ReCALL **26**(2): 121–127. 2014 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344014000068
First published online 14 February 2014

ReCALL special issue: Researching uses of corpora for language teaching and learning

Editorial

Researching uses of corpora for language teaching and learning

ALEX BOULTON

University of Lorraine and CNRS, France
(email: alex.boulton@univ-lorraine.fr)

PASCUAL PÉREZ-PAREDES

Universidad de Murcia, Spain
(email: pascualf@um.es)

Corpora, in a broad sense, have had a role to play in language teaching and learning for many decades. Of note are Thorndike and Lorge's *Teacher's Word Book of 30,000 Words* (1944), West's *General Service List* (1953), or Gougenheim (e.g., 1958) and colleagues' work on the *Français Fondamental*, but these mostly involved indirect applications, especially identifying frequent items (forms, meanings and uses) for inclusion in syllabuses and language programmes. Such work continues in lexicography, largely thanks to the pioneering Cobuild work led by the late John Sinclair (1987) (almost all major dictionaries, grammar books and manuals today are corpus-based to some extent, for major world languages at least), not to mention the proliferation of frequency lists (e.g., the series of Routledge Frequency Dictionaries[1]) and various academic research projects, from Coxhead's *Academic Word List* (2000) to Martinez and Schmitt's (2012) *Phrasal Expressions List*, as corpora have much to tell us not just about 'usage' but also about collocations, multi-word units and 'chunking' (e.g., McCarthy, 2004).

Obviously these applications were indebted to lexicographical work carried out by linguists with an interest in what we call today corpus linguistics. However, a separate current emerged in the 1980s where language teachers, often close to a research team working with corpus data, saw a connection between research and teaching and began tentatively to explore how corpora could be of more direct benefit to their learners. Such applications are closely associated with the work of the late Tim Johns, but the first academic publication seems to be from McKay (1980) in San Francisco, whose students

[1] http://www.routledge.com/books/series/RFD/

explored printed corpus materials to observe verb patterns in context. Ahmad *et al.* (1985) took this further, describing how their students benefited from direct access to an electronic corpus to pursue their own questions – a surprisingly ambitious approach given the capabilities of computers, corpora and software at the time. This is the scene to which Johns (1990) introduced the term *data-driven learning* or DDL (see also Johns & King, 1991). It is the same scene which saw the founding of the biennial *Teaching and Language Corpora* (TaLC) conference series in 1994, each giving rise to a high-quality volume of selected papers (Wilson & McEnery, 1994; Wichmann *et al*., 1997; Botley *et al*., 1996; Burnard & McEnery, 2000; Kettemann & Marko, 2002; Aston *et al*., 2004; Hidalgo *et al*., 2007; Kübler, 2011; Frankenberg-Garcia *et al*., 2011; Thomas & Boulton, 2012; Leńko-Szymańska & Boulton, forthcoming).

The first TaLC Conference and accompanying publications brought together not only researchers and teachers interested in the pedagogic applications of corpora, but also linguists who were already familiar with language description and corpus methodology. They outlined the main reasons why corpora, and corpus-derived data, were already perceived as an asset in language teaching and learning all those years ago: (1) computers and storage at the time were improving dramatically; (2) there was a new interest in authentic data and usage in language education; and (3) there was a consensus that learners were adopting new, more active roles in their learning process. These three aspects have remained at the heart of the academic debate and the professional practice of teachers using corpora ever since. They also established much of the research and teaching agenda of the next two decades, including direct and indirect uses of corpora for teaching, learning, testing and as a reference resource, the suitability of existing corpora for pedagogical purposes, the development of languages for specific (LSP) corpora, multilingual corpora, parallel corpora, learner corpora, *ad hoc* disposable corpora, and more user-friendly tools for learners and teachers alike. This special issue underlines the continued relevance of these issues while also showcasing new developments in researching uses of corpora.

Though the field has yet to reach full maturity (Thomas & Boulton, 2012: 8), there is today a tradition of research and practice which provides a reassuring background to uses of corpora in language teaching and learning. Researchers feel less need to explain the field and justify the benefits of corpora for pedagogical purposes, and are broadening their interests by applying research methodologies already in place in applied linguistics, particularly in second language acquisition (SLA) and fields such as error analysis (e.g., Granger *et al*., 2002). The focus is switching from corpus linguistics to language pedagogy; in other words, rather than trying to bring corpus linguistics into the language classroom, the emphasis is increasingly placed on the L2 user and how he or she might benefit from corpus linguistic tools and techniques. Much of the impetus for this comes from applied linguists and language teachers, as opposed to specialists in corpus linguistics; to paraphrase Widdowson (2000), we might thus talk of applied corpus linguistics rather than corpus linguistics applied. This change can be seen partly as greater attention is accorded to the learner rather than to the tools and techniques themselves; emic studies, for example, have highlighted a number of reservations (e.g., in deciding points to develop, formulating appropriate queries, and interpreting output in the unusual form of truncated KWIC concordances). Other work has focused on developing more pedagogically relevant corpora and software, as new generations of learners became more computer-literate too.

One alleged problem is the repeated allegation of a dearth of empirical studies in DDL and associated areas. The first review of empirical studies was provided by Chambers (2007a), who concluded that researchers and teachers typically overlapped their roles, that empirical research was mostly small-scale and quantitative, and confined to the university context. Though today it is possible to collate a hundred or more academic papers which seek to evaluate some aspect of corpus use for L2 users (Boulton, 2010), it is clear that more diverse and rigorous research designs are needed to focus on the complex phenomena covered. As Pérez-Paredes (2010: 54) puts it: "Can we really claim that direct transfers from [linguistic] research will be successfully implemented in a learning environment?" Our intention for this volume was therefore not to repeat a previous special issue of *ReCALL* (*Integrating corpora in language learning and teaching*) edited by Angela Chambers (2007b), but to focus specifically on empirical research and evaluation in DDL-like approaches.

This special issue is rather longer than usual, reflecting the high quality of papers received. The order of presentation is inevitably difficult, but generally goes from the more experimental to the more observational, from short-term to longer term, from etic to emic, from controlled to open, with each study bringing something new to the table. Many of the papers in this volume continue existing research traditions in DDL, with much of the work at university level among relatively advanced learners for writing purposes. However, some are conducted in less well-researched areas in secondary schools, at lower levels of proficiency, or for speaking purposes. Some of the papers are rigorous experimental / laboratory studies lasting just a few minutes (including one semi-replication study), generally on a specific (lexicogrammar) language point, while others are more ecological and cover an entire semester with a much more open language focus; these are generally backed up by questionnaires and other instruments to gather feedback from participants. Much use is made of control and experimental group comparisons to explore corpus use *vs*. no treatment or traditional treatment (teaching, dictionary use, etc.), or variations of DDL (e.g., inductive *vs*. deductive; for comprehension *vs*. production); several also feature delayed post-tests in addition to the more common immediate post-tests. The tests themselves may be highly controlled, but several look at more open-ended language use, especially in the form of writing. The corpora range from the large, publically available British National Corpus (BNC) or Corpus of Contemporary American English (COCA) to smaller, purpose-built corpora; some involve the learners' own productions, and are used in quite different ways, from learning aids to reference tools.

Ana Frankenberg-Garcia provides an extension or semi-replication (rather than a single-variable replication) of an earlier study comparing single and multiple concordance lines and dictionary definitions as a language reference resource. One crucial difference is that the present paper deals with secondary-school students, a hugely under-researched population in DDL work. Students were randomly assigned to one of four groups using materials featuring a single corpus example, multiple corpus examples, dictionary definitions, or no materials in a control group. The experiment clearly distinguished between comprehension and production in both materials and tests. The examples were selected from large corpora and featured relevant contextual clues to meaning (for the decoding task) and appropriate colligations / collocations (for the encoding task). The dictionary definitions were similarly chosen (mainly from Cobuild) to include the relevant senses only, and also provide the necessary colligational / collocational clues. In the test itself, the decoding task featured unfamiliar items in a multiple-choice gap-fill format; the encoding task required

https://doi.org/10.1017/S0958344014000068 Published online by Cambridge University Press

translation from L1 Portuguese of 'known' but error-prone items using lexical prompts. In the comprehension test, the group with dictionary definitions fared best, but not significantly better than the multiple corpus examples group; even the group with a single corpus example significantly outperformed the control. In the production test, all three experimental groups scored higher than the control, but only those with multiple corpus examples significantly so. The results support the findings of the earlier study, and are taken to mean that learners can derive useful information on both meaning and usage from corpus examples, that multiple examples are better than one, and that different types of examples are needed for encoding and decoding. There are clear implications for compilers of learners' dictionaries as well as for language teachers and materials developers.

Yukio Tono, **Yoshiho Satake** and **Aika Miura** report on how Japanese learners of English can use corpora to help revise their own writing. Texts from 93 undergraduates were error-coded, with each error being broadly classified as omission, addition or misformation, and two errors selected from each essay: one appropriate for correcting using corpus data, one without. Three weeks later, the participants were given a 20-minute initiation to the corpus tool and then required to revise their first drafts, using the tool for one of the errors as indicated by the codes. The results show that learners can make useful revisions to texts using corpus data, and that there is a significant difference in the accuracy rate among the three error types: omission and addition errors were more easily identified and corrected than were misformation errors. The analysis discusses error types in much more detail, and also compares learners at different levels of language proficiency.

Zeping Huang focuses on awareness of the patterning of abstract nouns among 40 Chinese students majoring in English. Following the pre-test writing assignment, an experimental group was provided with paper-based concordance lines to study the collocations of five abstract nouns, while the control group was allowed to consult dictionaries for the usage of the words involved. The results of the immediate post-tests showed that the writing of the experimental group contained not only fewer linguistic errors among the target abstract nouns, but also greater variety of collocational and colligational patterns; this tendency continued in delayed post-tests two weeks later. Data from questionnaires and learning journals show the learners have favourable opinions of the approach overall and are able to adapt to the inductive approach involved in such corpus work, but do bring to light some reservations.

Jonathan Smart compares different types of corpus-based instruction, using paper-based materials derived from Mark Davies' corpora[2] for the passive voice. The first corpus group experienced inductive DDL, analysing printed concordances in collaborative tasks; the second also used corpus-based materials but followed a deductive presentation–practice–production paradigm (PPP). The PPP format was also used with a third group based on conventional teaching materials. Students of L1 Arabic and Chinese participated in the 4-hour experiment during an intensive language course in the US, completing various tasks on form, meaning and use in immediate and delayed post-tests in the same format as a pre-test. All three groups improved their performance in the immediate post-test, but the two deductive groups returned essentially to their pre-test scores in the delayed post-test. The inductive DDL group, however, experienced the greatest gains (+51% in the immediate post-test), an increase which remained

[2] http://corpus.byu.edu

https://doi.org/10.1017/S0958344014000068 Published online by Cambridge University Press

significant in the delayed post-test (+43%). This is taken to suggest that the main benefit of DDL lies in its inductive nature rather than just the use of authentic language samples derived from corpora. It may of course still be that deductive DDL or traditional instruction would be of particular benefit to some students according to individual learning styles and preferences.

Elena Cotos examines the effects of corpora in noticing, exploring and reusing linking adverbials among graduate students from various disciplines enrolled on a 10-week writing course at a university in the US. One group used a 40K-word corpus of research articles relating to their own discipline; the other additionally had access to a similar sized corpus of academic writing produced by the participants themselves. Both groups virtually doubled their use of target items in their writing after the course (88%), an increase sustained and even extended four weeks later (93%); they also used a greater variety of linkers, and to better effect semantically and syntactically in context. Both groups also improved significantly from the pre- to post-tests with a very large effect size. Between groups, access to a local learner corpus gave rise to more frequent, varied and appropriate use of linkers than just using a native corpus, and provided significantly better scores on the post-test. While questionnaires elicited positive reactions from both groups, work on their own and peers' texts was received particularly favourably, and seems to have led to greater cognitive processing.

Joe Geluso and **Atsumi Yamaguchi** attempt to integrate corpus use into an original course design with the focus on spoken fluency. Lower-intermediate Japanese students were introduced to COCA over three weeks, then investigated it largely for formulaic sequences (FSs). Activities were designed around student-led lessons on favourite FSs, and preparation of a 'speaking journal' which encouraged them to use the FSs in unpredictable encounters with native speakers. Recordings of these interactions suggest the students were generally able to use the FSs appropriately in context. Questionnaires and interviews revealed largely positive reactions, though some reservations emerged about unfamiliar vocabulary and truncated concordance lines, as well as possible mistrust of corpus data (some students preferring to have a dictionary on hand to check their findings). Some reported that the class period was insufficient to complete their corpus queries, but that they used the corpus enough overall.

Ji-Yeon Chang provided an introduction to corpus use for post-graduates in engineering and computer science in Korea to help with writing. The students consulted corpora on their own for their individual needs, submitting their writing regularly and meeting weekly with the researcher over 22 weeks for corpus-based feedback (cf. Johns' kibitzers[3]). The work featured both a general corpus (COCA) and a local corpus of research articles in the relevant fields using AntConc[4]. The analysis uses grounded theory in an emic approach to examining recordings of the weekly individual sessions and the final questionnaires. Overall, the participants appreciated the use of corpora as examples of writing, especially at the level of lexicogrammar, though they did find it time-consuming and linguistically quite demanding. Some deemed the local corpus to be more relevant to their specific needs, though less credible as it included non-native writing (albeit in successfully published papers).

[3] A sample is available at http://lexically.net/TimJohns

[4] http://www.antlab.sci.waseda.ac.jp/antconc_index.html

https://doi.org/10.1017/S0958344014000068 Published online by Cambridge University Press

COCA scored points from its mere size, as more results could be retrieved even in micro-registers, but was found difficult to use by some. A further worry concerned the risk of plagiarism. Overall, support is found for students being able to use corpora successfully on their own for their specific writing needs, and for corpora as a useful complement to web search engines for linguistic purposes, even in an ambitiously autonomous programme such as this. However, it is suggested that further training would be useful in deciding what to search for and how to find it, and in helping students to compile their own corpora relevant to their specific needs.

Agnieszka Leńko-Szymańska describes a course designed for Master's students including trainee teachers in Poland, the objective being to promote corpus literacy for a variety of uses. A pre-course questionnaire shows almost no prior awareness of corpora, and limited motivation. The course itself ran over one semester (14 weeks) in a combination of lectures and hands-on work to introduce a wide range of corpora and associated tools (all simple, free, and readily available), as well as corpus-related materials and techniques from lexicogrammar to genre awareness; a useful link is provided to the online course page. A final questionnaire found generally promising evolution in opinions and positive reactions, especially for content geared towards the final project to compile a corpus and present a lesson plan for exploiting it. Some comments are more ambivalent, notably claims that there was too much to do but that they would have liked to do more to really familiarise themselves with the tools. The general conclusion is that an introduction to corpora can require far more than the 22 hours allotted, and that cross-fertilisation would be beneficial if students could recycle the tools and techniques in other courses.

Acknowledgements

The editors are indebted to the following scholars who reviewed papers for this special issue: Guy Aston, Angela Chambers, Maggie Charles, Frederik Cornillie, Fiona Farr, Robert Fischer, Ana Frankenberg-Garcia, John Gillespie, Michael Goethals, Detmar Meurers, Hilary Nesi, Hans Paulussen, Simon Smith and Chris Tribble. We would also like to thank the *ReCALL* editors June Thompson and Françoise Blin for their confidence in us, and would like to acknowledge the work of the Cambridge Journals team for making this special issue possible.

References

Ahmad, K., Corbett, G. and Rogers, M. (1985) Using computers with advanced language learners: An example. *The Language Teacher*, **9**(3): 4–7.

Aston, G., Bernardini, S. and Stewart, D. (eds.) (2004) *Corpora and language learners*. Amsterdam: John Benjamins.

Botley, S., Glass, J., McEnery, A. and Wilson, A. (eds.) (1996) *Proceedings of TaLC 1996. UCREL Technical Papers*, **9**. Lancaster: University Centre for Computer Corpus Research on Language.

Boulton, A. (2010) Learning outcomes from corpus consultation. In: Moreno Jaén, M., Serrano Valverde, F. and Calzada Pérez, M. (eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching*. London: Equinox, 129–144. Electronic supplement: Empirical research in data-driven learning – A summary. http://bit.ly/BoultonATILF

Burnard, L. and McEnery, T. (eds.) (2000) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang.

Chambers, A. (2007a) Popularising corpus consultation by language learners and teachers. In: Hidalgo, E., Quereda, L. and Santana, J. (eds.), *Corpora in the foreign language classroom*. Amsterdam: Rodopi, 3–16.

Chambers, A. (ed.) (2007b) Integrating corpora in language learning and teaching. *ReCALL*, **19**(3).

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly*, **34**(2): 213–238.

Frankenberg-Garcia, A., Flowerdew, L. and Aston, G. (eds.) (2011) *New trends in corpora and language learning*. London: Continuum.

Gougenheim, G. (1958) *Dictionnaire fondamental de la langue française*. Paris: Didier.

Granger, S., Hung, J. and Petch-Tyson, S. (eds.) (2002) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Hidalgo, E., Quereda, L. and Santana, J. (eds.) (2007) *Corpora in the foreign language classroom*. Amsterdam: Rodopi.

Johns, T. (1990) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, **10**: 14–34.

Johns, T. and King, P. (eds.) (1991) *Classroom concordancing*. *English Language Research Journal*, **4**: iii–iv.

Kettemann, B. and Marko, G. (eds.) (2002) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi.

Kübler, N. (ed.) (2011) *Corpora, language, teaching, and resources: From theory to practice*. Bern: Peter Lang.

Leńko-Szymańska, A. and Boulton, A. (eds.) (Forthcoming) *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins.

Martinez, R. and Schmitt, N. (2012) A phrasal expressions list. *Applied Linguistics*, **33**(3): 299–320.

McCarthy, M. (2004) This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga: The Irish Yearbook of Applied Linguistics*, **20**: 30–52.

McKay, S. (1980) Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly*, **14**(1): 17–26.

Pérez-Paredes, P. (2010) Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario. In: Harris, T. and Moreno Jaén, M. (eds.), *Corpus linguistics in language teaching*. Bern: Peter Lang, 53–73.

Sinclair, J. (ed.) (1987) *Looking up: An account of the Cobuild project in lexical computing*. London: Collins.

Thomas, J. and Boulton, A. (eds.) (2012) *Input, process and product: Developments in teaching and language corpora*. Brno: Masaryk University Press.

Thorndike, E. and Lorge, I. (1944) *The teacher's word book of 30,000 words*. New York: Columbia University.

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

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.) (1997) *Teaching and language corpora*. Harlow: Addison Wesley Longman.

Widdowson, H. (2000) On the limitations of linguistics applied. *Applied Linguistics*, **21**(1): 3–25.

Wilson, A. and McEnery, T. (eds.) (1994) *Corpora in language education and research. UCREL Technical Papers*. **4**. Lancaster: University Centre for Computer Corpus Research on Language.