HAL
open science

Data-driven learning and language pedagogy

Alex Boulton

To cite this version:

Alex Boulton. Data-driven learning and language pedagogy. S. Thorne & S. May. Language, Education and Technology: Encyclopedia of Language and Education, 3, Springer, pp.181-192, 2017, Encyclopedia of Language and Education: Language and Technology. hal-01854664

HAL Id: hal-01854664

https://hal.science/hal-01854664v1

Submitted on 7 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Alex Boulton. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (eds.), *Language, Education and Technology: Encyclopedia of Language and Education*. New York: Springer. DOI 10.1007/978-3-319-02328-1_15-1

Abstract
Language corpora have many uses in language study, including for learners and other users of foreign languages in an approach that has come to be known as data-driven learning (DDL). This boils down to the learner's ability to find answers to their questions by using software to access large collections of authentic texts relevant to their needs, as opposed to asking teachers or consulting ready-made reference materials. As such, not only do corpora contain the potential to answer many language questions, the consultation itself is likely to lead to improved language awareness and noticing. This chapter discusses the nature of corpora and their relevance in language learning, outlining the processes involved in DDL, and looks at the history and research development in the field from its beginnings to the present day, taking into account its limitations and gaps in our current knowledge with an eye to the future.

Keywords: Corpus, Data-driven learning, DDL, Corpus-based language learning

Introduction

The essential condition for any language learning is exposure to the language itself. In foreign language contexts in particular, such exposure may be inadequate: Zahar et al. (2001: 558) estimate that an hour's reading may lead to the incidental learning of just two words; at that rate, it would take decades to build up a sizeable vocabulary. Clearly exposure alone is not enough in such cases and may be complemented by formal instruction intended to help speed up the process by drawing learners' attention to important points, explaining, demonstrating, providing examples, and so on. Fortunately, syntheses show that instruction does make a difference (e.g., Norris and Ortega 2000), though that does not mean that any type of instruction works equally well for all learners in all contexts (Hattie 2009). Formal teaching may however oversimplify things: the contrived language that is presented, the all-purpose definitions provided, the abstract rules given, as well as the structured tasks to be completed. These can all have their uses, but if that is all there is then they may lead to dependence on the teacher and an inability to work with authentic language – i.e., to make the most of any real exposure.

This is where language corpora can be of use in what has come to be known as data-driven learning (DDL). The basic concept is to allow massive exposure that is still organized and focused. Using the power of computer software, learners can query large collections of texts relevant to their needs, looking at frequencies and distributions and multiple occurrences of target items in context. This essentially constructivist, inductive approach means they can then reach their own conclusions that are meaningful to them individually, and the cognitive processing should lead to longer retention than simply "being taught." This may be quite time consuming at the start, but the real advantages lie not so much in the explicit knowledge gained as in the processes involved – ability to deal with authentic texts in different genres; awareness of frequency, chunking, and collocation; noticing forms and

variation; formulating hypotheses and inferring meanings; and so on. In other words, it should help students become better language learners and users.

Early Developments

The word "corpus " can mean different things to different people in a variety of disciplines. In corpus linguistics, it is a large collection of authentic texts that has been deliberately sampled to be representative of the type of language one is interested in; it is accessed by software often called a concordancer, though it can usually do more than just concordancing as we shall see below. This is, however, a prototypical definition, and corpus linguistic tools can be used with just a few thousand words (wherever repeated searches can beneficially be conducted by computer rather than regular reading), a single text (e.g., a novel), "non-authentic" text (arguably simplified readers or textbooks, or learner essays), and collected automatically (in the case of web-compiled corpora) or at least partly serendipitously (depending on resources available). In language teaching, the overriding criteria are pedagogical rather than theoretical, and the ad hoc creation of a small, specialized corpus of texts can be much more relevant to learners' needs than some of the large, general-purpose corpora that are publicly available.

The first modern corpus is no doubt the Brown corpus, a million words carefully sampled from 500 extracts of texts that had been published in 1961 (Kučera and Francis 1967). This was partly motivated by dissatisfaction with the tools then available for describing English which derived largely from intuition and fortuitous examples. The goal here was to introduce greater scientific rigor from a more systematic base. Later, the Bank of English at Birmingham University was designed for linguistic purposes but also with pedagogical aims in mind (see Sinclair 1987, for a review). This monitor corpus was designed to increase over time to account for developments in British English, but originally just over 7 million words were used to produce the first Cobuild dictionary. The corpus could be sorted on the computer and then the short contexts (concordances) were printed out for every occurrence of every item; the lexicographical work took place entirely on paper, similar to earlier hand-compiled concordances from the Bible or Shakespeare, for example. The 100 million words of the British National Corpus (see Aston and Burnard 1998) represented a truly monumental undertaking when it was built in the early 1990s, but later advances made it possible for a single person to create the Corpus of Contemporary American English semi-automatically from the Internet, currently standing at 520 million words (see Davies 2009). Entirely automated procedures now mean that billion-word corpora are regularly compiled (e.g., Baroni and Bernardini 2006). At the top end of the scale, the searchable Internet has debatable status as a corpus but can be usefully queried via regular search engines or more specialized software for pedagogical purposes.

In the hands of experts, corpora can be useful in preparing all kinds of pedagogical materials and resources, from general and specialized dictionaries to grammar books and usage manuals, from syllabus design to testing, from wordlists to coursebooks. Such uses are beyond the scope of this chapter, which is concerned with how learners can use corpora directly.

This is a pre-publication version. For the version of record, please see DOI: 10.1007/978-3-319-02328-1_15-1
(or email me at alex.boulton@atilf.fr)

Many of the early attempts at learner corpus consultation are based around Birmingham with teachers in contact with the Cobuild project. McEnery and Wilson (1997: 12) mention uses dating back to 1969, though the earliest publication seems to be from San Francisco where McKay (1980) describes activities encouraging learners to identify grammatical patterns to distinguish semantically similar verbs, based on sentences printed out from a corpus. In Surrey, Ahmad et al. (1985) had their advanced learners using a computer to query a corpus directly, though such early software could be exasperatingly slow. Things really took off in the late 1980s back in Birmingham, with Tim Johns as a leading pioneer often cited as the founding father of DDL. He created a concordancer specifically designed with language learners in mind (MicroConcord, which later morphed into WordSmith Tools; see http://www.lexically.net/wordsmith), and published a number of papers explaining many different ways in which he used corpora for and with his students, with many widely-cited sound bites and frequent citations, especially from the seminal collection of papers he coedited (Johns and King 1991) and which included two papers of his own.

Johns created or used different types of corpora – scientific texts, parallel corpora of translations, a single novel – to be relevant to the learners' needs. Authenticity was important in terms not just of text but also of needs and indeed the task, since corpus consultation involves exploring and thinking about language – crucial to any language learning. In this way the learner was seen as a researcher with direct access to the data, and the teacher as guide rather than dispenser of linguistic knowledge. Proactive materials could be printed out for repeated use with lower-level students for general purposes, while more advanced students could explore the corpus individually or collaboratively, using the concordancer themselves for serendipitous browsing or focused querying. Johns would leave the concordancer on in his classes as an informant and used it in his one-to-one advising sessions to help with academic writing. For him, DDL was not just learner centered but also provided a means to keep language (especially lexicogrammar) firmly center stage. All of this was argued to lead to greater autonomy; indeed, his final paper (Johns et al. 2008) provides some evidence that the DDL participants outperformed the control group even on items that had not been explicitly covered, suggesting that the processes may improve language skills as a whole. The tremendous variety of uses of corpora promoted by Johns set the agenda for years to come, though of course he was not alone, especially in the UK and Europe. Of particular note is the biennial Teaching and Language Corpora (TaLC) conference series inaugurated in Lancaster in 1994, each event giving rise to a selected volume of papers; further information can be found in Thomas and Boulton (2012: 17–34).

Major Contributions

Before going any further, it may be useful to see what DDL actually looks like. Traditionally the user sees corpus data in the form of a concordance, typically in KWIC (key word in context) format. Fig. 1 shows a random selection of 20 concordance lines taken from a corpus of academic writing (110 papers focusing on DDL, 600 k running words) for the search term, "e.g." here centered and in bold. This very simple formal example highlights a number of features which may be useful to learners. On the left, it is immediately apparent that most occurrences occur within brackets (other searches show that this is true in 85% of cases), the implication being that it is unusual in the syntax of the main sentence and should not be overused in this way in this type of writing. On the right, the presence or absence of

a comma owes more to individual journal style guides than any generalizable pattern. Beyond that, it is often used to introduce references, which can lead to further searches for citation practices and discussion of whether in other disciplines or other languages, research cited is thus typically relegated to brackets or not. Most corpus analysis software offers this basic concordancing function. Other features include frequency counts of individual words or clusters, collocates, distributions, and so on, all of which can prove useful to L2 learners.

```
1.   Jackson 1997); and in translation studies (e.g. Pearson 1996; Aston 1999; Mallikamas 2001;
2.  t is against the ideal of learner autonomy (e.g., Johns, 1991a). As in real life, learners p
3.  ally favourable in this and other research (e.g. Johns 2002; Hadley 2002; Ciesielska-Ciupek
4.  er-initiated and teacher-initiated queries (e.g. Yoon 2008). As with corpus use in general,
5.  oncentrating on written academic discourse (e.g. textbooks and articles), and the other on o
6.  significantly over that of a control group (e.g. Goyette). Indeed the point of click-on reso
7.   offer more readily recontextualised input (e.g. EEL sub-parts of the EGAP and ESAP register
8.  r-, or under-use of particular L2 features (e.g., Granger, 1998; Granger, Hung, & Petch-Tyso
9.   studies have examined vocabulary learning (e.g. Kaur & Hegelheimer, 2005; O'Sullivan & Cham
10. ed according to formal linguistic criteria (e.g., verbs, nouns, prepositions) or according t
11. lass activities into the intermediate class,e.g., letting students have hands-on practice, w
12. evens, 1991; Tribble, 1991) or translation (e.g., Aston, Gavioli, & Zanettin, 1998; Bernardi
13. me lexical items were shared by both texts (e.g., export/shipping products, taking action, o
14. re should be minimal formal accountability (e.g. no required summaries or book reports). Ind
15. ies where English is not the main language (e.g., China and India). As an illustrative examp
16. were items which were felt to be too noisy (e.g. headlines).   5. Method   The overall aim w
17.  competent writers. Here corpus technology (e.g., general corpus concordancing) is a promisi
18.  of patterns, extrapolation to other cases (e.g. Scott & Tribble, 2006: 6; Gaskell & Cobb, 2
19.   1995; Louw, 1997) and of translation (see,e.g., Bowker, 1998; Zanettin, 2001). This sectio
20.  only be understood at the discourse level (e.g. Braun 2005; Hughes and McCarthy 1998). Ther
```

Fig. 1 KWIC concordance of, “e.g.”, in a corpus of academic writing

It is in the nature of innovations in computer-assisted language learning (CALL) that early publications tend to be descriptive of classroom practices and software developments. The situation gradually evolves, and Boulton and Cobb (2017) identified over 200 publications attempting some kind of evaluation of corpus use by L2 users; some of the more widely cited are briefly outlined below.

Much of the initial interest lay in emic studies to find out what learners thought about DDL. The data were often gathered through interviews, diaries, or especially questionnaires; the latter are still frequently used but often now as a complement to other aspects. Long-term ecological studies are particularly valuable here, such as Baten et al. (1989) who received overwhelmingly positive feedback from 400 Dutch economics students after 4 months. More recent is the frequently cited paper by Yoon and Hirvela (2004), who introduced corpora to their ESL students in the USA over several weeks. The questionnaires again revealed considerable enthusiasm, especially among those with comparatively low levels of linguistic proficiency, which opens up the question of who DDL is most appropriate for. Across all studies, the response is overwhelmingly positive which no doubt owes something to the novelty factor and the Hawthorne effect, given that most researchers/teachers were themselves enthusiastic. Nonetheless, it seems that DDL can appeal to a wide variety of learners, though Turnbull and Burston (1998) provide a detailed case study of two students needing English for a master’s degree in Australia: one was found to be field independent and took to corpora very quickly; the other was field dependent and found it largely a waste of time.

Another focus has been on the uses learners make of corpora, again mostly by asking learners about their practices; a notable exception is Pérez-Paredes et al. (2012) in Spain

who tracked their learners' searches. They found that lack of training led to fairly unsophisticated queries, with learners approaching corpora in much the same way as they did Internet searches; indeed, the most successful outcomes were found to be combinations of corpus and web searches. The types of queries formulated are analyzed by Kennedy and Miceli (2010), who usefully distinguish pattern hunting (i.e., search for inspiration) and pattern defining (i.e., checking specific questions); success was linked to trial and error, among other things. Charles (2014) had her graduate students compile their own discipline-specific corpora to help with academic writing and followed up use a year later. Eighty six percent of the respondents continued to use corpora at least occasionally in drafting or revising their academic writing and 38% of them regularly. The overall picture that emerges is that most students can use concordancers directly, though it remains controversial how much training is needed. Where time or resources are limited, or for students with lower levels of L2 proficiency, linguistic sophistication, or motivation, work with printed data can provide one solution (e.g., Boulton 2010).

Others have attempted to see whether DDL leads to measurable outcomes from a more etic perspective, i.e., whether it "works" or not. These again split into two groups, the first evaluating the use of corpora as a learning aid, focusing on learning outcomes usually of specified target items. The results generally derive from some kind of language test, whether pre/post or control/experimental designs. Among the earliest and most ambitious, Cobb's (1997) PhD thesis and papers derived from it showed that lower-level Arabic students were able to learn large numbers of words using DDL over a long period of time and were significantly more likely to retain them long term than control groups with word lists and dictionaries only. Most other studies come to similar conclusions for vocabulary and lexicogrammar in general, which may be what DDL is most suitable for, whatever the level of the learners (Lee and Liou 2003). Much of the work here is relatively ecological, being based on a regular course over several weeks or a semester. Chujo and Oghigian and their colleagues in Japan run a semester-long DDL course on a regular basis enabling different types of data collection and analysis, especially as they tweak the course each time. In a 2012 paper looking at noun and verb phrases over two semesters, the experimental group made significant gains in most areas compared to the control group; the results are found to be particularly promising when printouts and hands-on concordancing are combined.

The other group of studies interested in outcomes looks at the impact of corpus use not as a learning aid but as a reference resource, especially while writing (drafting or revising texts or translations). Some of it is short-term experimental work such as by Frankenberg-Garcia (2014), who provided her Portuguese high-school learners of English with dictionary definitions and multiple concordance lines. Both were found to be useful for comprehension, but as few as three carefully-chosen corpus examples proved significantly more effective in production. O'Sullivan and Chambers (2006) got their Irish students of French to correct their own essays; following training, they successfully corrected many underlined errors of grammar and lexis in particular, as well as syntax and even formal things such as spelling where dictionaries or other resources would have been quicker and just as effective. Geluso (2013) also had his learners produce essays especially for the study, but then got them to use Google frequencies as a test of formulaicity for sequences in

inverted commas which the students themselves chose as dubious. Four native English speakers rated the results as being significantly more "natural." Search engines were also used by Todd (2001), but here with the snippets as an equivalent to concordance lines to help correct errors; again, the results suggest that learners can indeed make significant use of such self-selected data.

Work in Progress

The essential ingredients in DDL are corpora and the software to query them, and users today have access to tremendous numbers of both. More and larger corpora can be compiled quickly and easily and distributed free or at small charge via the Internet for many different languages: SketchEngine alone currently lists over 50 languages, some with many different corpora (www.sketchengine.co.uk). However, the prevalence of (semi)automatic compilation aids means that few corpora are as rigorously compiled as the BNC, for example, and care inevitably needs to be taken in interpreting the results. Some tools such as BootCaT (bootcat.sslmit.unibo.it) are publicly available and mean that ordinary users can compile rough-and-ready corpora in a few minutes for specific purposes: all that is needed is to input a handful of "seed" words which are characteristic of the type of language required; the tool does the rest. The availability of large quantities of text via the Internet also means that teachers or learners can manually identify and download texts to build their own corpora for local use. These are often far smaller, which can be an advantage when the needs are highly specific. Software development has also led to increasing numbers of query tools often freely available on the web or for download, which again helps to make DDL much more accessible. Some of these are highly specific, some are intended for experienced researchers; others though are extremely simple and sufficiently general for ordinary L2 learners to be able to work with. AntConc (www.laurenceanthony.net/software/antconc) deserves a special mention here as it has been among the most widely used in recent DDL studies, including some of those mentioned above.

Technological advances have made DDL faster, simpler, more intuitive, prettier, more accessible, and so on. But in terms of methodologies, the essential aspects of DDL remain largely unchanged, typically featuring induction from multiple occurrences in context, augmented with lists and charts of frequencies, collocates, wordsketches, etc. This means that much of the research has been in piloting specific corpora or software, or in testing the basic approach with different learners in different contexts with different needs and questions in mind – all the while doing quite similar things. The advantage of this is that enough evidence has accumulated to be able to take stock. Boulton and Cobb (2017) have undertaken the first systematic meta-analysis of DDL with 88 unique samples from 64 separate studies. The results show large effect sizes overall, both within and between groups. Moderator analyses reveal gaps in the research agenda, including for languages other than English, spoken skills, long-term uptake and occupational uses, etc.

Three trends in recent years are of particular note. First, a number of studies apply essentially DDL-like practices to the web as corpus. The value here is that the web itself is large and varied enough to contain almost anything the user might want; the challenge of course lies in finding it using regular search engines as surrogate concordancers (cf. Boulton 2015). At the same time, users are already familiar with the web and with search engines,

which may go some way toward countering objections of technical difficulties, and further training in their use is more likely to be taken up long term precisely because the tools are so general purpose. A second way to help integrate DDL into learning is to graft them into CALL packages. Cobb's Compleat Lexical Tutor (www.lextutor.ca) provides a number of tools in addition to regular concordancing, allowing learners or teachers to create gap-fills automatically from multiple concordance lines, to visualize the frequency bands of words in a particular text, to click on a word in their own text for a concordance to pop up, to consult and share concordances during writing or error-correction, among other things. This is a way of bringing DDL to the learners rather than expecting them to come to corpus linguistics. Finally, the traditional interest in lexicogrammar is being complemented by more work at the level of discourse, especially using corpora as a reference resource for academic writing. This is not necessarily obvious, since many features of interest may be difficult to search for at surface level; having the students build their own small, specialized corpora increases ownership and familiarity and is one way forward suggested by Charles (2014).

Problems and Difficulties

The advantages of DDL notwithstanding, the fact that it is not mainstream practice suggests that there are difficulties involved. Various questionnaires have noted problems from the learner's perspective, but many of these have dissipated over time, and solutions exist for others. Despite copyright issues and questions of ownership, lack of access to appropriate data is far less a problem today with the increasing availability of large numbers of corpora, as well as the Internet itself. Technical problems can be eliminated if the teacher prepares printed handouts for activities, and software and interfaces have become far more user-friendly in recent years. The ubiquity of Internet search engines have gone a long way towards bridging the gap between everyday practice and DDL: the concept of data searching is familiar and the techniques are largely transferable; users are able to read concordance lines nonlinearly just as they are Google snippets and are less concerned with "drowning in data." Some learners may find the language in corpora difficult: smaller, more relevant corpora may make them more approachable, especially where learners are involved in choosing familiar texts. At lower levels of proficiency, learners may be more comfortable with parallel corpora of translations (see below) or even with corpora of simplified texts or graded readers (available for English on www.lextutor.ca/conc/eng).

Perhaps the biggest problem lies in simply knowing what to query in the first place: much work with error-correction, for example, relies on teachers indicating problem areas (e.g., O'Sullivan and Chambers 2006). One possibility is to rely on frequency data from the web as an indication, focusing on rare items except where they include proper names or highly technical items, as suggested by Geluso (2013). To the extent that DDL enhances language awareness, increased practice is likely to make this easier and more intuitive over time. There is still the problem of formulating the question as a query that the software can understand, and then interpreting the results. Training is recommended by many just to get the most out of Internet search engines, and more may be required for dedicated concordancers and other corpus software. How much training is needed for hands-on concordancing is a controversial issue, though it will ultimately depend on the learners' own needs and preferences, and how much they are likely to want to use corpora in the future.

This raises the further question of the types of learners that DDL is likely to suit best, given that there is considerable variation in their appreciation of the approach and the benefits they derive from it. By far the majority of studies to date have focused on university students, though there is no intrinsic reason why younger learners cannot also benefit. On the other hand, there has been considerable work with learners at lower-intermediate level who are majoring in disciplines other than language, suggesting that language proficiency and sophistication may not be insurmountable barriers. It may even be that DDL is more appropriate in such cases for learners whose previous experience with more teacher-centered, deductive approaches has left them uninterested or struggling (cf. Yoon and Hirvela 2004). All that can really be said at the moment is that further work is needed in a number of areas – which leads us to the final section.

Future Directions

Empirical DDL research has largely focused on university students with intermediate to advanced levels of English as a foreign or second language. It may be that this is where it is most useful and appropriate, though for a more comprehensive picture we would expect more work with younger learners, in secondary or even primary schooling, in private language schools, and outside formal education. This last point seems particularly important: if corpus consultation is argued to be useful for real needs, then we know to know what it can bring to professional situations. Interest in long-term uptake of DDL is at present limited (though see Charles 2014), and introducing it to the workplace seems to be nonexistent except for academic writing.

As far as the corpora themselves are concerned, English is likely to remain the major preoccupation for the foreseeable future, but we would expect more work on other languages too. More important, concordancers work only with written text (including transcriptions); since many learners are primarily interested in spoken language, we would expect the next few years to see development of aligned corpora with sound and even video. It is extremely time consuming to collect spoken data, and the few that currently exist tend to consist largely of interviews (e.g., www.uni-tuebingen.de/elisa/html/elisa_index.html or www.um.es/sacodeyl). An obvious bypath would be to use existing subtitled documents which are already aligned, albeit imperfectly: Aston (2015) describes such uses of the TED talks using WordSmith Tools; Quaglio (2009), among others, has shown that scripted dialogues are closer to "authentic" spontaneous conversation than might be thought, and thus also have their place in a spoken program of DDL for general language learning purposes.

Parallel corpora of translated texts may also be further developed: at the moment, they are relatively rare outside specialist translation courses, despite their obvious uses in many areas, as well as for learners at lower levels of proficiency. There are currently very few that are freely available and easy to use, and they often have their limits: of note is EuroParl, the proceedings of the European Parliament in 21 languages (www.statmt.org/europarl). While the status of Linguee (www.linguee.com) as a parallel corpus may be debatable, it can be used in ways compatible with DDL but with more than one language. Other initiatives can be expected as it becomes easier to align translations for analysis with free parallel concordancers (e.g., www.laurenceanthony.net/software/antpconc).

Technological advances have helped to bring DDL closer to its potential users, with numerous corpora and software designed with L2 learners in mind. At the same time, as technology and the Internet in particular become second nature, learners are already involved in everyday practices that bring them closer to DDL. Johns was originally determined to present DDL as radically different to traditional teaching; the time may have come for it to be seen as an extension of ordinary practice. It will be interesting to see if and to what extent web searches and DDL merge. Finally on the technological front, smart phones and other mobile devices may also bring about substantial changes, but interfaces will need to adapt to allow for screen size and processing speed in particular; entirely new practices may emerge. For the most part, the basic shape of DDL was formed quite early on: recent studies can in many cases be considered replications of earlier work.

DDL is in line with a number of theories of language, learning, and use, some of which derive from insights gleaned from corpus linguistics, but this is largely a one-way relationship. The future may usefully see more empirical studies explicitly designed to analyze the theoretical foundations in more detail. Among other things, we know that language consists of regular overlapping sequences in the form of chunks that are processed, stored, and retrieved as wholes rather than being constructed bottom-up from grammar "rules" as traditionally thought, meaning that any individual item is typically found in a limited number of contexts (Sinclair 1991, on the idiom principle; Hoey 2005, on lexical priming; Millar 2011, on psycholinguistic evidence for chunking). This breaks down the grammar/lexis divide suggesting that our language knowledge is the sum of the encounters we have with it, both receptively and productively, in line with emergentist, usage-based theories (Tomasello 2005). Taylor (2012) talks by analogy of the "mental corpus," highlighting that many of these theories not only support corpus linguistics and DDL but owe much of their origins to them.

Finally, new research practices are needed to test the real benefits of DDL – not just for learning specific items but in helping users to become better language learners, more sensitive to language as a whole. This is the central claim, but so far the only evidence is incidental and at best suggestive (Johns et al. 2008; Allan 2006). What is needed are careful longitudinal studies that specifically focus on this. Ideally, for any technology or approach to become really useful, it needs to be taken up outside the context of a single course – with teachers of other and subsequent courses, and after the end of the instruction period.

References

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

Allan, R. (2006). *Data-driven learning and vocabulary: Investigating the use of concordances with advanced learners of English*, *Centre for Language and Communication Studies Occasional Paper* (Vol. 66). Dublin: Trinity College Dublin.

Aston, G. (2015). Learning phraseology from speech corpora. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 65–84). Amsterdam: John Benjamins.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Baroni, M., & Bernardini, S. (Eds.). (2006). *Wacky! Working papers on the web as corpus*. Bologna: Gedit.

Baten, L., Cornu, A.-M., & Engels, L. (1989). The use of concordances in vocabulary acquisition. In C. Laurent & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 452–467). Clevedon: Multilingual Matters.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534–572.

Boulton, A. (2015). Applying data-driven learning to the web. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 267–295). Amsterdam: John Benjamins.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2).

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes, 35*(1), 30–40.

Chujo, K., & Oghigian, K. (2012). DDL for EFL beginners: A report on student gains and views on paper-based concordancing and the role of L1. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora* (pp. 170–183). Brno: Masaryk University Press.

Cobb, T. (1997). *From concord to lexicon: Development and test of a corpus-based lexical tutor*. Unpublished PhD thesis. Montreal: Concordia University.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–188.

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL, 26*(2), 128–146.

Geluso, J. (2013). Phraseology and frequency of occurrence on the web: Native speakers' perceptions of Google-informed second language writing. *Computer Assisted Language Learning, 26*(2), 144–157.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Johns, T., & King, P. (Eds.). (1991). *Classroom concordancing*, *English Language Research Journal* (Vol. 4). Birmingham: Centre for English Language Studies, University of Birmingham.

Johns, T., Lee, H., & Wang, L. (2008). Integrating corpus-based CALL programs and teaching English through children's literature. *Computer Assisted Language Learning, 21*(5), 483–506.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology, 14*(1), 28–44.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lee, C.-Y., & Liou, H.-C. (2003). A study of using web concordancing for English vocabulary learning in a Taiwanese high school context. *English Teaching and Learning, 27*(3), 35–56.

McEnery, T., & Wilson, A. (1997). Teaching and language corpora. *ReCALL, 9*(1), 5–14.

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

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics, 32*(2), 129–148.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417–528.

O'Sullivan, Í., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing, 15*(1), 49–68.

Pérez-Paredes, P., Sánchez-Tornel, M., & Alcaraz Calero, J. (2012). Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics, 17*(4), 483–515.

Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. natural conversation*. Amsterdam: John Benjamins.

Sinclair, J. (Ed.). (1987). *Looking up: An account of the COBUILD project in lexical computing* (pp. 104–115). London: Collins.

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

Taylor, J. (2012). *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.

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

Todd, R. (2001). Induction from self-selected concordances and self-correction. *System, 29*(1), 91–102.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard: Harvard University Press.

Turnbull, J., & Burston, J. (1998). Towards independent concordance work for students: Lessons from a case study. *ON-CALL, 12*(2), 10–21.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2. *Journal of Second Language Writing, 13*(4), 257–283.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review, 57*(3), 541–572.

This is a pre-publication version. For the version of record, please see DOI: 10.1007/978-3-319-02328-1_15-1
(or email me at alex.boulton@atilf.fr)