Data-driven learning for younger learners: Obstacles and optimism.

Alex Boulton

Data-driven learning for younger learners: obstacles and optimism

Alex Boulton, ATILF – CNRS & University of Lorraine
alex.boulton@atilf.fr
orcid.org/0000-0001-6306-8158

Foreward to P. Crosthwaite (Ed.). (2019). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Although I've been researching data-driven learning for many years now, I know virtually nothing about DDL with younger learners. The simple fact is that probably no one really does. An ongoing but certainly not exhaustive collection of research in the area currently brings up 378 separate publications featuring empirical study of DDL.[1] Of these, only 19 explicitly state that the participants are in high school, and none in a primary school context. The overwhelming majority are conducted in university contexts, for a variety of reasons. But before we explore these, perhaps we need to define what we mean by DDL. This is no easy task, and I often read published papers or listen to conference presentation where the authors explicitly situate their work as DDL, only to hear colleagues respond, "but that's not data-driven learning!"

That DDL means different things to different people is not new. The approach is closely associated with work by Tim Johns who coined the term in 1990; in his various publications, he gave examples of open-ended serendipitous corpus exploration, one-to-one writing correction, class use of hands-on concordancing on pre-determined language points, reusable paper-based materials for remedial grammar, and even 'blackboard concordancing' whereby each learner is given a page of a text, directed to find examples of a particular feature and write them on the blackboard in concordance-like format. So it should come as no surprise that there are many definitions of DDL. At the narrow end, we might adopt a prototype definition, i.e. a kind of core which everyone can agree on. I once suggested this might be something like "the hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning of advanced usage" (Boulton, 2011, p. 572). With a prototype definition like this, "the further an activity is from the central, prototypical core, the less DDL-like it is, but any cut-off point beyond which we might like to say 'this is no longer DDL' seems likely to be arbitrary rather than empirically grounded or based on a coherent, hermetic definition" (p. 563). While a prototype definition has its uses, in many cases a more general definition is preferable such as that proposed by Gilquin and Granger: "Data-driven learning (DDL) consists in using the tools and techniques of corpus linguistics for pedagogical purposes" (2010, p. 359).

DDL is not a theory of language or language learning, though it can be argued that corpus linguistics has radically changed our understanding of language, giving rise to a number of theories such as the idiom principle (Sinclair, 1991), lexical priming (Hoey, 2005), the mental corpus (Taylor, 2012), and norms and exploitations (Hanks, 2013). Further, it aligns with several important principles and theories in current thinking, such as authenticity,

[1] 'Empirical' here is defined as studies that "subject some aspect of DDL to observation or experimentation with some kind of externally validated evaluation other than the researchers' own intuition" (Boulton, 2010, p. 130).

Alex Boulton (Forthcoming). **Foreword. Data-driven learning for younger learners: Obstacles and optimism.** In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

autonomy, chunking, consciousness-raising, constructivism, critical thinking, complexity and dynamic systems theory, discovery learning, focus on form, individualisation, induction, learner-centredness, (meta-)cognitive skills, noticing, salience, task-based learning, usage-based learning, etc. It is beyond the scope of this short text to discuss these alleged advantages in detail; various aspects are raised in the papers present here (see also Flowerdew, 2015). But if DDL is so wonderful, why isn't it more widespread? And in particular, why has it been comparatively under-researched in the case of younger learners?

First, it could be simply that the majority of researchers in the field are themselves university academics using DDL in their own teaching (cf. Chambers, 2019). This certainly includes me, but the case is not exclusive to DDL, with university students being by far the most extensively studied population in second language acquisition (SLA) and applied linguistics as a whole. Although languages are of course widely taught to younger learners, research in that context tends to be more the domain of education rather than applied linguistics, each field having its own journals and cultures. Unfortunately, the two do not talk to each other very much, with the result that DDL work in applied linguistics remains relatively unknown in educational circles. The consequence of this is not only that DDL is understudied with younger learners, but also that DDL and corpus linguistics tend not to form a major part of teacher training programmes (see Hirata, this volume). Consequently, trainee teachers (outside applied linguistics programmes) are rarely introduced to DDL and cannot themselves use its tools or techniques in their own teaching. As Conrad (2000, p. 556) notes, "the strongest force for change could be a new generation of ESL teachers who were introduced to corpus-based research in their training programs"; until such time as researchers in applied linguistics talk to researchers and decision-makers in education, the current situation is unlikely to change much and DDL will remain on the margins, rarely integrated (Wicher, this volume), and never "normalised" (Chambers, 2019). Of course, the same can be said of the students themselves: school exams and teaching practices need to be constructively aligned for any innovation to have the opportunity to penetrate (Meunier, this volume), and it is understandable that language students (just like student teachers) might resent time spent on practices that do not directly help with exam results (Szudarski, this volume).

A second reason why DDL is less studied among younger learners might be that researchers and practitioners expect it not to work very well in the first place (see also Schaeffer-Lacriox, this volume). This could be because their young age is taken to be synonymous with insufficient levels of language proficiency, an issue much debated even in university contexts, where few projects have explored DDL among lower-level learners. One solution is to use prepared materials of appropriately selected items, and even for hands-on work the texts can be selected to be appropriate for the target users, or graded for level (e.g. Hadley & Charles, 2017). But it is certainly true that there is a lack of appropriate corpora and tools beyond academic corpus linguistics, as argued by Pérez-Paredes (this volume) who looks at uses of SACODEYL, a corpus of teenage interviews. If existing corpora tend to be geared to older users (or linguists), it is still possible for teachers to create their own corpora that are specifically relevant to younger learners, as shown in several other papers in this volume: corpora of children's news in Di Vito; corpora of the learners' textbooks in Papaioannou *et al*.; multimodal corpora in Hirata.

Alex Boulton (Forthcoming). **Foreword. Data-driven learning for younger learners: Obstacles and optimism.** In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

The language points discussed in DDL also tend to be relatively advanced, but this is not in itself an indication that DDL is inaccessible for other features appropriate to lower levels. Liontou (this volume), for example, provides successful results in the "notoriously difficult" area of idiomatic phrases. Indeed, most of the studies with lower levels have found promising results, although two recent meta-analyses of DDL come to slightly differing conclusions. For DDL as a whole, Boulton and Cobb (2017) analysed 88 unique samples from 64 studies and found that proficiency did not systematically correlate with effect sizes, suggesting that it is open to all levels. Lee *et al*. (2019) focused on DDL just for vocabulary in 38 samples from 29 studies, and did find a medium difference between advanced and low-level learners, but there was still a significant effect in all cases. Caution should always be a watchword in dealing with such findings since proficiency is not only difficult to assess, but is reported in so many ways that what is 'advanced' in one study might be considered 'intermediate' or even 'low' in another context. Of studies apparently with advanced learners in major CALL journals, Burston and Arispe (2016) found that fully 50% were at best B1 level on the Common European Framework of Reference for Languages, which is commonly thought of as lower intermediate. But coming back to the main point: "DDL works pretty well in almost any context where it has been extensively tried" (Boulton & Cobb, 2017, p. 386).

Another explanation might be that children lack the cognitive development necessary for sophisticated language work. A number of studies have addressed DDL with children in exploring their mother tongue (L1); some of these are mentioned in Crosthwaite and Stell (this volume). Over ten years ago, John Sinclair, often considered the father of modern corpus linguistics with the COBUILD project, planned the wide-ranging introduction of corpora to primary schools in Scotland via a tool known as PhraseBox:

> PhraseBox… is like giving each pupil real-time access to a huge memory of all the different ways in which thousands of people have expressed themselves over several years, all instantly available in a highly organised presentation. Gradually, students are expected to internalise what they need of the resource and gather confidence in their ability to express themselves publicly; but the resource will always be available when it is needed. (Sinclair, 2006, n.p.)

Sadly, the resource never did become available, as Sinclair died in 2007, and the project along with him. I was fortunate to see him demonstrate it with Ana Mauranen at the IVACS conference in Nottingham in 2006, but he wrote little about it (see also Stubbs, 2011). Nonetheless, the very fact that he and people in a national education system believed in it is revealing in itself: there is no *a priori* reason to assume that younger learners are unable to use the technology and analyse language in ways compatible with DDL.

This brings us specifically to the issue of ICT (information and communication technology), and whether younger learners have the skills necessary for DDL. To counter that, it might be pointed out that DDL can be conducted using prepared, paper-based print-outs which still give medium effect sizes (Boulton & Cobb, 2017). Further, younger learners who have grown up with ICT may actually be more relaxed about it than older learners, as discussed by Gatto (this volume). A case in point: some of my students in an English master's degree have

Alex Boulton (Forthcoming). **Foreword. Data-driven learning for younger learners: Obstacles and optimism.** In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

difficulty with the wildcard function which so impressed a ten-year-old in Crosthwaite and Stell (this volume).

Finally, the fact that there are few published studies of DDL for children does not necessarily mean that it is absent from common classroom practice. Johns and King (1991, p. iii) situated DDL "within the overall aim of developing students' ability to puzzle things out for themselves." For our purposes, this has to involve language data in one form or another, but otherwise that leaves the field very open. So in particular, while the internet is not a 'corpus' and Google is not a 'concordancer', it is quite possible that teachers are encouraging their learners to explore language without ever having heard of DDL. As McCarthy (2008, p. 566) puts it, "we are, all of us, corpus users, because we use the internet." And if we are using the internet as a source of repeated language examples, we are all doing DDL (Gatto, this volume; see also Boulton, 2015). As one student remarked of COCA in the study by Papaioannou *et al*. (this volume): "This is just another tool the internet offers us to learn new words". And of course it is possible that learners are doing this without their teachers' knowledge, or even against their recommendations. Meunier (this volume) discusses several tools that provide opportunities for very DDL-like activities, using accessible corpora (e.g. TV series such as Breaking Bad or Game of Thrones in www.playphrase.me), and with no mention of 'corpus' or 'concordance' or other problematic terms.

Research can be based on actual practice, and can certainly inform it, but should never be 'applied'. The main consumers of research are other researchers rather than teachers or decision-makers (e.g. Borg, 2009), but when teachers do read research, it is essential that they interpret and adapt research findings to their own local contexts and needs: existing reports of DDL aimed at students and adults should not be simply transformed into a 'corpus approach for kids'. If the search for a single 'best method' has been largely abandoned in recognition of the diversity of teaching goals, contexts, needs and profiles of the individuals involved (Brown *et al*. 2007), it cannot be desirable to impose DDL as the sole form of language instruction. Researchers thus also need to remain open to a variety of approaches, and not be dogmatic about what is or is not DDL and how it should be implemented across the board. As Meunier (this volume) argues, we need to think outside the box and go off the beaten path to promote creativity in DDL. It is refreshing that some of the examples given in this volume (e.g. Liontou; Papaioannou *et al*.) are quite different from usual expectations of DDL. What we have is a host of variants under the umbrella term of DDL, which have their place alongside others for some learners in some contexts and for some language questions.

I hope this short foreword has shown that the common assumption that DDL is not suitable for younger learners is just an assumption – an absence of research evidence merely means that it has not been widely tested, not that it is not or cannot be effective. DDL might even be more appropriate for children than for older learners, as their expectations are less fixed and the approach lends itself to the inquisitive; the children in Di Vito (this volume) remembered the methodology well because they felt involved and valued and encouraged to think for themselves. There certainly are obstacles and we should not be blasé about them (Schaeffer-Lacroix, this volume), but there are any number of reasons to think that DDL, appropriately implemented, can bring something to the table in a variety of contexts. The present volume thus addresses a genuine need for DDL research among younger

Alex Boulton (Forthcoming). **Foreword. Data-driven learning for younger learners: Obstacles and optimism.** In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

learners, with the various papers examining relevant arguments, describing potential uses and evaluating experimental or ecological examples.

References

Borg, S. (2009). English language teachers' conceptions of research. *Applied Linguistics*, 30(3), 358-388. DOI: 10.1093/applin/amp007

Boulton, A. (2010). Learning outcomes from corpus consultation. In M. Moreno Jaén, F. Serrano Valverde & M. Calzada Pérez (Eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching* (pp. 129-144). London: Equinox.

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.

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. DOI: 10.1075/scl.69.13bou

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348-393. DOI: 10.1111/lang.12224

Brown, H., Tarone, E., Swan, M., Ellis, R., Prodromou, L., Jung, U., Bruton, A., Johnson, K., Nunan, D., Oxford, R., Goh, C., Waters, A., & Savignon, S. (2007). Forty years of language teaching. *Language Teaching*, 40, 1-15. DOI: 10.1017/S0261444806003934

Burston, J., & Arispe, K. (2016). The contribution of CALL to advanced-level foreign/second language instruction. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture* (pp. 69-73). Dublin: Research-Publishing.net. DOI: 10.14705/rpnet.2016.eurocall2016.539

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, FirstView. DOI: 10.1017/S0261444819000089

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34, 548-560. DOI: 10.2307/3587743

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15-36). Amsterdam: John Benjamins. DOI: 10.1075/scl.69

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). London: Routledge.

Hadley, G., & Charles, M. (2017). Enhancing extensive reading with data-driven learning. *Language Learning & Technology*, 21(3), 131-152. http://llt.msu.edu/issues/october2017/hadleycharles.pdf

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: MIT Press. DOI: 10.1017/S1351324913000302

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge. DOI: 10.4324/9780203327630

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., & King, P. (Eds.) (1991). *Classroom concordancing. English Language Research Journal*, 4.

Alex Boulton (Forthcoming). **Foreword. Data-driven learning for younger learners: Obstacles and optimism.** In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, Advance Article. DOI: 10.1093/applin/amy012

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, 41(4), 563-574. DOI: 10.1017/S0261444808005247

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

Sinclair, J. (2006, January). A language landscape. *West Word*. www.westword.org.uk/jan2006.html

Stubbs, M. (2011). A tribute to John McHardy Sinclair (14 June 1933 – 13 March 2007). In T. Herbst, S. Faulhaber & P. Uhrig (Eds.), *The phraseological view of language: A tribute to John Sinclair* (pp. 1-16). Berlin: De Gruyter Mouton. DOI: 10.1515/9783110257014

Taylor, J. (2012). *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199290802.001.0001