EUROCALL

Nottingham 2011

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

The Call Triangle: student, teacher and institution

What data for data-driven learning?

Alex Boulton[*]

Crapel – ATILF / CNRS, Nancy-Université

Abstract

Corpora have multiple affordances, not least for use by teachers and learners of a foreign language (L2) in what has come to be known as 'data-driven learning' or DDL. The corpus and concordance interface were originally conceived by and for linguists, so other users need to adopt the role of 'language researcher' to make the most of them. Despite the alleged advantages of this, it does create a potential barrier for occasional or non-specialist users in particular. While researchers debate the status of the 'web-as-corpus', the Internet represents a vast bank of data already familiar to most people; less discussed is the status of 'Google-as-concordancer' – another familiar tool. This paper discusses some of the advantages and disadvantages of this approach from a pedagogical perspective.

Keywords: corpus; concordance; data-driven learning; DDL

1. Introduction

Tim Johns first introduced the term 'data-driven learning' (DDL) in 1990 to describe how language learners could become language detectives to explore language data themselves. Much of the early work was conducted in what seem now to be fairly primitive conditions (see e.g. Aston 1996), and exclusively with locally available resources. This invariably meant language corpora (i.e. carefully compiled samples of language intended to be representative of a particular variety) and dedicated concordancers, with Johns himself working with parts of the COBUILD (Bank of English) corpus, or small ESP corpora he compiled for use with his own MicroConcord (Johns, 1986) for just these purposes.

The Internet has brought many corpora and dedicated tools within reach of practising teachers and learners. However, a common criticism is still that many of them require considerable investment in terms of training for learners (and teachers) to understand the rationale as well as how to use them efficiently. Even accepting the potential benefits of a DDL approach, the technology is clearly perceived as a major obstacle to the implementation of DDL in classes around the world (Boulton, 2009).

One might however wonder whether a 'corpus' and 'concordancer' in the original sense are needed at all, or if they can be are replaced by more familiar resources – notably the web as 'corpus' and/or general search engines as 'concordancers'.[1] For the purposes of language teaching and learning, it is surprising that there have been so few attempts to apply DDL techniques to such general purpose data and tools (e.g. Sha, 2010), and so few reports on their actual use (e.g. Todd, 2001). This paper outlines some of the usual objections to this approach, and puts them into perspective based on pedagogical criteria.

[*] Contact author. Tel.: +33 (0)354 505 112
E-mail address: boulton@univ-nancy2.fr

[1] Cf. the similar on-going debate within corpus linguistics itself: *ICAME Debate*, Oslo, 1 June. http://www.hf.uio.no/ilos/english/research/conferences/2011/icame2011/workshops.html#WS2

2. The web as 'corpus'

Textbooks generally define a corpus as a large collection of authentic texts in electronic format designed to be representative of a language variety. But this is not uncontroversial even within the realm of corpus linguistics: there are "several criteria that, if met, define a prototypical corpus, but the criteria are neither all necessary nor jointly sufficient" (Gilquin & Gries, 2009: 6). For some, it is a simple fact that "the World Wide Web is not a corpus" (Sinclair, 2005: 21), while for others "the answer to the question 'Is the web a corpus?' is yes" (Kilgarriff & Grefenstette, 2003: 334). Problems frequently cited against treating the web as corpus relate to its unknown size, ever-changing composition, its hidden pages (e.g. Lüdeling et al., 2007). The web may not be "representative of anything other than itself," as Kilgarriff and Grefenstette (2003: 333) point out – "but then neither are other corpora."

It is not annotated, but nor are many other corpora, and while this limits some types of research is does not mean that no research is possible. And of course it contains considerable 'noise' in the form of reduplications, spam, lists, nonsense pages, and so on, with innumerable different types of texts from widely varying authors writing for different purposes all mixed up. But this is all part of "the mush of general goings-on" of real language in use (cf. Firth, 1957: 187); while carefully compiled corpora attempt to create a more principled and orderly bank of text for linguistic analysis, none of these objections stop linguists using the web as a 'quick and dirty' source of language data for everyday concerns.[2] Further, it increasingly serves as a useful point of comparison even in research papers (Joseph, 2004) – and for good reason. Firstly, "language is never, ever, ever random" (Kilgarriff, 2005), and even with all its noise and other problems, the sheer size of the web means that web data often give results that are close to traditional corpora (e.g. Rohdenburg, 2007), and even to native-speaker judgements (Keller & Lapata, 2003).

The main point here is that if even linguists can overcome qualms about using web data, then it would seem unreasonable to prevent others from using it, especially perhaps language learners who do not need to be as scrupulous in their requirements as researchers: the decision should be *pedagogically* driven rather than based on non-pertinent *research* criteria. Even the sceptical Sinclair recognises that "the web itself… [is a] huge source of language that is available in the classroom or the study at home" (Sinclair, 2004: 297), and that "it is important to avoid perfectionism in corpus building. It is an inexact science…" (Sinclair, 2005: 98) – especially for learners and teachers. While the web may not be a prototypical corpus in terms of linguist research, we can at least treat it as "corpus surrogate" (Bernardini et al., 2006: 10ff) which may be quite fit for purpose.

Its advantages in language teaching include its size, recency, variety (whatever you want is probably there somewhere), availability (free), reliability (the web itself doesn't crash, or impose limits on the number of simultaneous users), speed, flexibility, and so on. Just as importantly, it is already familiar to learners, especially via Internet search engines such as Google.

3. Google as 'concordancer'

Google (or any other general purpose search engine) is designed for information retrieval rather than linguistic research, and is therefore inevitably more limited than a concordancer for this purpose. It does not allow explicitly linguistic search syntax; though there are often ways round its limitations, these can be time-consuming. The presentation of responses is also not linguistically ideal, though the snippets are not entirely dissimilar to concordances. Google is something of a black box, where the user has little idea of how the results are retrieved or ordered and can do little to change this except submit a new query with different parameter settings or search terms (Bergh, 2005). It can also be difficult for learners to interpret the results – how reliable are they, how frequent is frequent 'enough'? And, of course, Googling is simply not a 'serious' pursuit; but again, if linguists can use it at least informally for this purpose, then a fortiori language learners whose requirements are less stringent. No concordancer is ideal (Kaszubski, 2006; Kosem, 2008), and general-purpose search engines may be the least ideal of all.[3] But there seems to be nothing stop us treating "Google as a quick 'n' dirty corpus tool" (Robb, 2003); it may even be that the messiness of web data and limitations of search engines will foster language awareness and critical thinking about language (Milton, 2006).

Google (or another search engine) is likely to be already familiar to learners, with a simple, intuitive interface that does not require vast linguistic or metalinguistic baggage. They may not be using it very

[2] Frequently for example on *LanguageLog*: http://languagelog.ldc.upenn.edu

[3] Many of the postings on Jean Véronis's blog (*Technologies du langage: Actualités, commentaires, réflexions*) highlight the deficiencies of Google in particular, e.g. '5 billion have disappeared overnight'; 'Yahoo's missing pages'; 'Crazy duplicates'; 'Google: Mystery index' and many more. Yet we are still left with 'Google: The largest linguistic corpus of all time'. (http://blog.veronis.fr)

well, but are already getting results, and a little further training is likely to increase their efficiency (cf. Acar et al., 2011). So using Google allows learners to draw on existing knowledge and techniques, and any further training will transferable back to their everyday lives where ICT literacy is an essential skill, not limited exclusively to corpus use. Indeed, there is some evidence that Google is already being used in this way for language teaching and learning (e.g. Clerehan et al., 2003; Conroy, 2010).

Would this constitute data-driven learning? In a way, the question is redundant: all that counts is whether it is beneficial to the learning process; but that is evading the question. DDL is a difficult beast to pin down: it is "not an all-or-nothing affair: its boundaries are fuzzy, and any identifiable cut-off point will necessarily be arbitrary" (Boulton, 2011: 575). It is notable however that although Johns was mainly working with corpora, he chose the term 'data-driven' rather than 'corpus-driven': "the data is primary (Johns, 1991: 3). And the web certainly constitutes language data. He was also largely working with concordancers, "one of the most powerful tools that we can offer the language user" (Johns, 1988: 15), but that was before the Internet and search engines even existed. One can only speculate as to how DDL would have developed had learners had such ease of access to data in the 1980s.

4. Conclusions

The objections here to using the web as 'corpus' and search engine as 'concordancer' have been shown to be largely theoretical, and based on criteria which are of little relevance in language teaching. The main conclusion is pragmatic and practical rather than dogmatic or ideological: if an approach or technique is of benefit to the learners and teachers concerned, it should not be ruled out automatically (Hafner & Candlin, 2007). As so often, there is likely to be a payoff between how much the teachers / learners are prepared to put in (ideally as little as possible) and how much they want to get out (ideally as much as possible). The optimum may be a point on the continuum, or more likely a movement along the continuum – gradually investing more and more, until such a time as the extra benefits do not justify the extra costs. Such a cost-benefit analysis will produce different results for different individuals and groups with different needs and preferences, facilities and constraints.

It seems likely that many learners around the world are already Googling the Internet in ways not entirely dissimilar to DDL, a practice which may be actively encouraged by their teachers while remaining invisible in the DDL research literature. The approach is in many ways attractive, offering as it does a familiar, intuitive, and easy way to begin simple DDL which brings immediate benefits (cf. Shei, 2008). To reach a wider audience, there is also something to be said for encouraging the perception of DDL as *ordinary* practice rather than radical or revolutionary (Boulton, 2010). For those who wish to go further, it provides a handy first step on the road to more 'hard-core' DDL (Conroy, 2010). Other intermediate steps might include using individual websites such as newspapers for DDL-like queries, or web concordancers (such as WebCorp and KWiCFinder[4]) which still use the web-as-corpus approach but provide output which is more linguistically relevant and useful for language learning.

Space does not permit an extensive presentation of how learners (and teachers) can use Google and the web for language learning, nor a detailed theoretical analysis of whether this constitutes DDL; but there are reasons to think that they can and it is, topics which will be the subject of future research.

5. References

Acar, A., Geluso, J., & Shiki, T. (2011). How can search engines improve your writing? *CALL-EJ*, *12*(1), 1-10. http://callej.org/journal/12-1/Acar_2011.pdf

Aston, G. (1996). The British National Corpus as a language learner resource. In S. Botley, J. Glass, A. McEnery & A. Wilson (Eds.), *Proceedings of TALC 1996. UCREL Technical Papers, 9*, 178-191.

Bergh, G. (2005). Min(d)ing English language data on the web: What can Google tell us? *ICAME Journal, 29*, 25-46. http://gandalf.aksis.uib.no/icame/ij29/ij29-page25-46.pdf

Bernardini, S., Baroni, M., & Evert, S. (2006). A WaCky introduction. In M. Baroni & S. Bernardini (Eds.). *Wacky! Working papers on the web as corpus* (pp. 9-40). Bologna: Gedit. http://wackybook.sslmit.unibo.it/

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL, 21*(1), 37-51.

[4] http://www.webcorp.org.uk; http://www.kwicfinder.com

Boulton, A. (2010). Data-driven learning: On paper, in practice. In T. Harris & M. Moreno Jaén (Eds.), *Corpus linguistics in language teaching* (pp. 17-52). Bern: Peter Lang.

Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Goźdź-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 563-580). Frankfurt: Peter Lang.

Clerehan, R., Kett, G., & Gedge, R. (2003). Web-based tools and instruction for developing it students' written communication skills. In *Proceedings of exploring educational technologies*. http://www.monash.edu.au/groups/flt/eet/full_papers/clerehan.pdf

Conroy, M. (2010). Internet tools for language learning: University students taking control of their writing. *Australasian Journal of Educational Technology, 26*(6), 861-882. http://ascilite.org.au/ajet/ajet26/conroy.html

Firth, J. (1957). *Papers in linguistics 1934-1951*. London: Oxford.

Gilquin, G., & Gries, S. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1): 1-26.

Hafner, C., & Candlin, C. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes, 6*(4), 303-318.

Johns, T. (1986). Micro-Concord: A language learner's research tool. *System, 14*(2), 151-162.

Johns, T. (1988). Whence and whither classroom concordancing? In P. Bongaerts, P. de Haan, S. Lobbe & H. Wekker (Eds.), *Computer applications in language learning* (pp. 9-27). Dordrecht: Foris.

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. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing. English Language Research Journal, 4*, 1-16.

Joseph, B. (2004). The editor's department: On change in Language and change in language. *Language, 80*(3), 381-383. http://www.ling.ohio-state.edu/~bjoseph/publications/2004EDchange.pdf

Kaszubski, P. (2006). Web-based concordancing and ESAP writing. *Poznan Studies in Contemporary Linguistics, 41*, 161-193.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics, 29*(3), 459-484.

Kilgarriff, A. (2005). Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263-275. http://kilgarriff.co.uk/Publications/2005-K-lineer.pdf

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on web as corpus. *Computational Linguistics, 29*(3), 333-347.

Kosem, I. (2008). User-friendly corpus tools for language teaching and learning. In A. Frankenberg-Garcia (Ed.), *Proceedings of the 8th teaching and language corpora conference* (pp. 183-192). Lisbon: ISLA-Lisboa.

Lüdeling, A., Baroni, M., & Evert, S. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 7-24). Amsterdam: Rodopi.

Milton, J. (2006). Resource-rich web-based feedback: Helping learners become independent writers. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 123-137). Cambridge: Cambridge University Press.

Robb, T. (2003). Google as a quick ‘n’ dirty corpus tool. *TESL-EJ, 7*(2). http://www.tesl-ej.org/wordpress/issues/volume7/ej26/ej26int/

Rohdenburg, G. (2007). Determinants of grammatical variation in English and the formation / confirmation of linguistic hypotheses by means of internet data. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 191-209). Amsterdam: Rodopi.

Sha, G. (2010). Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. *Computer Assisted Language Learning, 23*(5), 377-393.

Shei, C. (2008). Discovering the hidden treasure on the Internet: Using Google to uncover the veil of phraseology. *Computer Assisted Language Learning, 21*(1), 67-85.

Sinclair, J. (2004). New evidence, new priorities, new attitudes. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 271-299). Amsterdam: John Benjamins.

Sinclair, J. (2005). Corpus and text: Basic principles. / Appendix: How to build a corpus. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 5-24 / 95-101). Oxford: Oxbow Books. http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm

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