Corpus research for language teaching and learning

[A] Introduction

A corpus can be broadly defined as a collection of naturally occurring spoken or written data in electronic format, selected according to external criteria to represent a language, a language variety or a specific domain of language use (e.g. conversation or academic writing). Analyses of corpora using the methods and tools of corpus linguistics have contributed significantly to much more detailed and accurate descriptions of a wide variety of languages. They have also revolutionized language theory and description by placing special emphasis on (a) the frequency in which words and patterns occur in a language, (b) the ways in which words combine in collocations and other phraseological units, (c) the close connections between grammar and lexis, and (d) the ways in which situational factors (e.g. written versus oral communication, news writing versus fiction writing) act as explanatory variables for linguistic variability (Hunston, 2002).

The main objective of this chapter is to review how corpus-based descriptions of language use can have an effect on the teaching syllabus and the design of pedagogical materials by informing decisions about what to teach and when to teach it. Corpora can also be exploited directly by teachers and learners but so far such uses have largely been confined to advanced proficiency levels and higher education. The reader is therefore referred to Cobb and Boulton (2015) for an overview of direct uses of corpora in the L2 classroom.

[A] Current issues

[B] The Role of Frequency

Corpus-derived frequencies are often used to inform two main components of reference and instructional materials design: selection and sequencing. As put by Cobb and

Boulton (2015, p. 479), “the basic idea is that frequency of form and meaning is the most reliable predictor of what can be most usefully taught at different points in the learning process.” One of the areas where frequency, among other corpus-based information, first brought about an unprecedented revolution is pedagogical lexicography. It would be inconceivable for lexicographers today to produce a learner’s dictionary in a well-resourced language without making use of corpus-derived frequencies at the very least as a guide to the selection of the lexical entries (i.e. which words to define and illustrate), the ordering of word meanings, and the identification of a defining vocabulary, i.e. a restricted list of high-frequency words which can be used to provide simple definitions of words in the dictionary. When corpus-derived frequencies are used in textbook material development, they typically serve to sequence the introduction of new vocabulary words. For example, the authors of the Vocabulary in Use series published by Cambridge University Press used word lists derived from the Cambridge International Corpus to inform the selection of the most useful vocabulary that students need at each level.

With the increased availability of large reference and specialized corpora, work on frequency lists has been extremely productive over the last years[i] and several proposals have sought to replace West’s (1953) General Service List (GSL), i.e. an influential but dated list of 2,000 lexical items deemed to be especially useful for the purposes of language pedagogy that was compiled before the advent of computerized corpora. One such proposal is the New General Service List (Brezina & Gablasova, 2015), which, unlike the GSL, makes use of lemmas instead of word families (see Paquot (2010) or Gardner & Davies (2014) for a synthesis of criticisms levelled at the use of word families to determine word frequencies). Brezina and Gablasova (2015)

selected four general corpora to represent a variety of corpus sizes and approaches to sampling and representativeness and compared lists of the top 3,000 lemmas for each corpus based on the average reduced frequency (ARF), which is a measure that serves to weight the absolute frequency of a word in terms of its distribution in the corpus[ii].The results showed that there exists a stable core vocabulary of 2,122 lemmas (70.7%) among the four corpora.

Wordlists based on specialized corpora have also flourished over the last years, especially those targeting vocabulary needs in higher education. The Academic Keyword List (AKL, Paquot, 2010) contains 930 potential academic lemmas, i.e. words that are reasonably frequent in a wide range of academic texts but relatively uncommon in other kinds of texts as identified by the three corpus-based techniques of keyword analysis, range and dispersion. The AKL differs widely from Coxhead's (2000) Academic Word List as it is a list of lemmas, not word families, and was developed for productive purposes. As such, it includes high frequency words (e.g. aim, argue, because, compare, explain, namely, result) which have been shown to play an essential structuring role in academic prose but remain a source of difficulty for EFL learners. Gardner and Davies (2014) adopted a similar approach to Paquot (2010) to produce the New Academic Vocabulary List, a list of 3000 lemmas derived from the academic subcorpus of the Contemporary Corpus of American English (COCA). Specialized wordlists for disciplines such as medicine, agriculture, engineering and business have also been compiled on the basis of domain-specific corpora over the last ten years (see Lessard-Clouston, 2012/2013 for a selected overview).

The most recent wordlists can certainly be described as superior to previous efforts if only because they are based on ever larger corpora and compiled with the help

of more sophisticated corpus-based techniques and measures. However, the large majority still focus on single words despite convergent findings in corpus linguistics, psycholinguistics and cognitive linguistics that word combinations, be they framed in terms of phraseological units, formulaic sequences or constructions, play crucial roles in language acquisition, processing, fluency and idiomaticity (e.g. Sinclair, 1991, Ellis & Wulff, 2015). Three exceptions are:

1. The Academic Formulas List, i.e. a list of 200 core, 200 written-specific and 200 spoken-specific formulaic sequences which occur significantly more often in academic than in non-academic discourse (Simpson-Vlach & Ellis, 2010);
2. The PHRASal Expressions List, i.e. a list of 505 most frequent non-transparent multiword expressions such as of course, at least and I mean in the British National Corpus accompanied with frequency information for spoken general English, written general English and written academic English (Martinez & Schmitt, 2012);
3. The Academic Collocation List (ACL), i.e. a list of 2,468 most frequent and cross-disciplinary lexical collocations compiled from the written component of the Pearson International Corpus of Academic English (PICAE; Ackermann & Chen, 2013).

What these three lists have in common is that they all adopted a mixed-method approach that consists of corpus-derived frequency information and expert judgment to select and/or prioritize the most pedagogically useful word combinations.

It is still early to evaluate the impact of the newest corpus-derived lists of words and word combinations in the field of language learning and teaching. However, it is likely that they will be of more use in materials design[iii] than in the classroom proper

because, with an almost exclusive focus on frequency, they still require extensive pedagogic mediation to meet teachers' and learners' needs. One recent corpus-derived list that promises to be more useful for direct use in the classroom is the Phrasal Verb Pedagogical List, i.e. a selected list of the 150 most frequent phrasal verbs and their key meaning senses in the COCA (Garnier & Schmitt, 2015; see Sample Study 30_1). Similarly, the AKL words have been described in the Louvain English for Academic Purposes dictionary (LEAD) and students majoring in English at the University of Louvain (Belgium) use the web-based dictionary-cum-writing-aid in the EAP classroom and at home. The LEAD contains a rich description of academic words, with a particular focus on their phraseology (collocations and recurrent sequences) (cf. Granger & Paquot, 2015; Paquot, 2012). The lexical entries provide information derived from an analysis of a large corpus of academic texts (i.e. the academic component of the British National Corpus), as well as a range of discipline-specific corpora and English as a Foreign Language (EFL) learner corpora representing a wide range of first language (L1) populations.

[B] (Lexis-)grammar in Context

Analyses of raw and grammatically annotated corpora have led to a radical change in the way grammar is conceived of. First, corpus-based research has consistently shown that grammatical patterns differ systematically across registers, i.e. language varieties determined by their purposes and situations of use (e.g. newspaper writing vs. academic writing vs. fiction writing). To illustrate, Conrad (2000) shows that linking adverbials – words and expressions such as however, therefore and in other words that explicitly connect two units of discourse – are used much less frequently in newspaper writing than in academic prose. Second, corpus-based analyses have been instrumental in the

development of the concept of spoken or conversational grammar. Among the many contributions these studies made was "the incorporation into the purview of grammar of what, in previous centuries, had typically been regarded as marginal word-classes or as aberrations, items such as discourse markers [well, sort of, just] and interjections [oh, ouch]" (Carter & McCarthy, 2015, p. 5). Third, corpus-based analyses have demonstrated the patterned nature of language and close connections between grammar and lexis. Lexicogrammatical co-selection phenomena are ubiquitous in language (e.g. the verb deem is mostly used with the passive voice; cf. Biber, Johansson, Leech, Conrad, and Finegan, 1999), which implies that certain grammatical constructions ought to be described in relation to lexical items and vice versa. These developments have had impact on reference tools, notably grammars. For example, the Longman Grammar of Spoken and Written English (LGSWE, Biber et al. 1999), i.e. the first comprehensive reference grammar to adopt a corpus-based approach, describes the use of grammatical features in academic prose, news, fiction and conversation, with a special focus on lexicogrammatical patterning.

Textbooks for language learning and teaching lag behind. Quite a few studies have compared the use of language features in general reference corpora with the pedagogical descriptions of the same items in pedagogical material, and most notably course books (cf. Römer, 2005). Meunier and Reppen (2015) have recently shown that at least major publishers have tried to answer a repeated call for more corpus-informed materials. They carried out a case study on the treatment of the passive in corpus versus non-corpus informed grammar textbooks published between 2002 and 2012 and showed that unlike non-corpus informed textbooks, textbooks which are presented as corpus-informed do a better job at describing what we know from corpus research about the use

and the lexical associations of the passive. More generally, however, there is still room for improvement, especially in the treatment of collocations and lexicogrammatical patterns. Another area that deserves special attention is the role of learner corpus data.

[B] The Role of Learner Corpus Data

Learner corpora are systematic collections of authentic, continuous and contextualized language spoken or written by L2 learners stored in electronic format. Since the inception of learner corpus research at the turn of the 1990s, the field has always been driven by its desire to have an impact on language teaching and learning. The rationale is that while language descriptions based on native/expert corpora should typically be used to define the teaching agenda and ensure that the language taught reflects typical language use, the learner language descriptions should be used to modify this agenda to take into account the difficulties, needs, and unique trajectories of learners (e.g. Meunier, 2002).

One popular method to analyse learner corpus data is Computer-aided Error Analysis (CEA; Dagneaux, Denness, and Granger, 1998). This method focuses on the identification, correction and annotation of errors according to a standardized system of error tags with the help of a software tool, i.e. an 'error editor'. CEA has significant advantages over traditional error analysis. First, it relies on an error taxonomy that makes the error tagging procedure both more systematic and coherent. Error taxonomies are usually based on structural linguistic categories and include tags for errors that relate to form (i.e. spelling and morphology), grammar, lexis, lexico-grammar, punctuation, word order, etc. Second, it enables users to select an error category, for example that of morphology errors, extract a comprehensive list of all the errors of this type in the learner corpus, sort them in various ways and analyse them in context.

Despite the controversial status of errors in second language acquisition (SLA) and applied linguistics in general, results of computer-aided error analyses have been among the first types of learner corpus data to find their way in pedagogical materials, especially in monolingual learners' dictionaries. In the second edition of the Macmillan English Dictionary for Advanced Learners (MEDAL2, 2007), for example, 'Get it right' boxes contain authentic erroneous examples from the International Corpus of Learner English (ICLE; Granger, Dagneaux, Meunier, and Paquot, 2009), clear explanations of the source of the problems and clear and practical advice on how the errors can be corrected and avoided. With a view to ensuring maximum representativeness of the error notes, only errors that were both frequent and widespread across the sixteen mother tongue backgrounds represented in the ICLE were covered (Gilquin, Granger, and Paquot, 2007). The focus on errors is also in evidence in a range of recent pedagogical grammars mostly published by Cambridge. In Grammar and Beyond 4 (2013), for example, each unit features an 'Avoid Common Mistakes' box informed by careful analysis of the error-annotated subset of the Cambridge Learner Corpus (CLC) as illustrated in Figure 1. The CLC contains over 60 million words mostly of written exam scripts produced by English language learners taking Cambridge English exams all over the world, a large proportion of which is error-annotated.

<FIGURE 1: Grammar and Beyond 4, 'Avoid Common Mistakes' box, p. 75 ABOUT HERE>

The most large-scale research project based on the CLC, i.e. the English Profile Programme, however, adopts the opposite approach: instead of focusing on errors, it aims primarily at investigating what grammar and vocabulary can learners typically use correctly and appropriately at each level of the Common European Framework of

Reference for Languages (CEFR, Council of Europe, 2001). Results of the English Vocabulary Profile and the English Grammar Profile are available online in the form of a listing of known vocabulary (e.g. the verb 'to agree' is attested at A2 level in its meaning of 'to have the same opinion as someone' and at B1 in its meaning of 'to decide something with someone') and Can Do statements targeting core grammatical features such as tenses, articles, modal and auxiliary verbs, and word order by CEFR level (e.g. a B1 learner can use the past continuous with an increasing range of adverbs in the normal mid position) (http://www/englishprofile.org).

While the outcomes of the English Programme Profile should prove most useful for curriculum and syllabus design, material design and assessment, they are also limited in one important way, i.e. the exclusive focus on first appearance of specific vocabulary items or grammatical features in learner language. Thus, the conjunction 'because' and the adverb 'however' are described as A1 and A2 level words respectively; the approach adopted does not make it possible to highlight that these two connectors are then used with extremely high frequency by intermediate and advanced learners at the expense of other means of expressing cause and contrast. To uncover such patterns of overuse and underuse, the method of Contrastive Interlanguage Analysis (CIA; Granger, 1996) has been extensively used in learner corpus research. CIA involves two types of comparison:

1) A comparison of one or more interlanguage varieties with one or more reference varieties which aims to shed light on a range of linguistic features that characterize learner writing and speech, "not only errors but also instances of under- and overrepresentation of words, phrases and structures" (Granger, 2002, p. 12);

2) A comparison of different interlanguage varieties of the same language, i.e. the English of French learners, Spanish learners, Dutch learners, etc. so as to "differentiate between features which are shared by several learner populations and are therefore more likely to be developmental and those which are peculiar to one national group and therefore possibly L1-dependent" (Granger, 2002, p. 13).

Patterns of overuse and underuse shared by various learner populations should be used to inform generic tools such as pedagogical dictionaries and textbooks. However, this is still extremely rare. One exception is MEDAL2 where learner corpus-based information from ICLE was used not only to compile error notes but also to inform an extended writing section focusing on twelve rhetorical or organizational functions particularly prominent in academic writing (e.g. comparing and contrasting, expressing cause and effect). The analysis of learner corpora and their comparison with native corpora highlighted a number of problems which non-native learners from various mother tongue backgrounds experience when writing academic essays (e.g. lack of register awareness, phraseological infelicities, semantic misuse, and restricted lexical repertoire). These sources of difficulties are addressed explicitly in MEDAL2: Figure 2, for example, shows a "Be careful!" note which warns the reader against the excessive use of modal auxiliaries to express the function of expressing possibility and certainty. It is followed by a series of verbs, adverbs, adjectives and nouns which, though common in native academic writing, are rarely used by learners.

<FIGURE 2 'Be careful note' on the overuse of modal auxiliaries (MEDAL2, p. 17) ABOUT HERE>

The fact that most learners' dictionaries and other pedagogical materials are generic, i.e. target all learners whatever their first languages, has precluded the inclusion of notes and exercises that focus on difficulties that are specific to one particular L1 or family of L1s to date. With the generalized use of the Internet and online pedagogical tools, however, more customization is expected to become the norm and the LEAD, which contains both generic and L1-specific error notes that show up depending on the first language selected by the user, will probably not remain the only one of its kind (see above). This is highly desirable since results from learner corpus research have repeatedly demonstrated the influence of the mother tongue on correct and incorrect language use by (even advanced) learners, most particularly perhaps at the phraseological level (e.g. Paquot, 2015; see Sample Study Box 30_2).

[A] Challenges and Controversial Issues

As already highlighted above, corpus evidence does not necessarily equate with pedagogical relevance and the results of corpus studies should always be balanced with other factors such as learners' proficiency levels, learners' needs, and teaching objectives before they are used to inform curriculum design and teaching material development. Results of corpus studies, however, also crucially depend on the type of corpora used. What corpus should be used to inform English Language Teaching (ELT) around the world? Should a general reference corpus or a specialized corpus be used? Should the corpus be made up exclusively of native language production or is it acceptable if the corpus includes expert or near-native data irrespective of mother tongue background? What is the target language variety or norm? Should a corpus of British English or American English be favoured? In what circumstances should we opt for a corpus representing another English language variety? Is there a role for corpora of

English as a Lingua Franca in ELT? The way we answer these questions will impact heavily the type of corpus findings obtained. Similar questions also need to be addressed when a learner corpus is compared with another corpus to bring out learners' difficulties. Depending on the objectives of the study, learner corpus researchers also need to decide whether they want to compare EFL learner language with expert and/or student productions.

Another controversial issue relates to the use of authentic texts from corpora in language teaching. While Garnier and Schmitt (2015), for example, illustrate the different meanings of phrasal verbs with made-up examples, other researchers believe that corpus-derived (sometimes simplified) examples should be used because "intuition is often unreliable when it comes to patterns of language use" (Meunier & Reppen, 2015, p. 501). The debate is far from resolved and more research is needed into whether and to what extent learners benefit from exposure to authentic examples in the L2 classroom, especially as a function of variables such as the learners' proficiency levels, the teaching objectives and the topic of the course.

[A] Limitations and Future Directions

More insights from corpus linguistics, not just frequency information, need to find their way into curriculum design and material development. Future textbooks and other pedagogical tools should provide learners with a more integrated perspective on lexis and grammar; they should also place more emphasis on word combinations such as collocations and phrases and focus on variability in language use. I have also argued elsewhere for a corpus-based context-sensitive treatment of phraseology in (electronic) learners' dictionaries whereby collocations would not appear in undifferentiated lists but be organized by domain, mode or register (Paquot, 2015).

For such a ‘revolution’ to happen, however, there is a need for more pedagogical development and translation to make the results of corpus-based studies really usable in the classroom. Recent corpus-based resources such as the new GSL developed by Brezina & Gablasova (2015) are raw material and not as pedagogically useful as they could be (i.e. unlike West’s (1953) GSL, the new GSL comes with no rich description such as meaning senses or examples). More generally, the implications of current corpus-based research for pedagogy are usually not developed in any great detail. As put by Meunier in a discussion of learner corpus research and L2 language learning, “[t]here is an urgent need to go beyond the usual last paragraph of articles (or last slide of conference presentations) stating that ‘foreign language instruction could profit from this kind of investigation’ and efforts should be made towards providing teachers with ready-to-use teaching materials” (2011, p. 465). More systematic and large-scale comparisons of textbooks and other pedagogical resources with corpus data, of the type initiated by Römer (2005), should also be conducted, targeting not only materials produced by major publishing houses but also by local publishers, with a view to better document the discrepancies between textbook language and ‘real’ language.

Ultimately, however, the future role of corpus-based findings in L2 syllabus and material design will probably rest on the field’s ability to answer a repeated call for more pedagogically relevant corpora (cf. Meunier, 2011). Until very recently, learner corpus research has focused on English for academic and specific purposes in higher education. As a result, a large majority of learner corpora consist of exam scripts or argumentative essays; more learner corpora such as the Spanish Learner Language Oral Corpora (SPLLOC, http://www.splloc.soton.ac.uk/index.html) representing less advanced proficiency levels as well as other domains, registers and modes are definitely

needed. Importantly, for similar developments as the ones witnessed in ELT to occur in the teaching and learning of other foreign languages, there is an urgent need for more, better and bigger corpora of languages other than English. In a methodological synthesis of learner corpus research, Paquot and Plonsky (2017), found that 88% of the 378 studies in their sample examined English as the target language.

<SAMPLE STUDY 30_1 ABOUT HERE>

<SAMPLE STUDY 30_2 ABOUT HERE>

[A] Resources for Further Reading

Biber, D. & Reppen, R. (2015). The Cambridge handbook of corpus linguistics. Cambridge: Cambridge University Press.

The Cambridge Handbook of Corpus Linguistics offers a comprehensive survey of corpus-based linguistic research on English, including chapters on collocations, phraseology, grammatical variation, discourse and the description of English varieties (e.g. dialects, World Englishes, learner language), and a critical discussion of the state of the art in the field. Part IV presents corpus research into a variety of areas that are particularly relevant to language learning and teaching such as corpus-derived vocabulary lists and classroom applications of corpus analysis.

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

In this article, Granger discusses the major criticisms Contrastive Interlanguage Analysis (CIA), i.e. a comparative framework to analyse learner corpus data, has faced since its introduction in 1996 and presents a revised model, CIA², which, she argues, better acknowledges the important role played by variation in LCR and is generally more in line with the current state of foreign language theory and practice.

Granger, S., Gilquin, G., & F. Meunier (2015). The Cambridge handbook of learner corpus research. Cambridge: Cambridge University Press.

The Cambridge Handbook of Learner Corpus Research provides a unique account of learner corpus collection, annotation, methodology, theory, analysis and applications. Part III is devoted to the links between learner corpus research and SLA. Part IV will be of special interest to all readers who want to know more about the applications of learner corpora to teaching, pedagogical material development, and testing. Part V explores the diverse and growing natural language processing tasks that rely on learner corpora.

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), A taste for corpora: In Honour of Sylviane Granger (pp. 7-31). Amsterdam & Philadelphia: Benjamins.

In this chapter, Leech considers how the 'corpus revolution' made frequency information available in a totally unprecedented way from the 1960s onward. Then, he discusses the equation 'more frequent = more important to learn', what questions of frequency we really need to ask, and how far they can be answered in the present state of corpus linguistics.

Timmis, I. (2015). Corpus linguistics for ELT research and practice. New York: Routledge.

This volume is a user-friendly overview of corpus research in selected areas of language study relevant to ELT: lexis, grammar, English for Specific purposes, and spoken language. It provides a practical guide to undertaking ELT-related corpus research and bringing corpus material to the classroom.

[A] Keywords:

Frequency; collocations; phraseology; lexis-grammar; learner corpus

[A] Cross-referencing

Chapter 23

[A] Biography

Magali Paquot is a permanent FNRS research associate at the Centre for English Corpus Linguistics, Université catholique de Louvain. She is co-editor in chief of the International Journal of Learner Corpus Research and a founding member of the Learner Corpus Research Association. Her research interests include corpus linguistics, learner corpus research, vocabulary, phraseology, second language acquisition, linguistic complexity, crosslinguistic influence, English for academic purposes, pedagogical lexicography and electronic lexicography.

[A] References

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). The Longman grammar of spoken and written English. London: Longman.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. Applied Linguistics, 36(1), 1-22.

Carter, R., & McCarthy, M. (2015). Spoken grammar: Where are we and where are we going? Applied Linguistics, *amu080*, 1-21. doi:10.1093/applin/amu080

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), The Cambridge handbook of corpus linguistics (pp. 478-497). Cambridge: Cambridge University Press.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century. TESOL Quarterly, 34(3), 548-560.

Coxhead, A. (2000). A new Academic Word List. TESOL Quarterly, 34, 213-38.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. System: An International Journal of Educational Technology and Applied Linguistics, 26(2), 163–174.

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

Ellis, N. C. (2002). Frequency effects in language processing. Studies in Second Language Acquisition, 24,143–88.

Ellis, N. C., O'Donnell, M.B., & Römer, U (2014). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. Linguistic Approaches to Bilingualism, 4, 405–31.

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), Theories in second language acquisition: An introduction (pp. 75- 93). New York & London: Routledge.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. Applied Linguistics, 35(3), 305-327.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. Language Teaching Research, 19(6), 645-666.

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

Granger, S. (1996). From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), Languages in contrast: Text-based cross-linguistic studies (pp. 37–51). Lund: Lund University Press.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson & J. Hulstijn (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (Vol. 6, pp. 3–33). Amsterdam & Philadelphia: John Benjamins.

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

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

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

Lessard-Clouston, M. (2012). Word lists for vocabulary learning and teaching. The CATESOL Journal, 24(1), 287-304.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. Applied Linguistics, 33(3), 299-320.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, S. Petch-Tyson & J. Hulstijn (Eds.),

Computer learner corpora, second language acquisition and foreign language teaching (pp. 119-142). Amsterdam & Philadelphia: John Benjamins.

Meunier, F. (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths. Revista Brasileira de Linguística Aplicada, 11(2), 459-477.

Meunier, F., & Reppen, R. (2015). Corpus versus non-corpus-informed pedagogical materials: grammar in focus. In D. Biber & R. Reppen (Eds.), The Cambridge handbook of corpus linguistics (pp. 498-514). Cambridge: Cambridge University Press.

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. London & New-York: Continuum.

Paquot, M. (2012) The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. In S. Granger & M. Paquot (Eds.), Electronic lexicography (pp. 163-185). Oxford: Oxford University Press.

Paquot, M. (2015). Lexicography and phraseology. In D. Biber & R. Reppen (Eds.), The Cambridge handbook of corpus linguistics (pp. 460-477). Cambridge: Cambridge University Press.

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

Römer, U. (2005). Progressives, patterns, pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics. Amsterdam: Benjamins.

Simpson-Vlach, R., & Ellis, N. (2010). An Academic Formulas List: New methods in phraseology research. ELT Journal, 31(4), 487-512.

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

West, M. (1953). A general service list of English words. London: Longman.

Sample Study 30_1:

Garnier, M. & Schmitt, N. (2014). The PHaVE List: a pedagogical list of phrasal verbs and their most frequent meaning senses. Language Teaching Research, 19(6), 645-666.

Background: It is notoriously difficult for practitioners to select the items to be taught among the thousands of phrasal verbs in English. Available lists of phrasal verbs (PVs) are of limited value because they do not account for polysemy, which is a serious pedagogical shortcoming as research indicates that phrasal verbs have on average 5.6 meaning senses.

Research gap: The main objective of Garnier and Schmitt's (2014) study was to develop a pedagogical list of PVs that would be helpful to teachers and learners by providing their most frequent meaning senses together with example sentences.

Research method:

- Selection of a previously published corpus-based list of the 150 most frequently used PVs in American and British English
- Identification of their most frequent meaning senses in the COCA.
- Two criteria for inclusion in the PHrasal VErb Pedagogical List (PHaVE List): (i) the meaning senses for each item should account for at least 75% of all occurrences of this phrasal verb in their corpus search and (ii) all the meaning senses should account for at least 10% of the same occurrences.

Key results: Garnier and Schmitt showed that a large majority of the most frequent PVs in English are polysemous, and that, on average, around two meaning senses account for

at least 75% of all the occurrences of a single PV in the Corpus of Contemporary American English. The PHaVE list thus focuses on the key meaning senses of the 150 most frequent PVs and gives their percentage of occurrence, along with definitions and examples written to be accessible for second language learners.

Comment: Garnier and Schmitt used corpus data to inform the selection of key meaning senses to be included in the PHaVE list but created example sentence themselves. As discussed above, this will be regarded by some as a limitation. However, among other things, Garnier and Schmitt took care of modelling sentences from various sources found on the internet as well as from the COCA itself, with the aim to produce as natural and authentic sentences as possible. As a result, I believe the PHaVE is to be commended as the output of a truly corpus-based pedagogic entreprise.

Sample Study 30_2:

Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. Second Language Research, 33(1), 13-32.

Background: Empirical evidence suggests that learners are sensitive to L2 frequency (Ellis, 2002). This sensitivity applies not only to individual lexical items but to the frequency of L2 collocations, lexical bundles and constructions as well (e.g. Ellis, O'Donnell, and Römer, 2014). Cross-linguistic influence is another important variable in SLA but it is usually approached in terms of the potential influence on the L2 of the presence or absence of a linguistic phenomenon in the L1.

Research gaps: Despite the large number of studies focusing on cross-linguistic

influence and current interests in frequency effects in SLA, there has been very little research on the potential influence of L1 frequency on L2 acquisition and use. The study thus aimed to answer the following research questions:

1. To what extent is there a relationship between the frequency of a lexical bundle in the EFL learners' written productions and the frequency of its equivalent form in the learners' first language?

2. Can French and Spanish learners' different use of lexical bundles be explained by frequency differences in the two Romance languages?

Method:

-
- Extraction of statistically significant three-word lexical bundles in the Spanish and French sub-components of the second version of the *International Corpus of learner English* (ICLE; Granger et al., 2009)
- Manual selection of lexical bundles with discourse or stance-oriented function
- Use of Jarvis's (2000) unified methodological framework for the study of L1 influence
- Identification of L1 translational equivalent forms
- Frequency comparisons across the two L1s on the basis of web corpora of French and Spanish.

Results: Strong and positive monotonic correlations were found between the frequency of a lexical bundle in the EFL learners' written productions and the frequency of its equivalent form in the learners' first language. Results also confirm that different patterns of use across the two L1 learner populations (e.g. Spanish learners' frequent use of the lexical bundle 'in the case of') may be explained by frequency differences in

L1 French and Spanish (Spanish ‘en el caso de’ is much more frequent than French ‘dans le cas de’).

Comment: The lexical bundle approach represents “corpus linguistic methodology at its most heuristic, i.e. as a raw discovery procedure” (De Cock, 2004: 227). Coupled with Jarvis’s (2000) framework, it proved most useful to identify in learner corpora transfer effects that until now have been little documented in the SLA literature.

[i] Work on frequency lists has also spread to other languages, as seen in the ‘Routledge Frequency Dictionaries’ series which now offers corpus-based core vocabulary (alphabetical, frequency and thematically-organized) lists for learners of fourteen languages including Arabic, Dutch, French, German, Japanese, Mandarin Chinese, Portuguese, Russian and Spanish.

[ii] Thus, if a word occurs with relatively high frequency in the corpus but is only found in a limited number of corpus parts of the same length, the ARF will be low.

[iii] For example, the Academic Keyword List has been used to inform the Oxford Learner’s Dictionary of Academic English and is presented as a major asset of the five-level academic English course Skillful published by Macmillan Education. Similarly, the Academic Collocation List appears in an appendix of the Longman Collocations Dictionary and Thesaurus.