But where's the proof? The need for empirical evidence for data-driven learning.

Alex Boulton

But Where's the Proof? The need for empirical evidence for data-driven learning

Alex Boulton
CRAPEL-ATILF/CNRS, Nancy Université
alex.boulton@univ-nancy2.fr

Corpus studies of native speaker and learner language have been influential in informing syllabus design and course content in foreign or second language (L2) teaching. Corpora can also be explored directly by the learner, in what Johns (1991) has called "data-driven learning" (DDL). Such an approach is alleged to have many advantages, including fostering learner autonomy, increasing language awareness and noticing skills, and improving ability to deal with authentic language. Although such theoretical arguments may seem convincing, their power is mitigated by the fact that DDL has yet to filter down into mainstream teaching and learning practices. In other words, it may be fine in theory, but what about in everyday practice? Empirical evidence in support of the theory would seem essential.

Survey

In the search for such evidence, we looked back at several hundred papers linking corpora and L2 teaching / learning since the appearance of the seminal collection of papers by Johns and King in 1991. The sheer volume of publications attests to the interest in the field, at least at the research level. Among these papers, we found 39 studies which report some kind of evaluation of DDL beyond the researcher's opinion. These are analysed in this paper, not as a traditional overview of findings but rather in an attempt to sort out the main strands of the types of research conducted to date. The survey is of necessity succinct in the extreme; readers are referred to the original studies listed in the bibliography.

Background

The studies were conducted in 19 different countries, attesting to a variety of different contexts (although 24 of the studies were in Europe). Most used native speakers of the countries they were in, with occasional non-native learners (e.g. exchange students); only five (all in English-speaking countries) drew on foreign students of mixed nationalities. The L2 was English in 34 of these studies, and a European language in the other five. Obvious reasons include the current level of demand for English, but also the greater availability of tools and corpora, not to mention awareness of corpus linguistics and DDL. There is certainly room here for broadening the scope to other target languages.

Learners

Of the 39 studies, only two focused on younger learners, while 36 were conducted with students in higher education, including eight with postgraduates. Of these, 18 studies involved participants who may reasonably be regarded as "sophisticated": in 15 cases they were enrolled for a degree course in the L2, or in a translation degree including the L2; a further three studies involved linguistics students. The other 18 studies focused on students needing English for academic or specific purposes. It is thus apparent that the majority of the studies are concerned with more or less advanced learners. Only two claim "low" levels and two "beginners", although careful reading casts some doubt on this.

Corpora and software

A wide variety of corpora were used in these studies, from the very large to only 2000 tokens. Some translation-based studies used parallel or comparable corpora, but most were monolingual. Published corpora included the BoE, the BNC, Brown, ICE and MICASE, used where appropriate with the packaged software (VIEW, Sara, etc). Many researchers created their own corpora according to students' needs and preferences: some sourced the internet, some used CD-ROMs, others scanning printed works (e.g. textbooks). In some cases students created their own corpora, usually as part of a corpus project. Overall, WordSmith Tools was the most popular software, being used in 12 studies, followed by MicroConcord in five. A variety of others were also used, including some home-produced, but surprisingly the web was only used directly as a corpus in two studies. The majority of studies allowed learners direct access to computers, although a few provided only paper print-outs.

Aims

The studies as a whole have extremely diverse objectives in mind, often attempting several things at once, which makes any definitive summary difficult. However, it seems convenient to class them into three main groups. In one, the main focus is on learners' attitudes towards corpus use; in another, the emphasis is on learners' practices – what they do and how well they do it, i.e. whether they are capable of becoming amateur corpus linguists. While these are both important areas, it is worth noting that neither attempts directly to evaluate the efficiency of corpus use by learners. This is the aim of the last group, which does focus on the L2, and here we can detect two major currents. Firstly, there are those which look at the use of corpus tools for reference purposes, essentially for translation, writing or error-correction. This may indeed turn out to be the main interest of corpora for learners (as indeed it can be for native speakers), but in these studies little if any attempt is made to investigate whether learning takes place from such use. Only a handful of studies tackle this directly; here the focus is almost exclusively on aspects of lexis, with very controlled tasks between an experimental and a control group, or in before-and-after situations.

Design

The studies here seem to fall into two main categories: those which start with a course, which they then seek to evaluate; and those which start with a research question, which they then attempt to investigate. These basic paradigms are reflected in the research instruments used: apart from "informal feedback", among the most popular are classroom discussions (11 studies), questionnaires (10) and interviews (8), often in combination. A few made use of automatic tracking, learner diaries, and classroom observation, while others analysed language use in specific set tasks. Interestingly, the evaluation in 11 cases was based on a written or oral report which the students had to provide on a language point they had studied using corpus techniques. The diversity of research instruments reflects the different research questions focused on; this is on balance likely to be beneficial, although it does make comparison and overall conclusions difficult.

Scale

For the 34 studies which provide explicit information, the average size of the learner population is 38.97. Five studies involve more than 100 students, while at the other end of the scale, there are two case studies of a single learner; another six involve 10 students at most. Eight studies involved a control group per se; four others managed to construct the design so that the same informants switched between two different task types, experimental and control; nine compared performance in pre- and post-tests. 16 of the 39 papers are purely qualitative in nature; a further 11 involved only raw numbers or percentages, with no or extremely limited statistical analysis. The remaining 12 attempt a more or less quantitative approach, but only six evaluate language learning as such (cf. "aims" above), and these are exclusively lexical or collocational in nature.

Discussion

If one may attempt a sweeping synthesis of such a variety of research, the general conclusion is that both qualitative and quantitative studies produce highly encouraging results: learner attitudes are largely positive; in most cases they are remarkably capable of corpus techniques; corpora can be used as an effective reference tool, as well as for learning. But however enthusiastic, all the studies here are also careful to point out limitations; in particular, it seems that the use of corpora may not be appropriate for all learners, at all levels, for all language points. Careful study is needed not just to show that DDL works, but in what conditions.

The majority of research to date concentrates on fairly sophisticated students in a university environment at a relatively advanced level of L2 ability. It may of course simply be that researchers favour their local environments for practical reasons. Alternatively, it has frequently been claimed that DDL is most suitable in such cases, but this position seems to be essentially an intuitive one: as we have seen, there has been extremely little empirical research to date with younger, less sophisticated, lower level learners in school environments with limited resources (in particular, without regular class access to a computer laboratory). The sheer size of school populations might encourage researchers not to reject the possibilities out of hand, but rather to explore empirically whether DDL can have anything to offer in such cases.

Few would argue for a radical corpus revolution in the classroom, but it is easy to see how DDL activities (such as those in Tribble & Jones, 1997) could be integrated into course books, and how publishers could produce more learner-oriented software and pedagogic corpora, either available as companion websites to published material or as stand-alone sites, for private study or for teachers to access and print out for class work. However, there has been minimal uptake by publishers and other key decision-makers in L2 learning – education ministries, teacher training institutes, educational establishments, etc. More empirical evidence, if positive, might enable DDL to break out of its current research confines and into wider L2 contexts.

The call for more research can be found repeatedly in the DDL literature, empirical or otherwise. 39 studies may seem a reasonable number, but it amounts to just over two a year since the Johns and King (1991) collection, and is largely the work of a small number of researchers. As Angela Chambers (2007: 5) puts it in her survey of 12 DDL papers, "it is worth asking why there are not more large-scale quantitative studies" – or, indeed, more empirical studies of any kind in the field. One possibility is that researchers are understandably daunted by the prospect of implementing large-scale, longitudinal studies carefully controlling for large numbers of variables. However, there seems little obvious reason why such difficulties should be greater in DDL than other areas of L2 pedagogical research. While large studies are of course desirable, it is often possible to separate out subsidiary questions to be tackled individually on a more modest scale.

Given the number of variables involved, no single study is likely to "prove" very much, just as a single concordance line is not the best evidence for language use. To take the analogy further, corpus linguistics looks at many concordances to find the general tendencies of language patterning; what is needed here is a large number of studies in DDL to see where the weight of evidence takes us. Without empirical support, the most we can hope for are statements along the lines of "I think", "it seems to me", "in our opinion", etc. – which do indeed feature prominently in much of the DDL literature. While such statements may be based on reasonable arguments, they are perhaps insufficiently powerful to convince the major decision-makers to invest in the production of appropriate materials, or to allow DDL techniques significant place in teacher training or L2 curricula. It is at the least ironic that empirical evidence should be so lacking in a field relating to corpus linguistics, where the nature of evidence is crucial

References

Angela Chambers. 2007. Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana. (eds) Corpora in the Foreign Language Classroom. Rodopi: Amsterdam, pp3-16.

Tim Johns. 1991. From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning. In CALL Austria, 10, p. 14-34.

Tim Johns & Philip King. (eds) 1991. Classroom Concordancing: English Language Research Journal, 4.

Chris Tribble & Glyn Jones. 1997. Concordances in the Classroom. Athelstan: Houston.

Articles examined here (where a single study is reported in more than one paper, all references are given)

Rachel Allan. 2006. Data-driven Learning and Vocabulary: Investigating the Use of Concordances with Advanced Learners of English. Trinity College Dublin, Centre for Language and Communication Studies, Occasional Paper, 66.

Guy Aston. 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, pp178-191.

Guy Aston. 1997. Involving Learners in Developing Learning Methods: Exploiting Text Corpora in Self-Access. In P. Benson & P. Voller. (eds) Autonomy and Independence in Language Learning. Longman: London, pp204-263.

Sylvia Bernardini. 2000. Systematising Serendipity: Proposals for Concordancing Large Corpora with Language Learners. In L. Burnard & T. McEnery. (eds) Rethinking Language Pedagogy from a Corpus Perspective. Peter Lang: Frankfurt, pp225-234.

Sylvia Bernardini. 2002. Exploring New Directions for Discovery Learning. In B. Kettemann & G. Marko. (eds) Teaching and Learning by Doing Corpus Analysis. Rodopi: Amsterdam, pp165-182.

Alex Boulton. 2006a. Bringing Corpora to the Masses: Free and Easy Tools for Language Learning. 7th Teaching and Language Corpora Conference. BNF / Université Paris 7–Denis Diderot: Paris, France. 1-4 July.

Alex Boulton. 2006b. Tricky to Teach, Easier to Learn: Empirical Evidence for Corpus Use in the Language Classroom. American Association of Applied Corpus Linguistics. University of Northern Arizona: Flagstaff AZ, USA. 20-22 October.

Alex Boulton. 2007a. Looking (for) Empirical Evidence for DDL at Lower Levels. Practical Applications of Language and Computers. Lodz University: Lodz, Poland. 19-22 April.

Alex Boulton. 2007b. DDL is in the Details… and in the Big Themes. 4th Corpus Linguistics conference. University of Birmingham Centre for Corpus Research: Birmingham UK. 27-30 July.

Alex Boulton & Stephan Wilhelm. 2006. Habeant Corpus—they should have the Body. Tools Learners have the Right to Use. In Asp, 49-50, pp155-170.

Lynne Bowker. 1998. Using Specialised Monolingual Native-Language Corpora as a Translation Resource: a Pilot Study. In Meta, 43/4, pp631-651.

Lynne Bowker. 1999. Exploring the Potential of Corpora for Raising Language Awareness in Student Translators. In Language Awareness, 8/3-4, pp160-173.

Angela Chambers. 2005. Integrating Corpus Consultation in Language Studies. In Language Learning & Technology, 9/2, pp111-125.

Winnie Cheng, Martin Warren & Xu Xun-feng. 2003. The Language Learner as Language Researcher: Putting Corpus Linguistics on the Timetable. In System, 31/2, pp173-186.

Maria Ciezielska-Ciupek. 2001. Teaching with the Internet and Corpus Materials: Preparation of the ELT Materials Using the Internet and Corpus Resources. In B. Lewandowska-Tomaszczyk. (ed) PALC 2001: Practical Applications in Language Corpora, Lodz Studies in Language, 7. Peter Lang: Frankfurt, pp521-531.

Tom Cobb. 1997a. From Concord to Lexicon: Development and Test of a Corpus-Based Lexical Tutor. Unpublished PhD thesis. Concordia University: Montreal.

Tom Cobb. 1997b. Is there any Measurable Learning from Hands-on Concordancing? In System, 25/3, pp301-315.

Tom Cobb. 1999a. Breadth and Depth of Lexical Acquisition with Hands-on Concordancing. In CALL, 12/4, pp345-360.

Tom Cobb, Chris Greaves & Marlise Horst. 2001. Can the Rate of Lexical Acquisition from Reading be Increased? An Experiment in Reading French with a Suite of On-line Resources, translated in P. Raymond & C. Cornaire (eds) Regards sur la Didacdtique des Langues Secondes. Editions Logique: Montreal, pp133-153.

Andy Cresswell. 2007. Getting to "know" Connectors? Evaluating Data-driven Learning in a Writing Skills Course. In E. Hidalgo, L. Quereda & J. Santana. (eds) Corpora in the Foreign Language Classroom. Rodopi: Amsterdam, pp267-287.

Alejandro Curado Fuentes. 2002. Exploitation and Assessment of a Business English Corpus through Language Learning Tasks. In ICAME Journal, 26, pp5-32.

Alejandro Curado Fuentes. 2003. The Use of Corpora and IT in a Comparative Evaluation Approach for Oral Business English. In ReCALL, 15/2, pp189-201.

Alejandro Curado Fuentes. 2007. A Corpus-Based Assessment of Reading Comprehension in English. In E. Hidalgo, L. Quereda & J. Santana. (eds) Corpora in the Foreign Language Classroom. Rodopi: Amsterdam, pp309-326.

May Fan & Xu Xun-feng. 2002. An Evaluation of an Online Bilingual Corpus for the Self-Learning of Legal English. In System, 30/1, pp47-63.

Ana Frankenberg-Garcia. 2005. A Peek into what Today's Language Learners as Researchers Actually Do. In International Journal of Lexicography, 18/3, pp335-355.

Delian Gaskell & Tom Cobb. 2004. Can Learners use Concordance Feedback for Writing Errors? In System, 32/3, pp301-319.

Gregory Hadley. 2002. Sensing the Winds of Change: an Introduction to Data-driven Learning. In RELC Journal, 33/2, pp99-124.

Marlise Horst, Tom Cobb & Ioana Nicolae. 2005. Expanding Academic Vocabulary with an Interactive On-line Database. In Language Learning & Technology, 9/2, pp90-110.

W-R. Ilse. 1991. Concordancing in Vocational Training. In T. Johns & P. King. (eds) Classroom Concordancing, English Language Research Journal, 4, pp103-113.

Tim Johns. 1997. Contexts: the Background, Development and Trialling of a Concordance-based CALL Program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles. (eds) Teaching and Language Corpora. Addison Wesley Longman: Harlow, pp100-115.

Claire Kennedy & Tiziana Miceli. 2001. An Evaluation of Intermediate Students' Approaches to Corpus Investigation. In Language Learning & Technology, 5/3, pp77-90.

Claire Kennedy & Tiziana Miceli. 2002. The CWIC Project: Developing and Using a Corpus for Intermediate Italian Students. In B. Kettemann & G. Marko. (eds) Teaching and Learning by Doing Corpus Analysis. Rodopi: Amsterdam, pp183-192.

Mansour Koosha & Ali Akbar Jafarpour. 2006. Data-driven Learning and Teaching Collocation of Prepositions: the Case of Iranian EFL Adult Learners. In Asian EFL Journal Quarterly, 8/4, pp192-209.

Julia Lavid. 2007. Contrastive Patterns of Mental Transitivity in English and Spanish: a Student-centred Corpus-based Study. In E. Hidalgo, L. Quereda & J. Santana. (eds) Corpora in the Foreign Language Classroom. Rodopi: Amsterdam, pp237-252.

David Lee & John Swales. 2006. A Corpus-based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-compiled Corpora. In English for Specific Purposes, 25, pp56-75.

Belinda Maia. 1997. Making Corpora: A Learning Process. In G. Aston, L. Gavioli & F. Zanetti. (eds) Proceedings of Corpus Use and Learning to Translate. (http://www.sslmit.unibo.it/cultpaps/paps.htm, accessed via http://web.archive.org/ April 2006)

Cynthia Mparutsa, Alison Love & Andrew Morrison. 1991. Bringing Concord to the ESP Classroom. In T. Johns & P. King. (eds) Classroom Concordancing, English Language Research Journal, 4, pp115-134.

Barbara Seidlhofer. 2000. Operationalizing Intertextuality: Using Learner Corpora for Learning. In L. Burnard & T. McEnery. (eds) Rethinking Language Pedagogy from a Corpus Perspective. Peter Lang: Frankfurt, pp207-223.

Barbara Seidlhofer. 2002. Pedagogy and Local Learner Corpora: Working with Learner-driven Data. In S. Granger, J. Hung & S. Petch-Tyson. (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. John Benjamins: Amsterdam, pp213-234.

Elke St John. 2001. A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language. In Language Learning & Technology, 5/3, pp185-203.

Vance Stevens. 1991. Concordance-Based Vocabulary Exercises: A Viable Alternative to Gap-filling. In T. Johns & P. King. (eds) Classroom Concordancing. English Language Research Journal, 4, pp47-61.

Yu-Chin Sun & Li-Yuch Wang. 2003. Concordancers in the EFL Classroom: Cognitive Approaches and Collocation Difficulty. In CALL, 16/1, pp83-94.

Richard Watson Todd. 2001. Induction from Self-selected Concordances and Self-correction. In System, 29/1, pp91-102.

Jill Turnbull & Jack Burston. 1998. Towards Independent Concordance Work for Students: Lessons from a Case Study. In ON-CALL, 12/2, pp10-21.