Corpus-based study of language and teacher education

Alex Boulton, Henry Tyne

Alex Boulton & Henry Tyne. (2014). Corpus-based study of language and teacher education. In M. Bigelow & J. Ennser-Kananen (eds), *The Routledge Handbook of Educational Linguistics*. New York: Routledge, p. 301-312.

Historical Perspectives

Linguistic data can be analyzed in a number ways, ranging from the description of isolated or invented examples to the study of collected samples or datasets. The analysis of collections of linguistic data is generally associated with modern techniques, though there are examples of work as early as the Middle Ages with concordances of important texts (notably the Bible) to highlight how words are used in context. Later, we see various lexicographical projects during the Renaissance, and, in the late nineteenth century, the work of dialectologists who sought to gather and record spoken data in particular. This work was painstaking in the extreme, and computers made an early contribution to making the various processes involved vastly simpler, faster and more rigorous, resulting in the creation of the landmark Brown corpus in the 1950s (1 million words of written data), the Bank of English for COBUILD projects in the 1980s (evolving from 7 million words, written and spoken), the British National Corpus in the 1990s (100 million words, written data with 10% carefully collected spoken samples) and, in the field of dialectology and sociolinguistics, projects such as ALAVAL or PFC, containing only spoken data. Further evolution has led to the semi-automatic collection of texts via the Web for even larger corpora in the new century: the Corpus of Contemporary American English (425 million words, written plus transcripts of unscripted speech), the WaCKy corpora (2 billion words on line), GoogleBooks (500 billion words of scanned text). Meanwhile, further work has been conducted with new types of corpora: parallel corpora in translation, sound-aligned spoken corpora, multimedia corpora with video, and learner corpora of non-native speaker output. As corpora have become bigger, so the software has come to play a more important role, providing simpler, faster, more powerful options, without which the sheer quantity of data could quite simply not be accessed as a corpus. Even 15 years ago, downloading 50 concordance lines could take several minutes, and for common items Aston (1996: 179) advised users to “go and have lunch while you’re waiting.” Today thousands of texts can be processed virtually instantaneously on line (e.g. BYU corpora) for general needs, ad hoc corpora of tens of thousands of words can be generated in minutes (e.g. WebBootCat) for one-off needs, and increasingly user-friendly software can be downloaded free from the web (e.g. AntConc) as smaller corpora find increasing uses for specific local needs.

The implications of corpus-based study in language teaching are manifold and can be seen today largely in the form of work upstream in informing dictionaries and other reference resources, from the COBUILD resources to the present day. Indeed, it is barely conceivable today to undertake new work in the provision of mainstream language teaching materials which is not corpus-based in some way. Corpora have also been used in devising syllabuses derived from evidence of frequency of actual usage in context from both native and learner corpora (e.g. the English Profile project at Cambridge University). While course books and teaching materials have been slower to make full use of corpus data, a number of major new textbooks (for English at any rate) do take up the challenge (e.g. Touchstone; McCarthy et al. 2005). Such uses of corpora remain largely invisible to the end users, but teachers in university environments have been quick to see the potential for themselves and their learners, notably in the work of Tim Johns at Birmingham University who promoted the term

"data-driven learning" (Johns & King 1991). Johns himself was working on DDL throughout the late 1980s, but McEnery and Wilson (1997: 2) credit even earlier work to Peter Roe in Aston University (also in Birmingham); the first published paper in the area appeared in 1980 by McKay in San Francisco. The biannual Teaching and Language Corpora conferences in Europe are testament to the continued research interest in the field, as are the number of books and articles in the area. Today, corpora are widely used in university environments, especially in translation studies and for learners needing languages for specific or academic purposes, and even for younger learners in school in a number of individual initiatives.

Core Issues and Key Findings

Allowing learners access to language corpora is alleged to have numerous advantages. Firstly, in terms of the language itself (i.e. as input for learners), there is a clear benefit in terms of both the quality and the quantity of data: the use of authentic materials for language teaching has been defended ever since the 1970s-1980s, in particular with the shift to communicative methods, and, as Chambers (2009) has pointed out, the possibility of using of corpora as authentic spoken materials came as a refreshing change in language teaching. Also, we see how authenticity through the quantitative approach can be addressed, i.e. the fact that specific forms or structures only occur in certain genres or discourse styles can be highlighted (see the work of Biber for English). In this way, the use of corpora has dramatically improved our knowledge of how language is actually used by different people in different contexts, genres, text types, and so on, the findings being used to inform all kinds of pedagogical materials which previously tended to be based on the author's intuitions or ad hoc collections of samples amassed over the years.

With regard to the question of learner access to language, numerous studies have demonstrated that the types of methods typically employed by corpus linguists and sociolinguists can be targeted for pedagogical purposes: they may include fieldwork, data-collection and other 'hands-on' activities such as transcription, or they may involve using a search engine or web-crawler to gather online data; they may involve using a concordancer to isolate forms or regular expressions, or simply to see how language behaves in a patterned manner. What is common to these methods is the idea that target language data is not presented to the learner as 'text'. This is important for teaching since many traditional learning materials and activities use texts, i.e. complete language samples. Here, however, learners may be attending to isolated details or forms (e.g. through transcription) or they may be learning from concordance lines, i.e. isolated samples of language extracted from their textual environments. Finally, it should be pointed out that engaging learners with data presupposes a host of attitudes and learning activities which focus on the principles of learning by doing and learning by discovery, a heuristic and often inductive process culminating in the inductive approach of DDL.

Research Approaches

Over 100 empirical studies to date have attempted to evaluate some aspect of corpus use: learner behavior (what they do with corpora), receptivity (their attitudes towards corpus use), success using corpora as a learning tool or as a reference resource (see Boulton 2010 on-line update for an overview). These studies are in the main relatively small-scale and qualitative, with an emphasis on questionnaires for collecting learner reactions to corpus

use. Qualitative studies make it difficult to know how well the type of research is likely to transfer to other contexts; while quantitative ones have the advantage of ironing out individual differences, they thereby underestimate the individualization aspect as learners can each explore very different things.

There is however some quantitative analysis, especially for learning outcomes, the results of which are generally encouraging if not necessarily statistically significant in the majority of cases for reasons well known in dynamic systems theory or complex adaptive systems research: overemphasis on single variables is unlikely to produce convincing results in any area of FLT. From this perspective it is essential to avoid attributing too much importance to any individual study; the 'bigger picture' overview of dozens of studies is more representative, and here is seen as extremely positive.

Many of these papers cover similar populations and contexts for similar questions: English L2 (but some other languages including different scripts, e.g. Chinese); university contexts (but a few in secondary schools or elsewhere); relatively motivated, sophisticated learners at higher levels of proficiency (but some others); etc. This leaves open a number of new areas for greater exploration, as discussed in the following section.

New Debates

As we have seen, there has been considerable empirical research in corpus use for language learning and teaching, despite claims to the contrary. However, the majority of these have similar focus – questionnaires to collect learners' representations of using corpora or tests of short-term learning outcomes on specific language points for university students needing English. There is certainly a dearth of empirical research in other areas, notably:

- New contexts: How does DDL work in other contexts – in secondary and even primary education, with adults in continuing education, in private language schools or self-access centers? Given that the overwhelming majority of corpus-based studies are concerned with English and ELT, is it possible to teach languages other than English using corpus-based techniques (for technical, cultural, linguistic reasons); can corpus-based teaching be used by teachers of languages where teaching materials are lacking or considered to be unsatisfactory, etc.?
- Longitudinal studies: How does DDL fare in the long term, not just on specific language items covered but on language as a whole? What use is made of corpora outside class or after the end of the course? Do learners apply it to other areas of study in addition to language learning, such as in information retrieval, literary or cultural studies, translation, etc.? What evidence is there that DDL really does promote autonomy, life-long learning, motivation, language awareness, metacognitive skills, learner-centeredness, etc. – all things commonly claimed in the literature but largely unexamined? How would it be possible even to go about researching these?
- Tools and technology: For the purposes of DDL, can the web be considered a surrogate 'corpus', and search engines even as surrogate 'concordancers'? Can Internet searches for language purposes be seen as related to DDL, as a way in to DDL, or are they totally different – and if DDL can build on Internet search techniques, does it have a washback effect on this in promoting ICT literacy in general? What are

the relative advantages of different types of corpora (written, spoken, multimodal, learner corpora, translation corpora, large / small, generic / LSP corpora, etc); are they best accessed via 'linguistic' tools (concordancers, even ones designed for pedagogical purposes), or via more comprehensive software packages (eg the Hypertext package in LexTutor); and so on.

- Practice: Is DDL best presented as a radical technique, or as 'normal' practice building on existing regular classroom activities? Are corpora best seen as a learning aid per se, or as a reference resource (especially for writing and translation) alongside dictionaries and other tools? Are there any limitations to the effectiveness of DDL in different contexts, for different purposes, for different learner profiles? How can teachers be encouraged to explore DDL for themselves and for their learners in both pre-service and in-service training?

Implications for Education

Corpora have still to make substantial inroads in mainstream teaching. It has been argued that the inductive processes involved in exploring authentic language are too demanding for younger or less sophisticated users, an argument undermined by research with primary school students for L1 education (e.g. Sealey & Thompson 2004), and by work with university students with lower levels of L2 language ability (e.g. Chujo et al 2009). A more convincing explanation is perhaps that teachers lack awareness of corpora and training in their use (cf. Conrad 2000; Cobb 2009). With corpora absent from teacher training qualifications, they are inevitably underexploited in initial teacher education. Mukherjee (2004) shows how practicing teachers are receptive to the benefits of corpora for their own use, but are slower to recognize applications for their learners.

Corpora by no means contain the solutions to all teaching problems, and massive use with any group of learners is likely to be counter-productive. However, the evidence suggests that, sensitively used, they can provide an additional set of tools and techniques for a variety of purposes for at least some learners in some contexts, can increase language awareness and metacognitive skills, building on and promoting existing ICT skills, are potentially highly motivating as they allow exploration of individual questions and are thus learner-centered, fostering autonomy with potential for life-long learning.

Additional References

2-3 most important historical books and articles

- Sinclair, John McHardy (ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins, p. 104-115.
  Sinclair is often called the father of modern corpus linguistics, and had a strong interest in pedagogical issues; this volume covers various aspects of the issues involved in compiling, annotation and exploiting the revolutionary COBUILD Bank Of English corpus.
- Johns, Tim & Philip King (eds.) 1991. *Classroom Concordancing. English Language Research Journal*, 4.
  Similarly, Johns can be called the founding father of DDL, putting corpora in the hands of learners; he had already published several papers in the field, but this collected volume is considered seminal.

- Tribble, Chris & Glyn Jones. [1990] 1997. *Concordances in the Classroom*. 2nd edition. Houston: Athelstan.
  The first attempt to show in concrete terms how corpora can actually be used for teaching / learning.

2-3 most important recent books and articles

- O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
  A comprehensive look at uses of language corpora with pedagogical aims in mind, though with little on DDL itself.
- Reppen, Randi. 2010. *Using Corpora in the Classroom*. Cambridge: Cambridge University Press.
  One of several recent books intended to bring corpora to a wider audience, including teachers.
- O'Keeffe, A. & M. McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
  An indispensable introductory set of readings for all aspects of corpus linguistics, especially here sections 5 and 6: "Using a corpus for language pedagogy and methodology" and "Designing corpus-based materials for the language classroom".
- Thomas, James & Alex Boulton (eds). Forthcoming 2012. *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.
  The latest volume of collected papers from the biennial *Teaching and Language Corpora* conferences, founded in 1994.