missing that could drive analyses of the language of non-native speakers of Polish (Zasina, 2019). Therefore, the goal of our poster is to present the first attempt to compile such a learner corpus and report on the ongoing project.

The primary goal of the project is to collect learners’ writings in Polish as a foreign language at various levels of language proficiency. The collected material will be a basis for analysing the learners’ language, identifying the most common language errors, creating classroom materials, and improving modern teaching methods.

In the first step, we are going to collect all available electronic texts, so as to gather a sizeable amount of starting material in the shortest possible time. In the second step of our project, we intend to focus more on hand-written texts and the rules for transcribing such texts in a computer-readable format, and on balancing the entire corpus in terms of first language and language level (CEFR).
In the initial phase, the corpus will be prepared in the TeiTok environment (Janssen, 2016), where it can be easily modified during its creation. TeiTok will also be used for text transcription and managing all collected learners’ writings.

References


The purpose of this research is to find out whether the duration difference between stressed and unstressed vowels can serve as a cue to determine English proficiency of Korean learners.

Linguistically, content words, such as nouns, verbs, adjectives, and adverbs contain messages speakers want to convey while function words which include articles, prepositions, and conjunctions create grammatical or structural relationships between words, phrases, or clauses (Fear et al., 1995; Selkirk, 1996). In English, a stressed-time language, the content words have stress, by which they are produced longer in duration, higher in pitch, and louder in amplitude. For this reason, speakers pronounce the content words in English with great care than the function words. However, in the case of Korean, individual word stress is not as salient as that of English, since Korean belongs to syllable-timed languages.

In the previous study, MacAllister et al. (2002) showed that difficulty in learning L2 contrastive category is relevant L1’s phonetic features on which the L2 category depends. Based on this past study, this research was aimed at testing a hypothesis that the higher English proficiency the Korean learners have, the greater the duration difference would be seen between the vowels in nouns and the vowels in articles. For example, when speakers produce a noun phrase such as "the bus", the stressed vowel in the noun of "bus" should be longer than the unstressed vowel in the article of "the". In this study, this duration difference between these vowels was expected to increase as their English proficiency gets higher.
To verify it, the researchers analyzed K-SEC (Korean-Spoken English Corpus) where 112 Korean elementary school students spoke English sentences. In the data, 853 sentences articulated by these speakers were targeted to measure the duration of each vowel in the monosyllabic noun phrases.

From the study's findings, it turned out that the native English speakers produced the vowels in the nouns 2.956 times longer than those in the articles. As for the Korean speakers, those who were categorized in the advanced proficiency group were found producing the vowels in the nouns 2.453 times longer than the vowels in the articles. Those who were categorized in the intermediate high and intermediate low group also produced the vowels in the nouns respectively 1.261 and 1.035 times longer than the vowels in the articles. However, the speakers in the novice group showed the different aspect - they produced the vowels in the nouns as much as about four-fifths of the vowels in the articles, which means that the vowels in the articles were produced much longer than the vowels in the nouns among the speakers in the novice group.

As a result, the research supports the previous hypothesis that the more proficient the learners are in English, the longer the vowels in content words are produced than the vowels in function words. It was also proven statistically significant (***p<.001) and therefore it is concluded that in speech production, the duration difference between stressed and unstressed vowels can be a cue in determining Korean learners' English proficiency.

References


Juliane Schopf. Authentic spoken data and applied conversation analysis as a resource for the improvement of communicative competence in healthcare [6]

Doctors hold around 200,000 conversations with their patients over the span of their career (cf. Silverman et al., 2013: 7). These conversations serve as primary diagnostic tools (cf. Koerfer & Albus, 2018: 13) and successful communication is known to improve patient compliance (cf. Lutfey, 2004; Pollok, 2005). Despite the acknowledged importance of communication in medical context, not enough attention is being given to the development of communicative competence in medical education and continuing medical training. A closer look into the existing materials, which touch upon the relevance of doctor-patient communication, suggests that they mostly consist of general recommendations and checklists which have neither empirical basis, nor information for the contextual embedding of the reported conversational situations.
The planned poster presentation – based on data from vaccination consultations – aims at showing the importance of using authentic conversations in the medical education and continuing training, as well as suggesting the ways of implementing applied conversation analysis methods for the empirically based communication training in the healthcare sector. For instance, the findings from the conversation analysis in the medical context and training materials based on them (cf. Coussios et al., 2019) can help introduce students to “professional meta-competence” (Koerfer et al., 2008: 41). This, in turn, will allow future doctors adjust their communicative skills in accordance with the situational needs of their patients. Additionally, corpus-based trainings may offer miscellaneous help for managing multilingual communicative situations, as demonstrated by Bührig & Meyer (2009a) and Bührig et al. (2012) in the example of corpus “Dolmetschen im Krankenhaus” (DiK – Interpreting in Hospitals; Bührig & Meyer, 2009b).

The usage of authentic audio data is particularly important for developing the teaching materials aimed at medical students who are receiving training in their second or foreign language. Frequent exposure to the authentic doctor-patient conversations can help language learners not only learn the communicative norms and features certain types of conversation, but also acquire knowledge of phonotactic and prosodic patterns of the target language. Spoken materials also enable non-native speakers better perceive and produce phonological features and constructions which are characteristic of colloquial speech. The existing language courses for non-native medical students mostly focus on learning the necessary terminology and developing writing skills, e.g. writing medical reports, while the doctors’ routine to the great extent consists of conversing with patients, their relatives and colleagues. To conclude, providing the access to the corpora of authentic doctor-patient conversations (corpora within the CLARIN-/CLARIAH-Network and Plattform Gesprochenes Deutsch [engl.: Platform Spoken German]) can complement the existing materials in the field of medical training and improve communicative skills of the medical students.

References


Evaluating the appropriateness of a written text for classroom use is an activity that language teachers are frequently engaged in. Evaluation processes, however, are usually conducted subjectively, thus leading to potential discrepancies among different evaluators.

This poster presents a new resource for assessing text difficulty automatically, in the context of Italian as a second or foreign language learning. The resource is called MALT-IT2, i.e. Measuring automatically the level of texts for second or foreign language learners of Italian, and it combines corpus construction with machine learning techniques. It is accessible at the following link: https://lol.unistrapg.it/malt/.

The corpus it is based on contains texts gathered from the language certification exams conducted by a specialised language examination centre in Italy. The texts included in the corpus cover proficiency levels from B1 to C2. Each level contains a range of 119 to 249 texts, 15709 to 7494 tokens, and 104679 to 45695 tokens, reaching a total of 692 texts, 29983 types and 336022 tokens overall. The corpus was then used as a gold standard to train text classification systems, with an integration of a wide set of linguistic features (e.g. Dell’Orletta et al., 2011; Graesser, et al. 2004, Grego Bolli et al., 2017; Xiaobin and Meurers, 2016), in order to create a system that would be able to automatically assign a text to a specific CEFR level.

Although a number of research projects have dealt with the automatic classification of texts for second language learning purposes (e.g. Xia et al., 2011), to the best of our knowledge, MALT-IT2 is the only resource available to date that is specifically aimed at Italian L2 teaching and assessment (Forti et al., 2019).

The poster will describe the different components forming the architecture of the system, as well as the influence that the different linguistic features considered have on differentiating between the various CEFR levels. It will also provide some examples of the information that the system is able to provide teachers needing to assess the level of a text.

References


