Research in data-driven learning

Alex Boulton

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

Chapter 1

Research in data-driven learning

Alex Boulton

Abstract

Data-driven learning (DDL) typically involves language learners consulting corpus data, either directly or via prepared materials, to answer questions about language. The approach has been mooted since the beginning of the modern era of corpus linguistics, and has come to be associated with work by Tim Johns who coined the term in print in 1990. Since then, hundreds of studies have attempted to evaluate some aspect of DDL, giving rise to several reviews and syntheses. This paper introduces DDL and discusses the syntheses to date, before analysing a rigorous collection of 351 studies published up to and including 2018. While previous syntheses have evaluated the field, the objective here is to provide an overview of how researchers see DDL across the board, to identify more clearly what DDL actually looks like today, how it has evolved from its early beginnings in the 1980s, and to suggest avenues for future research in underexplored areas.

Keywords: DDL, data-driven learning, concordancing, synthesis

Data-driven learning

Language teaching is about more than just furnishing students with language, since we can never give students all the language needed for every future situation. It is also important to offer them the cognitive and technological tools that will allow them to face the unpredictable range of language situations they may encounter in their professional and social lives with greater autonomy. One approach in this direction involves the use of corpora in what has become known as data-driven learning or DDL (Johns 1990), whereby learners can query large collections of language data to find recurring patterns and thus find answers to their own questions 'in the wild' rather than being dependent on teachers or published resources. As such, it is compatible with current trends towards individualisation and learner autonomy for work outside class and indeed for life-long learning.

An exact definition of DDL is open to debate. Johns himself provided many definitions, most famously perhaps in the introduction to Johns and King (1991): "the use in the classroom of computer-generated concordances to get students to explore the regularities of patterning in the target language" and develop their "ability to puzzle things out for themselves" (p. iii). In practice, this is generally via a corpus in the linguistic sense, i.e. a large collection of authentic texts designed to be representative of the language variety targeted, in electronic format for use with corpus tools. However, DDL may be used with just a single text, as proposed by Johns et al. (2008) with lower-level learners exploring chapters in a novel. Johns even described 'blackboard concordancing' (1993), whereby learners each have a page of text and identify a feature such as prepositions to write up on the board. As for authenticity of text, Johns (1997) admitted to editing computer-generated output, while Allan (2009) and others

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

have recommended the use of simplified readers; there seems to be no obvious reason why such practices should be excluded. Similarly, Google can be used as a substitute concordancer to query the 'web as corpus'; even if neither aspect is acceptable to many corpus linguists, the procedures may be very similar, and maybe if such resources had been available at the time, that might have been the earliest form of DDL by virtue of accessibility, familiarity, size, range and speed among other things (Boulton, 2015). DDL is also often associated with 'serendipitous' learning, but this too means different things to different people: for Johns (1988) it involved printed concordances and seeing what learners would find in ten minutes to report back to class; for Bernardini (2000) it seems to mean autonomous exploration using a concordancer for learners' own language questions.

For many people, indeed, a major advantage of DDL is that each user can pursue their own queries at advanced level, but Johns (1991) also described 'proactive' materials that could be prepared, printed and used repeatedly to target a recurrent problem of language use, for example in remedial grammar. The use of a concordancer is thus not central to the learner, as it might be used only by the teacher in the preparation of materials. And of course, a concordance is only one function of most corpus query software, which can also provide lists of words, clusters, morphemes, collocations, or key words in comparing two corpora, or show regular and irregular distributions, and so on. The corpus input is necessarily highly visible, although there may be switching back and forth between a concordance (especially in KWIC format: keyword in context) and the text it comes from. If there is no such visibility – if the learners are only reading texts or long extracts selected from a larger data set, or working with single occurrences selected to illustrate a grammar point – then this is generally not considered DDL, even if they are being encouraged to examine text and come to their own conclusions. Finally, the bulk of research in DDL seems to be with advanced learners of a second or foreign language (L2), but there have been studies with native speakers for academic purposes, and even in primary schools (e.g. Crosthwaite & Stell 2019).

So where does this leave us in defining DDL? A prototype definition might be appropriate here, beginning with a core definition that all would agree upon, such as the hands-on use of corpus data in the form of concordances derived from authentic texts, by proficient L2 learners in higher education for inductive, self-directed learning of advanced lexicogrammar (cf. Boulton 2011, p. 572). The advantage of a prototype definition is that it makes no pretence at exclusivity: from this common core, it is possible to deviate in virtually any direction (as we have seen above), becoming progressively less recognisable as DDL in the process. The boundaries are fuzzy: some deviations would be considered so far from the prototype as not to be considered DDL at all by some practitioners (e.g. the use of graded readers or printed materials), while others would disagree. An alternative, more traditional approach to defining DDL is usefully provided by Gilquin and Granger (2010, p. 359) as "using the tools and techniques of corpus linguistics for pedagogical purposes"; to this we might add that the use must directly involve the learners as some point in interacting with the corpus data (hands-on or in prepared materials), and typically for L2 rather than L1.

The assumption underlying much DDL is that examples may be more conducive to learning than information from a dictionary or usage manual since the learner is deriving his or her individual 'rules' that are personally meaningful, and consequently easier to assimilate, retain and apply, especially as a chunk. As such, albeit often implicitly, it reflects a usage-based, exemplar-driven approach to language learning, as well as general constructivist, discovery-

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

based learning in education more broadly. This is something which occurs naturally in everyday life, but formalizing the procedure can promote noticing and language awareness, ultimately creating better language learners (or so one might hope). It also coincides with research findings that show considerable advantages in chunking, whereby natural language use depends not on the construction of each sentence at the moment of use, but prefabricated blocks and frames of language that require less 'assembly' and thus reduce cognitive load (see e.g. Wray, 2002). It is perhaps not surprising that corpus study has given rise to a number of language theories, from Sinclair's (1991) idiom principle to Hoey's lexical priming (2005), Hanks's norms and expectations (2013) and Taylor's mental corpus (2012).

Criticisms of DDL: Theoretical and empirical support

A criticism of DDL is that it is not based on an underlying theory. Pérez-Paredes (2019, p. 17), for example, notes a "lack of theorization" in his survey of DDL in five major CALL journals (computer-assisted language learning). That said, journal articles are generally more concerned with presenting new data from a specific study rather than detailing theoretical underpinnings, which tend to be more the remit of books and chapters. A case in point is Flowerdew (2015), who examines the alignment between DDL and three language learning theories: the noticing hypothesis, constructivist learning, and sociocultural theories. It is clear though that DDL originally arose from empirical practice in teaching and learning situations; any theorising follows from that rather than vice versa. Chambers (2019, p. 464) attributes the continued enthusiasm among the research community to the way that DDL "corresponds perfectly with a number of contemporary paradigms in research and practice in education in general, and in language teaching in particular".

Ironically, an older criticism of DDL is that it lacks *empirical* support, and that the approach is "more talked about than tested" (An & Xu, 2013, p. 695). An obvious question then is: if papers on DDL provide no theoretical underpinning and no empirical support, then what do they talk about? Typically, in early days of a new approach, method or technology, publications are descriptive of initial practice, outlining instances of use to exemplify the potential, or speculate about potential uses and rationales (Shintani et al., 2013). The first DDL papers seem to do just that, such as McKay (1980), who outlines examples of uses with university students for teaching the syntactic, semantic and pragmatic dimensions of verbs. If this is apparently the first publication relating to corpus use in language learning – ten years before Johns coined the term DDL, and four years before he first mentioned such uses in print (Higgins & Johns, 1984) – it is certainly not the first time corpora had been used in this way. McEnery and Wilson (1997, p. 6) attribute this to Peter Roe who was apparently using corpora in a language for specific purposes (LSP) course at Aston University in Birmingham as far back as 1969. McKay's paper offers no evaluation of the approach adopted, hence the numerous lamentations that "most work on [classroom concordancing] tends to slant towards the speculative rather than the evaluative end" (Ma, 1993, p. 24). However, as we have seen, this is not to be expected in the earliest work. More troubling is that far more recent papers still note a dearth of empirical support (e.g. Abu Alshaar & Abuseileek, 2013; Luo, 2016). As we shall see, though, rigorous trawls bring up hundreds of studies that do indeed attempt to evaluate some aspect of DDL, and such complaints are tending to tail off. More recently, researchers have noted the lack of research on specific issues such as particular learner profiles, specialisations, languages and questions, and the pedestrian nature of designs and aims.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

Other criticisms and doubts tend to be pragmatic, including lack of access to the necessary resources (corpora, software, computers in the classroom, adequate wifi, etc.); technical skills and the need for extensive learner training; the linguistic sophistication of the learners vis-à-vis authentic, truncated concordance lines; the loss of teacher control and concrete 'rules' as the fuzzy nature of language comes to the fore; the reasoning skills involved in noticing a problem, translating this into an appropriate query, and interpreting the results (i.e. hypothesis formulation and testing) without drowning in data; the focus on written language, and the emphasis on lexicogrammar at the expense of contextualised, coherent discourse; the boring and mechanical nature of corpus queries, and the time involved in pursuing each query to a satisfactory conclusion. Such concerns should not be brushed aside, but that does not mean they are insurmountable (see Boulton, 2009), and some have reduced importance today. Among other things, vastly more corpora of various types are now available, as well as free, stable, simple software for work on- or off-line; in many countries, learners are very familiar with other types of search engines and 'snippets' or abbreviated results, and appreciate the abundance of examples; the fuzzy nature of authentic language needs to be addressed, as well as the importance of chunks and patterning over rules; the time invested should diminish with practice, and represents an investment for the future in terms of language learning. There is no reason to suppose that DDL should be appropriate for all learners in all contexts and for all language problems, but no approach can (or should) claim this. What is important is that it brings a useful addition to the repertoire of some learners some of the time.

Surveys of DDL

By 2007, enough empirical studies had been published for Chambers to produce the first synthesis of DDL research. In total she finds 12 studies which she divides into quantitative (four papers) and qualitative (nine),[1] with a further subdivision into direct or hands-on consultation via a concordancer in ten cases, and indirect corpus use with printed corpus-based materials in three others. Chambers' interpretation of the results – even with such a small sample at that point in time – is remarkably perspicacious and, disturbingly, may be largely still true today. Her main conclusion is not just that there is a surprising lack of research, but that it tends overwhelmingly towards qualitative studies akin to action-research. This leads her to wonder "why there are not more large-scale quantitative studies" (207, p. 5). She notes considerable variation on many dimensions, in corpus size and scope, ranging from large general corpora to small, locally-built corpora for a specific topic or genre focus. She acknowledges a number of limitations in the spread of research, with very few papers featuring lower-level learners, or outside the university system, or for languages other than English, or for discourse rather than lexicogrammar, though noting that there is no inherent reason why DDL should not apply there. She surmises that it may simply be because researchers work with the students they teach on a regular basis, i.e. enrolled in intermediate or advanced university classes for English. Support for this can be found in the fact that the researchers were also the teachers, as is common practice in much research in language teaching. Most of the studies she examined focus strongly on individual corpus consultation rather than pair or group work, though they occur largely in class rather than outside. While

[1] One paper is counted in both categories; one study is actually discussed in two separate papers.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

refraining from quantifying learning outcomes, she is cautiously optimistic about the impact of DDL, especially in terms of learners' overall positive reactions, while recognising difficulties such as those outlined above. She concludes:

> If corpus consultation … is to become a common activity for learners across the broad spectrum of language studies (general language learning, literary studies, languages for specific purposes, translation, etc.), it would seem necessary for developments to take place in a broader context than that which has been examined here, namely the classrooms of researchers with expertise in corpora and concordancing… It is perhaps outside the classroom that the next important step in research in this area will take place. (Chambers, 2007, p. 13)

This focus on a small number of papers allows for in-depth attention to each individual study; the downside is that it misses a lot of empirical DDL work: the collection outlined in the present chapter features 61 publications up to and including 2005, the date of the most recent paper included in Chambers (2007). This is partly because the inclusion criteria were quite strict, eliminating for example studies using comparable or parallel corpora as well as learner corpora, or studies which also feature linguistics or translation as main objectives. Improved access to research makes it far easier to search for papers today, especially among less well-known or widely-disseminated sources such as national journals, book chapters and conference proceedings. It is of note that most of the papers included are from leading international journals (three in *System*, two in *Language Learning & Technology*, one each in *ReCALL*, the *British Journal of Educational Technology*, the *Journal of Second Language Writing* and *ELR Journal* – the volume edited by Johns and King); the remaining four are all papers from collected volumes published following the biennial TaLC conferences (Teaching and Language Corpora).

At almost the same time as Chambers, Boulton (2008) surveyed 39 empirical DDL studies, though his inclusion criteria were somewhat broader. The main conclusions here are that both qualitative and quantitative results are encouraging, but Boulton recognises the limited scope of research objectives until that time. A later paper (Boulton 2010) looked specifically at learning outcomes from 27 studies, finding "grounds for optimism" (p. 143) for a variety of learner profiles in different context and for a range of purposes. Boulton (2012) reviews 20 papers in English for Specific Purposes (ESP), again with a positive appreciation, though he specifically concludes that "one of the most striking things to emerge is the diversity of the various studies, in terms of research designs and questions, corpora and tools, aims and implementations, which inevitably makes a proper meta-analysis impossible" (p. 277). We shall return to the *inevitability* of this below. Also in a specialized sub-field of DDL, Chen and Flowerdew (2018) reviewed 37 DDL studies relating specifically to English writing in language for academic purposes (LAP) from the year 2000, though two-thirds of them were after 2010-2017, reflecting an increase in research output in the field. Again, the main findings note tremendous variation in the tools and corpora used, with most studies featuring more than one. Large online corpora such as COCA seem to be used more for self-correction, while smaller, specially-compiled corpora tend to be for vocabulary and discourse with offline tools such as AntConc (Anthony, 2019). The authors also note few studies conducted outside the classroom or with lower levels, though the latter is unsurprising given the focus on EAP writing.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

More recently, Chambers (2019) returned to a number of her earlier themes, repeating her 2007 concerns that the vast majority of empirical studies are small scale and feature a tiny community of teacher-researchers enthusiastic about the corpus 'revolution'. While there certainly is a measure of proselytising on the part of DDL practitioners (cf. Römer's repeated call for a "corpus mission", e.g. 2011), others go to considerable lengths to hedge their findings, as this quotation from Cresswell (2007, p. 280) testifies: "Overall, given that the students were advanced and the items already partially known it is possible to conclude, albeit tentatively, that, given language items at the right level, DDL has an observable (though slight) positive effect on actual use." Both of these (an excess of enthusiasm or humility) may contribute to a lack of uptake in mainstream teaching practice; Chambers frames this as corpus use not becoming "normalised" in Bax's sense of technology being used "without our being consciously aware of its role as a technology, as a valuable element in the language learning process." (2003, p. 1). Pérez-Paredes (2019) also focuses on the normalisation of DDL in his systematic review which is rigorous and exhaustive inasmuch as he focuses exclusively on five major CALL journals over a limited time period (2011-2015). He adopts a more objective stance by comparing specific features in line with Chambers and Bax's (2006) taxonomy. While not concluding that the studies are generally qualitative, he notes that they do overwhelmingly feature at least some analysis of learners' attitudes to the use of DDL via questionnaires or interviews. As in other reviews, he finds the scope limited to university contexts, with a focus on corpora as an aid in writing, but with tremendous variation in the main areas under investigation. Although his aim is not to evaluate the pedagogical effects of DDL per se, intriguing is his use of the collection of papers as a corpus in its own right, providing the top multiword keywords as an indicator of the major topics covered.

The surveys discussed so far all share a narrative component insofar as they depend on linear reading and interpretation of the studies involved. The advantage of this is that any and all types of study can be included; the disadvantage is that the synthesist's interpretation is "inevitably idiosyncratic" (Han, 2015, p. 411):

As such, for all its merits – depth of analysis, latitude in sampling primary research, specificity of argumentation, promise of insights, and the like – we must recognize that each review comes from an individual who brings to it his or her own conceptual baggage or particular epistemological stance, which mitigates against objectivity.

An alternative can be found in meta-analyses, which extract the data from quantitative studies which are then pooled together to produce effect sizes for the target feature. It is important to stress that this approach also retains an element of subjectivity in the process, although in theory this is transparent and replicable. More limiting is that it can only include quantitative data and thus misses out on many crucial issues. Nonetheless, there have been a number of meta-analyses of DDL to date which can contribute to our overall understanding of the field. Mizumoto and Chujo (2015) examined 14 studies, all featuring the second author. They found a medium effect size overall ($d = 0.97$) in their pre/post-test comparison, with the best results featuring work on lexicogrammar among these lower-level learners of English as a Foreign Language (EFL). The study thus provides an excellent picture of one specific context in Japan, though the impact is limited for other countries, cultures, languages, proficiency levels and objectives.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

Also in 2015, Cobb and Boulton meta-analysed 21 papers for a broader sweep of DDL research, clearly separating within-groups (pre/post-test) from between-groups (control/experimental) designs since the size of the effect is likely to be very different: any teaching will lead to increased scores between tests, but an experimental group will not necessarily score much higher than the control (typically a comparison group with traditional teaching on the same language items). They found large effects in both (d = 1.68 and d = 1.04 respectively). 'Large' here is derived empirically from Plonsky and Oswald (2014) who looked at 91 meta-analyses in second language acquisition and suggested the top quartile for each design be considered large (i.e. the 25% of largest effect sizes), the second quartile medium and the third small; the bottom 25% of studies producing the smallest effect sizes are of limited practical significance. Cobb and Boulton noted at the time that this was only a preliminary study, and produced a fuller meta-analysis (Boulton & Cobb, 2017) that increased the pool to 88 unique samples derived from 64 separate studies. The more rigorous trawls and inclusion criteria reduced the effect sizes somewhat, though they can still be considered large, both between groups (d = 1.50) and within groups (d = 0.95). Having pooled the quantitative data to produce a general effect size, the next step is to break it down to examine moderator variables. Contrary to the expectations of one of the authors, they found larger effect sizes for hands-on DDL with a concordancer than for using prepared paper-based materials, and larger ones in terms of learning from DDL as opposed to using corpora as a reference resource. On the whole though, they reached the "somewhat surprising and possibly encouraging conclusion that DDL works pretty well in almost any context where it has been extensively tried" (Boulton & Cobb, 2017, p. 386), with medium or large effect sizes on most moderator variables which had been repeatedly studied.

Another major meta-analysis was conducted by Lee et al. (2019), this time limited to DDL studies on vocabulary, and featuring only control/experimental designs. Their analysis of 38 unique samples drawn from 29 studies produced medium sizes overall for both immediate and delayed post-tests (g = 0.74 and 0.64 respectively). The multi-level analysis found large effects particularly for in-depth knowledge (referential meanings and uses) over precise knowledge (definitions) and productive use in context. Like Boulton and Cobb (2017), they detected effects across many conditions, regardless of prior training, corpus type and length of the intervention, for example. On the other hand, this study did find an effect for proficiency, with higher-level learners scoring better; however, conclusions based on just one or two studies are less convincing than those with ten or more. Both these meta-analyses helpfully provide a spreadsheet in MS Excel for others to check their findings or pursue other comparisons, or to add new data as it arises.

This overview highlights the different approaches adopted in narrative synthesis and meta-analysis. Each can contribute to the debate, and neither is complete on its own. What distinguishes most of these syntheses from a traditional literature review is their aim to be as comprehensive as possible given their transparent inclusion criteria. This entails clearly defining the field and the types of publications sought, the search terms, the sources and dates, among other things. There are of course exceptions: as late as 2019, Al-Gamal and Mohammed Ali covered just five examples of "recent literature" (p. 473), dating from 2004 to 2018, with no other criteria specified. Boulton (2017) selected 46 DDL publications for a timeline paper in *Language Teaching*; in line with the remit for this rubric, the choice reflects perceived importance within the field, but he acknowledges a substantial element of subjective bias in the choices. Nonetheless, most of them were primary research studies

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

drawn from a systematic database. The introductions to journal articles, on the other hand, almost never justify the choice of papers for discussion, which seems to depend on serendipitous awareness of research to date combined with a deliberate attempt to support their own study – by citing papers that either confirm the ideas, or which contain some perceived lack or problem to be explicitly countered.

Methodology

The objective of the rest of this paper is not to evaluate DDL further but to outline the state of the field today based on as large a collection of publications that can be assembled up to and including 2018. DDL is defined as use of the tools and techniques of corpus linguistics for pedagogical purposes, with L2 learners interacting with corpus data in some way. While it is certainly the case that some studies are more rigorously designed than others, for present purposes the net was cast as wide as possible and publications were included as long as they featured a full written text in English that outlined an empirical evaluation of some aspect of DDL, while acknowledging that variability of quality is an issue. Ignoring large bodies of work due simply to source of publication cannot be the most satisfactory solution: in CALL as a whole, Gillespie (2020, p. 128) notes that this frequently leads to "reinventing the wheel." Some studies are described in more than one paper; as the focus here is on publication, both sources were retained. Papers in languages other than English were occasionally encountered, but these were not included as it would not be feasible to repeat the procedure rigorously in all languages. It should be borne in mind however that there is at least a small body of DDL research in Chinese, Korean, Japanese and European languages in particular that cannot be included here. The main categories included journal articles, book chapters and conference proceedings from major and minor sources. Excluded were conference slides and master's or doctoral theses due to idiosyncratic indexing and a tendency to be excessively complicated, forming a separate genre (see Lee et al. 2018: 739).

The collection has been added to over time and many items have been located serendipitously. However, rigorous trawls were conducted in late 2014 and 2019 of major databases including LLBA, MLA, JSTOR, DOAJ, and the catalogue system at the author's university (Université de Lorraine, France), as well as Academia, ResearchGate and Google Scholar. Keywords included *DDL*, *data-driven learning*, *corpus*, *concordance(r)* and *Johns*, with *language* and *learning* as contextualisers. Potential items were identified by reading the title, abstract and then the whole paper. The reference lists were scoured for further items and websites checked for any source that contributed more than one paper. This process ultimately produced 351 publications of 'empirical DDL' which are subjected to the analysis below.[2] This number represents an increase of approximately 90% over the initial pool of 205 papers in Boulton and Cobb (2017), given that some sources were now excluded (notably PhDs and texts in languages other to English).

The papers were coded to allow an overview of DDL in its different guises, notably for publication source and date; country, L1 and L2; the number of participants, their context,

[2] The full list of references is available on the author's homepage at https://perso.atilf.fr/aboulton/wp-content/uploads/sites/22/2020/09/BOU_list.pdf

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

specialisation and proficiency; the duration of the course; the corpus type and software used (for hands-on DDL only) and the type of interaction; the research focus and data collected. This often involved guessing where the information is not given explicitly, and the various calculations should be taken as an indication rather than an absolute figure – there is clearly much scope for improving reporting practices and transparency (cf. Plonsky & Ziegler 2016). A detailed individual analysis of 351 publications would clearly be beyond the scope of a single paper, the goal here being rather to provide an overview of DDL to date through the researchers' eyes.

Results and discussion

The total output of empirical DDL studies identified is shown in Figure 1. The trend line reveals a substantial increase over time (the correlation coefficient is $r = .879$), though some studies will inevitably have been missed, especially in later years as indexing can take some time. That 70% have been published in the last 10 years is a sign of healthy interest in different parts of the world. The total volume is more than one might think, especially if trawls are limited to major journals.

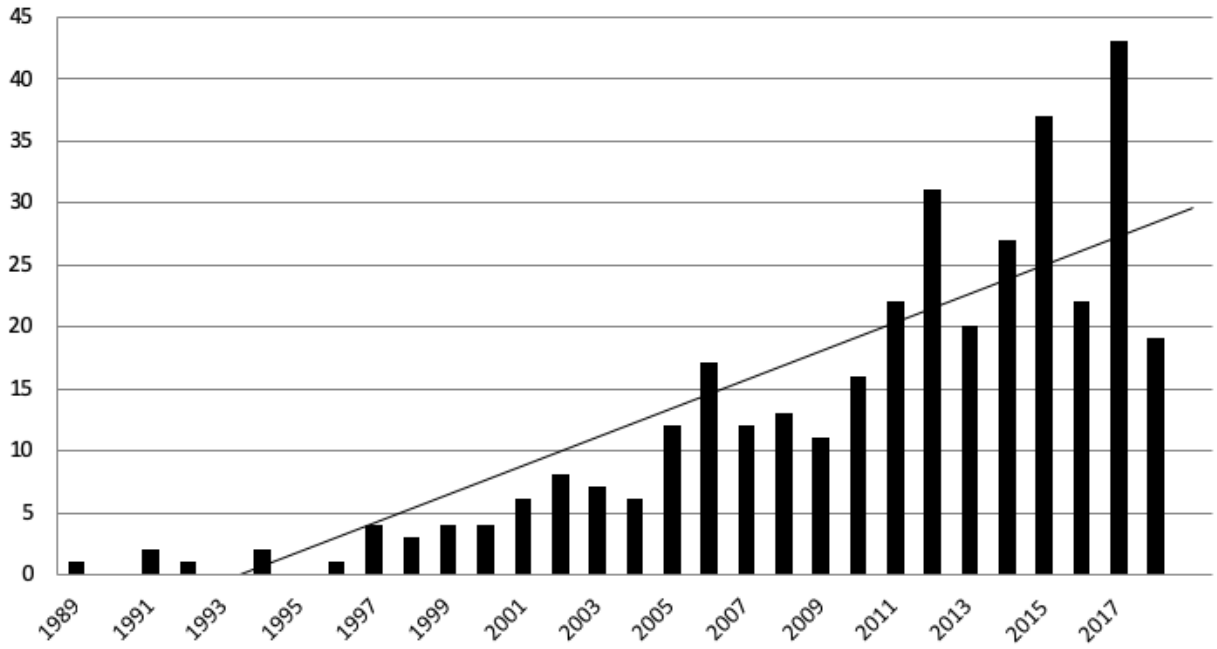

Figure 1. Dates of publication

Publication sources

266 of the publications are journal research articles (76% of all items), 46 are book chapters (13%), 33 are conference proceedings (10%) and 4 are from miscellaneous sources – occasional papers and apparently unpublished conference proceedings (1%). The conferences are quite varied, with only EUROCALL (5 publications), Corpus Linguistics (3) and Asialex (2) occurring more than once each. Other conferences give rise to collected volumes of

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

selected papers, 23 from the biennial Teaching and Language Corpora. The major publishers of book chapters are Peter Lang (13), Rodopi (7) and Benjamins (5).

The most popular journals for publishing empirical DDL papers are generally within the field of CALL, with the top ones being *ReCALL* (26 papers; impact factor 1.361), *Computer Assisted Language Learning* (23; 2.018) and *Language Learning & Technology* (20; 2.571), followed by *System* (8; 1.930) and the *CALICO Journal* (6); then a number of other major language teaching journals: five each in *ELT Journal* (impact factor 1.351) *English for Specific Purposes* (1.704) and the *Journal of English for Academic Purposes* (1.732). All of these currently have a respectable impact factor which places them in the top 60 journals for linguistics on the latest Journal Citation Reports list for 2018, with the exception of the *CALICO Journal* for historical reasons.

The only other source to have published more than five papers is *Procedia – Social and Behavioral Sciences* for a total of 8; although arguably a journal, the issues tend to be more akin to conference proceedings of short papers with highly variable selection procedures. Many of the other journals are national or regional, though some reach a genuinely international audience, such as *Multimedia-Assisted Language Learning* with an overwhelmingly Korean contribution but cited well beyond that sphere. The specific remit is sometimes explicitly acknowledged in the title and again does not automatically limit the readership (e.g. the *British Journal of Educational Technology* or the *JALT-CALL Journal*, where the 'J' stands for Japan). Conversely, including the word 'international' is no guarantee of international reach. Readers should always be vigilant of quality, however defined, and are likely to be especially suspicious of lesser-known sources. However, even major journals do on occasion publish papers of dubious methodological design or which tread essentially the same ground as previous research and thus contribute little new substance,[3] while some excellent studies are published in virtual unknown outlets. These may be from researchers who would otherwise not perhaps attempt to submit to a major international journal, providing an alternative to the English-speaking or Eurocentric bias evident in major journals. In making judgements about the quality of journals (let alone individual studies), it is extremely difficult to avoid a large dose of subjectivity – hence the widespread use of bibliometrics such as impact factor despite their flaws. It is thus a reality that papers published in those journals are likely to be more influential, even if the quality cannot be guaranteed even there. To provide a picture of recent, high-impact research, the 46 papers published only in those journals mentioned in the previous paragraph during the last five years (2014-2018) will feature alongside the overall picture in the rest of this analysis; they are labelled RMJ for 'recent major journal articles'. Potentially misleading percentages are not systematically provided when dealing with small subsets of papers .

Language and geography

English is the sole target language in 308 of the 351 papers (88%; RMJ 85%) and is alongside one or more other languages in 3 more. Of the remaining 11%, it is worth noting that very few studies have been conducted with non-European languages: 2 each for Mandarin (both in

[3] Replication studies notwithstanding. Several papers here call for their study to be repeated, but only two explicitly claim to be formal replications.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

RMJ) and Cantonese. Otherwise, several papers targeted other European languages: 12 each for French and German, 7 for Italian, 4 for Spanish, 1 each in Catalan and Czech. The vast majority of studies (80%; RMJ 70%) are carried out in a foreign-language setting, i.e. where the target language is not the language of the country.[4] Of those in a second-language environment, 77% were in an English-speaking country. 83% were carried out with native speakers of the local language, with one major exception: 42 of the 48 studies with participants of mixed language backgrounds (88%) were conducted in English-speaking countries; of the other 22 studies where the participants' L1 was not a local language, 16 (72%) were also in an English-speaking country, especially with Asian students.

Overall, 80 papers (23%) were conducted in an English-speaking country, mainly North America (USA 30, Canada 13) as well as the UK (12), Ireland (9), Australia (9) and New Zealand (3). Africa is, typically for CALL, barely represented, with just one study each in South Africa, Zimbabwe, Uganda and Egypt, though it is encouraging that three of these were published in 2017 or 2018. The rest of the Americas are also underrepresented with just one study in Mexico. Conversely, just over a third (126) were conducted in Asia, especially in Taiwan (34), Japan (30), China (20 + 9 in Hong Kong), South Korea (17) and Thailand (7), with Malaysia, Indonesia, Vietnam and Singapore producing between one and three papers each. A further 62 (18%) were conducted in the Middle East, especially Iran (27) and Turkey (23), with Oman, Saudi Arabia, Yemen, Jordan, Syria and the United Arab Emirates also putting in an appearance. Europe has a good showing with 99 publications (28%), especially from countries speaking Romance (Spain 17, France 12, Italy 10, Portugal 5) and Germanic languages (Germany 6, Austria 4, the Netherlands 3). Eastern and Northern European countries include Poland with 7 studies, and between one and four each from the Czech Republic, Croatia, Denmark, Macedonia, Russia, Slovenia and Sweden. The strong representation from various parts of the world shows that DDL is not a purely European phenomenon, though the picture is somewhat different if we look only at the RMJ papers: 28% are from English-speaking countries (USA 8, Canada 2, the UK 2, Australia 1), but 13% only from Europe (the 2 from the UK plus Poland 2, and 1 each from Italy, Macedonia and Portugal). The best-represented region is Asia which accounts for fully 50% of the publications: Japan and South Korea 6 each, Taiwan 5, Hong Kong 3 and China 2, alongside 1 from Vietnam. It is notable that the Middle East is virtually unrepresented in the RMJ papers, the only ones being 3 from Iran. This may not necessarily reflect the quality or the ambition of the research but a preference to publish locally for greater impact: although 22 of the 23 papers from Turkey appear in journals, none of these are in major international sources.

Demographics

Research in language learning and CALL in general is overwhelmingly carried out in university contexts where the authors teach, and involve their own students. This is also the case with DDL, with 305 papers (87%) reporting on a university context. Of these, 86 do not specify the level of study; among the others, 138 involve undergraduates (either explicitly stated or inferred from the context), 56 target graduates enrolled on master's or doctoral

[4] Language status is a complex issue and no offence is intended if readers have different criteria; the same applies to countries as regional borders are open to different interpretations.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

programmes. The remaining few are pre-sessional (8) or involve mixed years or participants who are not students in the usual sense. Comparatively few papers look outside university: 24 (7%) in high schools at different ages, 6 in language institutes or language centres (the status of which is not always clear), 7 as continued professional training, and 3 with internet volunteers. Among the 46 RMJ articles, 83% are in university with just 4 papers in high school, 1 in a private language institute. The year of study is not specified in 11 papers at university, but there is comparatively more research at graduate than undergraduate level: 15 compared to 12. Why this should be is not clear, though the upshot is that one cannot gain a complete picture of research by reading only papers in major international journals, and there is more research in high schools and at undergraduate levels than is sometimes claimed (although this of course says nothing about the success of these studies).

Of the 252 papers that report the students' discipline, 116 (46%) are involved in language study, either the target language itself or in translation or applied linguistics, sometimes as trainee or in-service teachers. Other disciplines are split fairly evenly between STEM subjects (science, technology, engineering, mathematics), biology or medicine (42 studies, 17%) and HSS (human and social sciences such as business, management, economics, tourism, law, etc.) with 41 publications (16%). The remaining 53 (21%) report studies with groups of students from mixed disciplinary backgrounds. A similar distribution can be found in the RMJ papers: 11 (37%) in the target language or related fields, 5 each (17%) in STEM and HSS, 9 mixed (30%). The discipline is not always given, or participants are referred to simply as 'EFL learners', which probably implies mixed backgrounds in many cases. There thus seems to be a spread across disciplines, though with a comparative bias towards students majoring in language-related disciplines, which may say more about the researchers than about DDL.

Where it is possible to infer, it seems that 232 papers overall (66%) describe DDL in language for general purposes (LGP), with the rest split between LAP (59, 17%) and LSP (54, 15%). A similar proportion of the RMJ papers are for LGP (29, 63%), but there are relatively more on LAP (13, 38%), while LSP is virtually absent – just 2 papers out of the 46. Again, it is difficult to know what to make of this, other than that it may reveal the researchers' interests rather than the potential of DDL. More studies relating to LSP (mostly discipline- and genre-specific) would be welcome in major journals.

Learners' language proficiency is reported in many different ways. 11 claim or can be inferred to be beginners, 3 false beginners, 7 elementary and 17 low(er)-level learners for a total of 12% of the publications. There are substantially more at intermediate level: 33 lower intermediate, 80 intermediate with no further specification, and 66 higher intermediate (total 179, 51%), leading up to the 93 with advanced levels (26%). Though the balance is towards the upper end, there is a body of empirical DDL work at lower levels. Surprisingly, perhaps, given that DDL is supposed to adapt to each learners' own needs and queries, only 9 publications acknowledge a wide range of proficiencies within the group. The RMJ papers similarly show 1 beginner, 1 low and 4 elementary groups; 7 lower-intermediate and 4 intermediate; 8 upper-intermediate to advanced and 16 advanced – again, skewed towards higher levels, which might explain the common perception that there is little research on learners at lower levels of proficiency.

These figures must be interpreted with great caution. Proficiency is often not stated and the reporting practices vary widely in the labels used or the information given in support. Often it

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

seems to be based on the researcher's intuition alone, sometimes on a pre-test of the target feature (which is not in itself a proficiency test), or course marks, institutional expectations, the learners' age or length of study of the target language, and so on. Only rarely is some kind of internationally recognisable label used such as the Common European Framework of Reference for Languages, though it is rare that a proficiency test or certificate is provided in support. Sometimes the labels seem not to correspond to this synthesist's experience, with sample extracts or test items at levels way above or below what one would expect. Burston and Arispe (2018) examined CALL studies with 'advanced' learners and found that vagueness and imprecision meant that sizeable numbers had to be eliminated, with many overstating the level of proficiency. This is a major problem for readers and synthesists, since without robust meta-data, any conclusions rest on very shaky ground.

Aims, corpora and tools

The majority of empirical DDL studies have an overt language teaching focus rather than being entirely open-ended, though the ways this is described vary such that it is impossible to provide accurate figures. There is certainly an emphasis on 'vocabulary', a word mentioned over 4,000 times in 286 papers (81%), though it seems to be the major focus of perhaps a fifth overall. The question is, what is meant by vocabulary? This may include delexical 'light' verbs, noun phrases or verb phrases, prepositions and colligation, collocations and chunks, usage and lexicogrammar, which together seem to be addressed in about a third of the papers. As vocabulary blends into lexicogrammar, so lexicogrammar blends into grammar, with accuracy and error-correction a major concern in about a fifth of papers. This is mostly in connection with writing, which is a major focus in a fifth, while reading is far less frequent and a focus in perhaps 10 papers overall (under 3%), with translation or interpreting at around 7%. Other topics mentioned only in a tiny minority of papers include speaking and listening skills, discourse and rhetorical structure, citation practice, and so on. Similar patterns can be observed in the RMJ papers. Overall then, it seems that DDL is (or is perceived to be by the researchers involved in these studies) mainly of use for vocabulary and lexicogrammar rather than wider areas of grammar, discourse and pragmatics, and especially for writing and error correction. The absence of other areas may be that it is less suited there, or that the techniques require greater language awareness or corpus literacy (e.g. for discourse), or that there are not enough resources, especially for speaking skills – or simply a lack of imagination.

To access these language areas, the majority involved learners' direct use of a concordancer (204, 58%), even more so in the RMJ (30, 65%). This contrasts with a sizeable number that used prepared materials for consultation on paper or, occasionally, projected on to a screen (65, 19%; RMJ 6, 13%). A third option is where corpora or corpus data are integrated into some kind of wider CALL package, the case in 34 papers (10%; RMJ 3, 7%). Some adopted a mixed approach, generally beginning with paper-based materials before moving to hands-on concordancing (39, 11%; RMJ 5, 11%).

Of the 318 papers that gave some indication of the corpora used, 102 (32%) used more than one corpus, 32 had 3; 21 had 4 or more (M = 1.4). The most popular corpora are the BNC (in 75 papers) and COCA (52). Brown (16) is less common nowadays presumably because of its age, as is the Bank of English (14) as the online sampler is no longer free. While some of these are large corpora (100m words in the BNC, 560m in COCA at the time of writing), there seems little enthusiasm in DDL for mega-corpora such as ukWaC or enTenTen which

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

number in the billions of words. Surprisingly, perhaps, only 10 papers use the internet (or a website) directly as a corpus queried by a search engine such as Google. Numerous and varied attempts have been made to make corpora more amenable, notably through the use of parallel or comparable corpora (39 papers, 11%) for various language pairs, from Linguee to Sinorama, though these can be more difficult to create than other corpora. Five papers have graded readers or simplified texts, and one uses a single novel (Swallows and Amazons) combined with regular reading. Tailor-made corpora feature in at least 80 papers (23%), especially for LSP/LAP, often featuring research articles (35 papers, 10%); very little use is made of published academic corpora, just 7 for MICASE and 4 for MICUSP, 1 for BASE and 0 for BAWE, for example. Despite quite numerous suggestions early in the DDL literature, only 19 papers (5%) use learner corpora, generally from the students themselves, with none of the main published learner corpora (e.g. ICLE) featuring here; their main use seems to be for analysis of learner language rather than directly in DDL. Similarly, corpora built from student textbooks would seem to be an obvious way to bring corpora closer to the learners, but only 8 studies here adopt this approach. In 13 cases the students themselves were involved in the corpus-building process, usually for groups of students from mixed disciplinary backgrounds needing English for academic writing. Perhaps the most glaring omission though is spoken corpora: 1 each for Elisa and Backbone, with transcriptions from MICASE (7) or occasionally from subsections of the BNC or COCA, or locally compiled from movies, for example. The RMJ papers generally follow these patterns, with 11 using the BNC, 8 COCA, and 4 still using Brown which suggests that the age of the corpus is not necessarily the overriding issue, depending on the learning objectives. There seems to be no overriding impulse to use novel corpora in these potentially influential papers, with just 8 of the 46 compiling their own, 3 involving student-built corpora, 3 learner corpora, 2 simplified texts, and 3 parallel corpora.

Not all papers state the size of the corpora they use. For the hands-on use of the 248 corpora where such information is available, or can be gleaned on line from information about published corpora, the mean size is 205 million tokens, with a standard deviation of nearly 1 billion, such is the variation. Figure 2 shows an increasing number of studies favouring corpora of relatively large sizes. The smallest can number as few as 1200 words – enough though for corpus tools to be of use in querying the data. These are generally locally-built corpora which can contain up to a few hundred thousand or a few million tokens, at which point there is a drop as the compilation process becomes excessively onerous; it may also reflect the fact that local corpora tend to have specific aims which may not require large corpora. A particularly large number (116, 47% of hands-on designs) have between 100m and 1b tokens, the range that includes the numerous studies that feature the BNC and COCA. Finally, only 4 studies use corpora of over a billion words, not counting the unfathomable size of the web-as-corpus. The RMJ papers follow similar patterns but with fewer extremes – no papers below the 10k threshold or above 1b words. The average size is 131m tokens (SD = 197m, revealing of the tremendous variation), with the biggest chunk (15) again in the 100m range, 11 between 1m and 100m, and 7 below 100k tokens.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

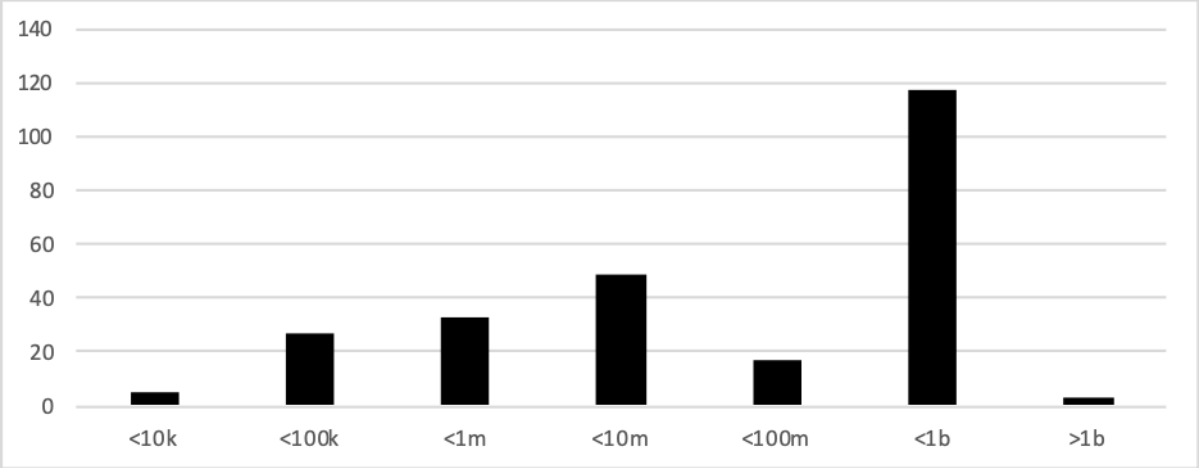

Figure 2. Corpus size in all publications

The corpora influence the software used in hands-on concordancing. Those featuring the BYU corpora (notably COCA and the BNC) use the BYU interface in 62 publications (18% overall; RMJ 11, 24%); other BNC interfaces are used in only 12 cases including 2 in the RMJ papers. Other integrated interfaces are used for MICUSP and MICASE in particular in 7 papers (RMJ 2). Stand-alone tools include the online LexTutor (19, 5%; RMJ 4, 9%), and the offline AntConc (27, 8%; RMJ 7, 15%) and WordSmith Tools (24, 7%). Curiously, perhaps, WordSmith Tools is not used in any RMJ paper, perhaps reflecting a preference for free tools; SketchEngine, for example, is only used in 5 of the 351 publications, while SkELL (a free version specifically for learners of English) has yet to appear. Other tools used in earlier studies include MicroConcord (7), MonoConc (5), ParaConc (8) and VLC Media Player (8); some of these appear almost exclusively in papers by one author.

Design

Of all 351 publications, 13 do not state the number of participants. Those that do involve 17,492 participants, 77% of them in experimental groups (Figure 3). The mean number of participants in experimental groups is 40, compared to 76 in the RMJ papers, partly because some studies also feature more than one experimental group, but otherwise suggesting that smaller studies are not considered appropriate for major journals, whether by the authors or by the editors. The high standard deviation (62) again reveals considerable variation. So while it is certainly true that there is quite a lot of small-scale DDL research, particularly for in-depth analysis of how they use corpora or attitudes towards them (4 case studies have only one learner experiencing DDL, 24 have no more than 5), there are also a number of quite large-scale studies. Over half however fall in the middle in the expected range for group sizes at university with between 15 and 35 students.

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

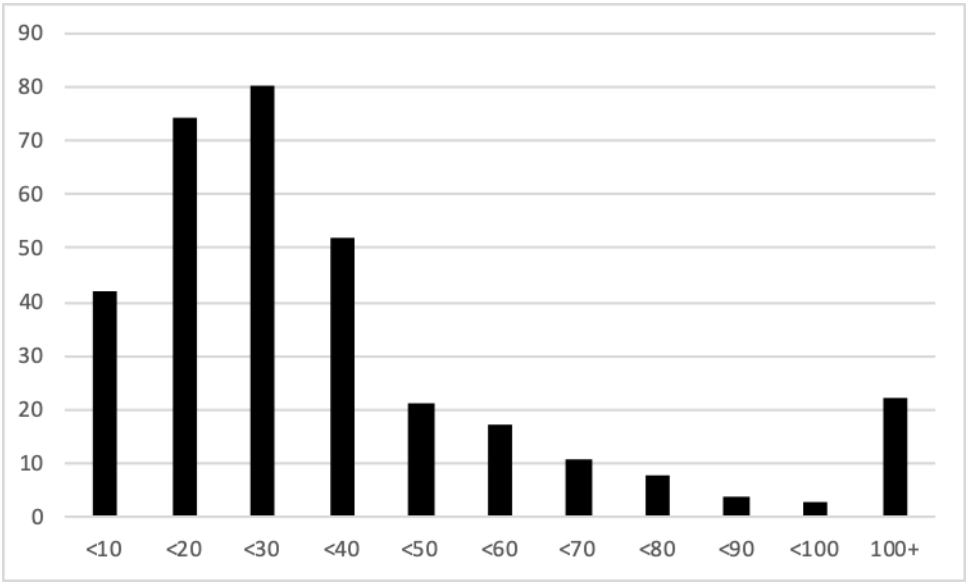

Figure 3. Experimental participants

Only 121 of the publications overall (34%; RMJ 16, 35%) have control groups; the others are either pre/post-test designs or involve qualitative data on the approach. The objectives reflect this overall, with learners' representations collected mainly by questionnaire or interview in over half the papers (185, 53%; RMJ 33, 72%). These instruments are often used alongside observations, think-aloud protocols, diaries, retrospective reports, videos, screen recordings, tracking, etc. to collect data about behaviour in 93 publications (26%; RMJ 12, 26%). In analysing outcomes, corpora are used as a learning aid in 148 papers (42%; RMJ 17, 37%) and as a reference tool in 89 (25%; RMJ 15, 33%). The studies generally compare pre/post or control/experimental designs (only 24 featured delayed post-tests, 3 in the RMJ papers), often with very constrained responses such as gap-fills or multiple-choice questions, though a small number involve detailed analysis of (written) production. These figures indicate that most publications use multiple instruments for more than one type of analysis. 56 (16%; RMJ 8, 17%) provide a purely qualitative analysis; 124 (35%; RMJ 11, 24%) give raw numbers of percentages only, while the remaining 171 (49%; RMJ 27, 59%) provide some kind of statistical analysis, usually of outcomes of corpus use though also of questionnaires and other data types.

The duration of the studies is variously given. For the 197 where duration is stated or can be inferred in minutes or hours, the mean is 12'51" (RMJ 13'51"), but the high standard deviation (15'39"; RMJ 15'24") again shows considerable variation (Figure 4). The shortest studies are lab-like experiments: overall, 75 (21%) lasted under 5 hours, of which 19 less than an hour, the shortest being just 10 minutes. 10-30 hours typically involves a weekly session of an hour or two over a semester, though it is very rare to be explicitly told how much of this was given over to DDL – one study, for example, does mention that it used DDL for just 10 minutes per session. This picture underestimates the reality, since longer studies, which generally favour a more ecological approach, do not always give an idea of time in minutes or hours. 48 papers mention only the number of sessions: 29 of those occur during a single class for an average of 3.13 sessions (SD = 4.16). 45 publications measure the duration in weeks (M = 9.2, SD = 6.2), 5 in months (M = 4.4, SD = 2.1) and 33 in semesters (M = 1.6, SD =

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

1.8). Among the RMJ papers, 7 give the number of sessions only (M = 3.3, SD = 3.2); 7 mention only weeks (M = , SD = 6.1), and only 1 lasts a semester.

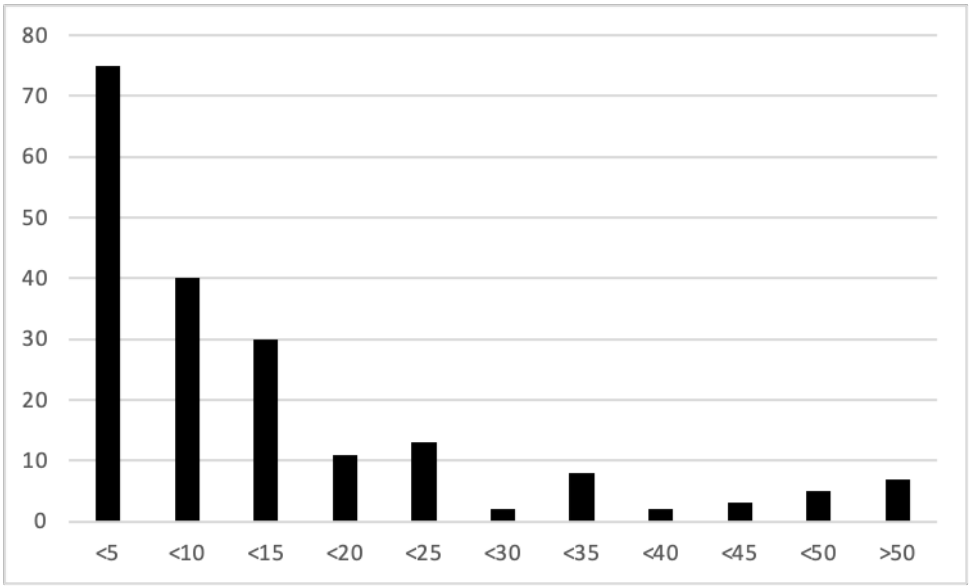

Figure 4. Duration in hours

Discussion and conclusion

This chapter is based on a rigorous, ongoing trawl of 351 separate publications which seek to empirically evaluate some aspect of DDL over the last three decades. The inclusion criteria cover 'grey' literature from less well-known journals, book chapters and conference proceedings in particular, but exclude student dissertations and slides, notes or other incomplete texts, as well as studies published in languages other than English. This relatively large number is increasing over time, suggesting that DDL is a healthy field of research, although it is not the objective here to examine the quality of the research or the outcomes. It is notable however that the reporting practices are frequently incomplete, with essential information such as the level of proficiency, duration, corpora and tools used, etc., missing or inadequately described in both minor and major publications – a real problem for the synthesist and reader. The entire body of research is contrasted with 46 papers from major recent journals published in the last 5 years, thus allowing a comparison of the field as a whole against likely influential papers in the near future. In some cases, the two collections mirror each other; in others there are notable differences, since the 15 countries where research has been conducted in the RMJ cannot reflect the variety of work in the 41 countries in the corpus as a whole. In particular, the 62 studies in Middle Eastern countries are represented by just 3 papers from Iran in RMJ, with the 23 conducted in Turkey disappearing entirely.

Overall, it is clear that there is comparatively little research on languages other than English, with most of the rest targeting European languages. This is perhaps unsurprising given the global demand for English today, especially in a university context, though it may also reflect

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins.
https://doi.org/10.1075/scl.102.01bou

the availability of resources (corpora and tools) or awareness of DDL; it could even be that English is particularly suited to such an approach. In language teaching research in general, it is usual for researchers / authors to use their own students as the object of their study, so it is unsurprising that most studies do occur in university contexts. This means that there is a rich area waiting to be explored in secondary education (just 7% of existing research) and beyond, in continuing education and in-company training, and in out-of-class contexts in general – especially as DDL is promoted as fostering autonomy over the long term. Overall there are 33 studies with lower-intermediate participants and 27 at still earlier stages, though participants in the RMJ papers tend to be at higher levels of proficiency. Nor is DDL just for students majoring in languages, which account for 46% of all publications, the rest being split between STEM and other HSS fields.

In terms of objectives, two thirds of the publications are concerned with language for general purposes, with EAP being relatively more frequent than ESP among the RMJ papers. The focus does seem to be on vocabulary and lexicogrammar as is often claimed, favouring production over reception, especially for usage and error-correction. There is very little research to date on DDL for speaking or listening, with some work on purely transcribed data but virtually none featuring texts aligned with audio or video. The corpora themselves vary from just over a thousand words to several billion, though both extremes are rare. The larger corpora in the range of 100m words or more tend to be popular public corpora such as COCA or the BNC, while smaller corpora up to a few million words are usually locally compiled for specific purposes. There is little use overall of novel corpus types – from parallel corpora to learner corpora, graded readers and single long texts, the web-as-corpus, etc. Smaller local corpora tend to be queried via downloadable software such as AntConc, with online interfaces such as LexTutor less frequent except for published online corpora which often have integrated interfaces. There is currently very little research featuring paying sites such as SketchEngine (which may be a good thing if learners are to continue after class), or even its related SkELL; at the same time, many local corpora or tools are unavailable beyond the institution – a different kind of barrier. The majority of work to date requires participants to interact with the software, though 30% of all studies make at least some use of printed materials, either exclusively or as a way in to hands-on corpus consultation, and another 10% integrate corpus data into some kind of CALL package.

In terms of design, there does seem to be some truth in the claim that there are many small-scale, qualitative studies. However, the average number of experimental DDL participants seems to be larger in the RMJ papers, and there are substantial numbers of quantitative studies that seek to test if, where, and to what extent DDL may bring measurable benefits. One problem here is that many test instruments are highly constrained and artificial; more work on user-generated queries is crucial. Longer, experimental studies are present, but the methodological design does not sufficiently cater for this: longitudinal and delayed post-tests are essential if we are to test the claim that DDL leads to long-term learning, along with other alleged benefits such as increased autonomy, noticing and language awareness, and that the skills acquired can be transferred to new, self-initiated areas. Many studies do not include control groups, but this is understandable in qualitative studies which can provide a more detailed view of the uses learners make of corpus resources. And since meta-analyses have shown that DDL does have an effect, there is a strong case today for abandoning 'control' groups in favour of comparing different types of intervention; this might involve direct comparison of different corpora or tools, different user profiles or preferences, populations in

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins. https://doi.org/10.1075/scl.102.01bou

different institutions or countries, different types of DDL, varying amounts of training or scaffolding, and so on. More genuinely mixed-methods studies would also be hugely valuable. At the most basic level, improved reporting is essential for nearly all the categories coded here, notably proficiency and duration. Without this, studies cannot be properly interpreted or compared. This message is not limited to DDL; Plonsky and Ziegler (2016) in particular find that CALL as a whole lacks methodological rigour and transparency.

The main observation must be that there is tremendous variation at all levels – in the participants and the contexts, in the corpora and tools, in the pedagogical goals and procedures, in the study instruments and designs. This can only be a good thing since one of the selling points of DDL is its flexibility – there is no single 'best' version of the approach which should apply to all. However, as we have seen, there is plenty of room for greater originality rather than repeating existing work with minor variations, partly due to ignorance of large quantities of work leading to reinventing the wheel (Gillespie 2020: 128). In particular, there is a strategic imperative for future research to feature languages other than English; novel corpora especially for speaking, but also the potential of learner corpora as input for DDL, or graded/simplified texts or parallel corpora; free resources (as opposed to producing and testing one's own corpora or software), including even the web-as-corpus; DDL for mobile tools; 'regular' teachers who are not the researchers/authors for greater integration into existing courses; and outside university classrooms, whether during a course or at some later time. The rarity of such studies seems to reflect a lack of imagination as researchers focus on what DDL can bring to their local situation rather than asking what research is needed and how they can go further, essential considerations if the field is to progress.

References

Abu Alshaar, A., & Abuseileek, A. F. (2013). Using concordancing and word processing to improve EFL graduate students' written English. *JALT CALL Journal, 9*(1), 59–77. https://doi.org/10.29140/jaltcall.v9n1.148

Al-Gamal, A. A. M., & Mohammed Ali, E. A. M. (2019). Corpus-based method in language learning and teaching. *International Journal of Research and Analytical Reviews, 6*(2), 473–476.

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal, 63*, 23–32. https://doi.org/10.1093/elt/ccn011

An, X.-H., & Xu, M.-Y. (2013). An empirical research on DDL in L2 writing. *US-China Education Review A, 3*(9), 693–701.

Anthony, L. (2019). *AntConc* [version 3.5.8m]. Waseda University. https://www.laurenceanthony.net/software

Bax, S. (2003). CALL: Past, present and future. *System, 311*, 13–28. https://doi.org/10.1016/S0346-251X(02)00071-4

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225–234). Peter Lang.

Boulton, A. (2008). But where's the proof? The need for empirical evidence for data-driven learning. In M. Edwardes (Ed.), *Technology, ideology and practice in applied linguistics*

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins. https://doi.org/10.1075/scl.102.01bou

(pp. 13–16). Scitsiugnil Press. https://www.baal.org.uk/wp-content/uploads/2017/12/proceedings_07_full.pdf

Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics, 35*(1), 81–106.

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). Equinox.

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

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and applications* (pp. 261–291). John Benjamins. https://doi.org/10.1075/scl.52.11bou

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). John Benjamins. https://doi.org/10.1075/scl.69.13bou

Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching, 50*(4), 483–506. https://doi.org/10.1017/S0261444817000167

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

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 3–16). Rodopi. https://doi.org/10.1163/9789401203906

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching, 52*(4), 460–475. https://doi.org/10.1017/S0261444819000089

Chambers, A., & Bax, S. (2006). Making CALL work: Towards normalisation. *System, 344*, 465–479. https://doi.org/10.1016/j.system.2006.08.001

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics, 23*(3), 335–369. https://doi.org/10.1075/ijcl.16130.che

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of English corpus linguistics* (pp. 478–497). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.027

Cresswell, A. (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* (pp. 267–287). Rodopi. https://doi.org/10.1163/9789401203906_018

Crosthwaite, P., & Stell, A. (2019). It helps me get ideas on how to use my words: Primary school students' initial reactions to corpus use in a private tutoring setting. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 150-170). Routledge. https://doi.org/10.4324/9780429425899-9

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). John Benjamins. https://doi.org/10.1075/scl.69.02flo

Gilquin, G. & S. Granger. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *Routledge handbook of corpus linguistics* (pp. 359–370). Routledge. https://doi.org/10.4324/9780203856949.ch26

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins. https://doi.org/10.1075/scl.102.01bou

Han, Z. (2015). Striving for complementarity between narrative and meta-analytic reviews. *Applied Linguistics, 36*(3), 409–415. https://doi.org/10.1093/applin/amv026

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Higgins, J., & Johns, T. (1984). *Computers in language learning*. Collins.

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

Johns, T. (1988). Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe & H. Wekker (Eds.), *Computer applications in language learning* (pp. 9–27). Foris. https://doi.org/10.1515/9783110884876-003

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 samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom concordancing. English Language Research Journal, 4*, 1–16.

Johns, T. (1993). Data-driven learning: An update. *TELL&CALL, 2*, 4–10.

Johns, T., & King, P. (1991). Editors' preface. In T. Johns & P. King (Eds.), *Classroom concordancing. English Language Research Journal, 4*, iii–iv.

Johns, T., Lee, H., & Wang, L. (2008). Integrating corpus-based CALL programs and teaching English through children's literature. *Computer Assisted Language Learning, 21*(5), 483–506. https://doi.org/10.1080/09588220802448006

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics, 40*(5), 721–753. https://doi.org/10.1093/applin/amy012

Luo, Q. (2016). The effects of data-driven learning activities on EFL learners' writing development. *SpringerPlus, 5*, n.p. https://doi.org/10.1186/s40064-016-2935-5

Ma, B. (1993). Small-corpora concordancing in ESL teaching and learning. *Hong Kong Papers in Linguistics and Language Teaching, 16*, 11–30.

McEnery, T., & Wilson, A. (1997). Teaching and language corpora (TALC). *ReCALL, 9*(1), 5–14. https://doi.org/10.1017/S0958344000004572

McKay, S. (1980). Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly, 14*(1), 17–26. https://doi.org/10.2307/3586805

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies, 22*, 1–18.

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2019.1667832

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Ziegler, N. (2016). The CALL–SLA interface: Insights from a second-order synthesis. *Language Learning & Technology, 20*(2), 17–37. https://doi.org/10125/44459

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–225. https://doi.org/10.1017/S0267190511000055

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning, 63*(2), 296–329. https://doi.org/10.1111/lang.12001

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

Pre-publication version; where possible, please refer to the final published text:
Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9–34). John Benjamins. https://doi.org/10.1075/scl.102.01bou

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

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772