Wanted : Large corpus, simple software. No timewasters

Alex Boulton

Wanted: Large Corpus, Simple Software. No Timewasters

Alex BOULTON
ATILF, Equipe Crapel
CNRS, University of Lorraine
Nancy, France
alex.boulton@univ-lorraine.fr

Acknowledgement: part of this paper was originally given at EuroCALL 2011 in Nottingham

1. Introduction

Data-driven learning involves the use of dedicated concordancers to explore large language corpora. Or does it? Willis (1998) reported on the use of manual concordancing involving nothing more than texts, learners and a blackboard, a technique which Johns (1993) recognised as DDL, and similar activities can be found described as DDL today (e.g. Tyne in press). The appeal of such activities is of course that they reduce or eliminate the need for technology altogether. This is certainly an advantage, as the technology itself is frequently cited as a major obstacle, and possibly the prime reason why DDL has not become more mainstream practice (e.g., Boulton, 2010). The problem with this approach is that it does not make use of the potential of computers to help in the inductive, discovery-based noticing process. An alternative path is to simplify the technology as much as possible – either the corpus itself or the associated software. Firstly, there has been some reaction against the early corpora (large, general-purpose, linguistically oriented) in favour of small, *ad hoc* corpora of a specific language genre (e.g. Ghadessy et al., 2001), often designed with explicitly pedagogical aims in mind (Braun, 2007). Similarly, great strides have been made in making concordancers more user-friendly, but technology still remains a difficulty cited in many contemporary papers (e.g. Rodgers et al., 2011). What is it about the concepts of 'corpus' and 'concordancer' that keeps them out of mainstream classroom practice?

A third way is to begin not with corpus or concordancer, but with the learners and what they already do – in other words, to bring our work closer to them rather than making them come to us. In what ordinary, everyday activities are learners involved in using computers to search for information outside the language classroom? Most obviously, in browsing the Internet. A small number of studies attempt to show how this can be used for language learning (e.g. Todd, 2001; Chinnery, 2008; Sha, 2010); the temptation then is to wonder whether the Internet might not serve as a substitute corpus, and Google as a substitute concordancer in a more-or-less DDL approach. This might seem inappropriately iconoclastic at a TaLC conference, but DDL and corpus linguistics have broken more than one ideological barrier themselves. If these resources had been available in the 1980s, we would likely be working in a very different field today.

This paper sets out to see whether such an approach can be considered a form of DDL, at least in abstract terms, then goes on to test the potential and limits of the Internet and Google for linguistic searches relevant to language learners. In conclusion, it is argued that the processes are in fact not entirely dissimilar, and that they can provide one way in to DDL

– immediately useful for all, and potentially leading some on to more prototypical DDL activities with corpus and concordancer.

2. The web as 'corpus'

A textbook definition of a corpus outlines a large collection of authentic texts in electronic format designed to be representative of a language variety. But this is not always as straightforward as it might sound, even within the field of corpus linguistics itself, where 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). Clearly, the Internet fails the textbook definition on several counts, but then so would many other corpora. 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." Its size and composition are unknown (e.g. Lüdeling et al., 2007) and probably unknowable in any meaningful sense, but the end-user may not be aware of what texts exactly are in, say, on-line versions of many accepted corpora, and can only take the providers' word for it. It is also unstable insofar as it fluctuates over time, but so also do monitor 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. Finally, and perhaps most unsettlingly, it is extremely "noisy" with its endless 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 the same can be said (if to a lesser degree) of many semi-automated corpora compiled from the web, from CoCA to the WaCKy corpora. Besides, this is arguably all part of "the mush of general goings-on" of real language in use (Firth, 1957: 187).

Ultimately, the issue of whether the web is (or can be used as) a corpus is a subjective one. For Sinclair (2005: 21), "the World Wide Web is not a corpus"; for Kilgarriff and Grefenstette (2003: 334), "the answer to the question 'Is the web a corpus?' is yes". The 'web-as-corpus' is not perfect, but nor are any other corpora; indeed, "it is important to avoid perfectionism in corpus building. It is an inexact science…" (Sinclair, 2005: 98). None of the objections outlined here stop linguists using the web as a 'quick and dirty' source of language data for everyday concerns; as Lou Burnard recently remarked, "as a non-native speaker of French, if I meet a word I don't know, I Google it to see how it's used…". In more rigorous research, the web 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).

In the end, if even (some) linguists can overcome qualms about using web data for everyday concerns as well as serious, rigorous research purposes, then it would seem unreasonable to prevent 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 *research* criteria which are of little relevance to them. 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). 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: 10) 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'

General purpose search engines such as Google are designed for information retrieval rather than linguistic research, and are therefore inevitably more limited than a concordancer for this purpose. It does not allow explicitly linguistic search syntax, and ways round its limitations 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.[1] 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 well, but are already getting results, and a little further training is likely to increase their efficiency (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

[1] 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 the 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)

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

I have argued here that the objections to using the web as 'corpus' and search engine as 'concordancer' are largely theoretical, and based on criteria which are of little relevance in language teaching and learning. 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 will be at some variable point in between, 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 (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) 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.

References

Acar, A., Geluso, J. & Shiki, T. (2011). 'How can search engines improve your writing?'. *CALL-EJ*, 12(1), Avaiable from: http://callej.org/journal/12-1/Acar_2011.pdf. 1-10

Aston, G. (1996). 'The British National Corpus as a language learner resource.' In: Botley, S., Glass, J., McEnery, A. & Wilson, A. (eds.). (1996). *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. Avaiable from: http://gandalf.aksis.uib.no/icame/ij29/ij29-page25-46.pdf. 25-46.

Bernardini, S., Baroni, M. & Evert, S. (2006). 'A WaCky introduction.' In: Baroni, M. & Bernardini, S. (eds.). *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit. Avaiable from: http://wackybook.sslmit.unibo.it/. 9-40.

Boulton, A. (2009). 'Testing the limits of data-driven learning: Language proficiency and training'. *ReCALL*, 21(1). 37-51.
Boulton, A. (2010), 'Data-driven learning: Taking the computer out of the equation'. *Language Learning*, 60(3). 534-572.
Boulton, A. (2010). 'Data-driven learning: On paper, in practice'. In: Harris, T. & Moreno Jaén, M. (eds.) (2010). *Corpus linguistics in language teaching*. Bern: Peter Lang. 17-52.
Boulton, A. (2011). 'Data-driven learning: The perpetual enigma'. In: Goźdź-Roszkowski, S. (ed.). *Explorations across languages and corpora.* Frankfurt: Peter Lang. 563-580.
Braun, S. (2007). 'Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora'. *ReCALL*, 19(3). 307-328.
Chinnery, G. (2008). 'You've got some GALL: Google-assisted language learning'. *Language Learning & Technology*, 12(1). 3-11.
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.* Avaiable from: 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). Avaiable from: http://ascilite.org.au/ajet/ajet26/conroy.html. 861-882.
Firth, J. (1957). *Papers in linguistics 1934-1951*. London: Oxford.
Ghadessy, M., Henry, A. & Roseberry, R. (eds.). *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.
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: Bongaerts, P., de Haan, P., Lobbe, S. & Wekker, H. (eds.). *Computer applications in language learning*. Dordrecht: Foris. 9-27.
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: Johns, T. & King, P. (eds.). *Classroom concordancing. English Language Research Journal*, 4. 1-16.
Johns, T. (1993). 'Data-driven learning: An update'. *TELL&CALL* 2. 4-10.
Joseph, B. (2004). 'The editor's department: On change in Language and change in language'. *Language*, 80(3). Avaiable from: http://www.ling.ohio-state.edu/~bjoseph/ publications/2004EDchange.pdf. 381-383.
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). Avaialbe from: http://kilgarriff.co.uk/Publications/2005-K-lineer.pdf. 263-275.
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: Frankenberg-Garcia, A. (ed.). *Proceedings of the 8th Teaching and Language Corpora Conference* Lisbon: ISLA-Lisboa.183-192.

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

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

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

Rodgers, O., Chambers, A. & LeBaron, F. (2011). 'Corpora in the LSP classroom: A learner-centred corpus of French for biotechnologists'. *International Journal of Corpus Linguistics*, 16(3). 392-358.

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

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: Sinclair, J. (ed.). *How to use corpora in language teaching*. Amsterdam: John Benjamins. 271-299.

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

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

Tyne, H. (in press). 'Corpus work with ordinary teachers: Data-driven learning activities'. In: Thomas, J. & Boulton, A. (eds.). *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.

Willis, J. (1998). 'Concordances in the classroom without a computer'. In: Tomlinson, B. (ed.). *Materials Development in Language Teaching*. Cambridge: Cambridge University Press. 44-66.