DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning.

Peter Crosthwaite, Alex Boulton

Crosthwaite, P., & Boulton, A. (in press). DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning. In H. Tyne, M. Bilger, L. Buscail, M. Leray, N. Curry & C. Pérez-Sabater (Dir.), *Discovering language: Learning and affordance*. Peter Lang.

DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning

Peter Crosthwaite and Alex Boulton

Introduction

30 years since Tim Johns (1990) coined the term 'data-driven learning' (DDL) to describe direct engagement with language corpus data, tools and techniques for pedagogical purposes, DDL has become a popular area of applied linguistics research, spawning several recent syntheses, surveys and meta-analyses into its effectiveness for language learning across a wide range of teaching and learning contexts (Boulton & Cobb, 2017; Chen & J. Flowerdew, 2018; H. Lee et al., 2019; Pérez-Paredes, 2019; Boulton & Vyatkina, 2021; Boulton, in press, among numerous others). The results of these syntheses have suggested that DDL leads to a number of empirically-proven benefits for learning (at least under pre/post-test experimental conditions, Boulton & Cobb, 2017), notably for vocabulary (H. Lee et al., 2019), English for academic purposes (Chen & J. Flowerdew, 2018), and L2 writing (Pérez-Paredes, 2019). In a synthesis of these syntheses, O'Keeffe (2020, p. 2) notes each has pointed toward the value of DDL, together with an "undying enthusiasm" about DDL as an aid to learning, and an "aspiration that [DDL] should become more mainstream."

So, if everything is going as well as the syntheses suggest, then why hasn't DDL yet achieved these grand aspirations? In our second order synthesis of these syntheses, the findings so far point to the following:

1. A lack of a real definition of the theoretical underpinnings of DDL, e.g., determining DDL's place within weak/strong interface theories on implicit/explicit learning (L. Flowerdew, 2015; O'Keeffe, 2020). In line with much CALL and indeed applied linguistics research, theoretical bases are often missing, or at best seem to be tacked on *post facto* to frame and justify the study.
2. A lack of DDL studies focusing on the impact of DDL at the cognitive level, alongside purported (but as yet rarely directly tested) improvements to learner autonomy, language awareness, noticing, etc. (O'Keeffe, 2020; Boulton & Vyatkina, 2021). Though such things are clearly difficult to test directly, they should not be impossible.
3. A general lack of information on the training provided in the use of corpus tools and integration of DDL to syllabus design, alongside a prevalence of studies with the researcher as the main stakeholder rather than teachers, or indeed the students or institutions (Pérez-Paredes, 2019). This is a niche group since nearly all researchers in language education are teachers, but very few teachers are researchers.
4. Overreliance on concordancing and frequency information from public or locally-compiled corpora at the expense of other software, data and corpus types (Pérez-Paredes, 2019; Boulton & Vyatkina, 2021). If DDL is to prove beneficial in the long term, students need access to free, stable, and relevant tools and language that can be used for a variety of future purposes after the end of the course and, indeed, the educational programme.
5. A need for DDL studies with learners other than tertiary students, especially for English (Boulton & Cobb, 2017; Pérez-Paredes, 2019). This includes younger learners at primary/secondary levels, but also outside initial education and in professional contexts,

and further work is needed to see if English is a special case or if DDL is appropriate for other languages.

6. A lack of methodological rigour in data collection, analysis, instruments and methods preventing replication, often coupled with poor reporting practices for methodology including context, proficiency, duration, and what the learners actually did (Boulton & Vyatkina, 2021). However, this is not exclusive to DDL but applies to SLA/applied linguistics as a whole (Plonsky, e.g. 2014).

While we recognise the value of the previous thirty years' research in getting us where we are today, there is a suspicion that research, zombie-like, is covering well-trod paths rather than breaking substantial new ground. This leaves DDL in danger of never truly breaking into the mainstream. In response, we present a number of useful DDL studies and initiatives that have the potential to 'reanimate' the field, demonstrating the kind of outside-the-box thinking required if DDL is to remain a viable pedagogical approach over the next 30 years.

Theory and DDL: (Still) never the twain shall meet?

The question of the (lack of) theoretical underpinnings of DDL has been discussed in the literature for some time now, with L. Flowerdew's (2015) treatise an influential initial attempt at getting the ball rolling: "Data-driven learning and language learning theories: Wither the twain shall meet." But the ball rolls slowly: only 11 of 209 recent journal papers on DDL include *theor** in the abstract (all given in Figure 1), much less take this as a genuine starting point for the study itself. As Boulton and Vyatkina (2021, p.17) note, "an approach survives on its merits and works because it works (or doesn't), regardless of theory – but this does leave us rather hungry for more". The impression we are left with is that many studies begin with a class and an objective, fill it in with some DDL and then write up what happened.

Figure 1.

1. of the study can be used as a framework for further **theoretical** and empirical research into the effect of the corpus-
2. ising from them. This paper concludes by discussing **theoretical** and pedagogical implications of the findings. The p
3. heir writing. The study concludes by discussing the **theoretical** and pedagogical implications of the results. Recent
4. g was enhanced by corpus-based instruction. Some **theoretical** and pedagogical implications of the study were then
5. e definition of probability, application prospect and **theoretical** explanation, this paper expatiates on the process of
6. riven learning method. This paper provides certain **theoretical** guidance to the improvement of the English writing
7. encies of the student. The purpose of this study is to **theoretically** substantiate and empirically confirm the effective
8. act, DDL good practices perfectly align with current **theories** and practices of SLA, namely the constructivist and le
9. f corpus-based learning tools. The gap between the **theories** of second language acquisition and their practice in m
10. pproach based on post-positivism underpinning the **theory** of noticing hypothesis (Flowerdew, 2015) has been appli
11. ain Findings: This study is conducted based on the **theory** of source-based teaching, while the process of utilising t

Most DDL studies point to 'constructivism' as the main learning paradigm under which DDL operates, where engagement with corpus data results in students' consolidation of multiple forms of information to arrive at data-driven conclusions when testing their hypotheses about language use (Cobb, 1999). However, as O'Keeffe (2020) rightly points out, this focus on constructivist learning at the individual level has come at the expense of other paradigms under which learning may occur, such as the sociocultural benefits of DDL when implemented in actual classroom practice and aided with peer- or teacher-provided scaffolding during focus-on-form(s) activities. P. Lee and Lin (2019) also note that constructivist DDL requires considerable inductive reasoning

skills, which students raised in educational systems favouring deductive, teacher-led reasoning may struggle with and even personally reject. Comparing both approaches, the researchers found no significant difference between inductive and deductive learning approaches for DDL; and although each treatment only lasted two weeks, their finding emphasises the need to look beyond constructivism as the sole explanation of how learning takes place during DDL.

At the cognitive level, it tends to be assumed, rather than tested, that DDL encourages noticing and pattern-recognition, exemplar-driven learning, acquisition through frequency and statistical information, recognition of the psycholinguistic reality of chunks and collocations, input flood, deeper cognitive processing, and so on. Positive results from various studies are taken to confirm this, but the evidence is generally indirect rather than the focus of the research itself. For example, one of the more popular paradigms in DDL is Schmidt's (1990) 'noticing' hypothesis: typically in these studies, corpus consultation is successful, therefore noticing has happened, and this is taken as evidence that DDL fits perfectly. The 'noticing' is often presumed to happen at some point during corpus consultation (e.g. Saeedaktar et al., 2020), often as the learner engages with concordance information in KWIC view (keyword in context), centred around the search term for ease of identification of patterns, as in Figure 1. KWIC view, as well as other forms of input enhancement made possible via the particular corpus tool used (e.g. colouring/highlighting/hyperlinking of the target word), is claimed to make the input more salient, thus aiding the noticing required for meaningful processing of the input to occur (Van Patten & Benati, 2010).

The link between DDL and noticing is often positioned together with theories of usage-based learning, with DDL helping learners to internalise statistical information about language in use in the form of frequency and collocation information from corpus data (Ellis et al., 2016; see Pérez-Paredes et al., 2020, for detailed discussion of its application for DDL). P. Lee et al. (2019) also suggest that as learner engagement with corpora requires a higher 'involvement load' (Laufer & Hulstjin, 2001), the conditions for vocabulary learning are improved under DDL. However, despite the claim that DDL is "well placed" to "lead to cutting-edge insights into the cognitive processes of language learning" (O'Keeffe, 2020, p. 2), as noted in Pérez-Paredes' (2019) survey, the development of actual cognitive abilities arising from DDL is "not represented in the body of research examined" (p. 16). O'Keeffe (2020) also notes this problem, stating that while DDL is "likely" to lead to improvements in a range of cognitive processes associated with constructivist learning, the link "has seldom been tested" (p. 3). There are a number of reasons for this, including a relative dearth of follow-up studies determining if DDL has led to sustained learning outcomes (only 30% of the studies reported in H. Lee et al., 2019, contained delayed post-test data), as well as almost no studies determining if improved learning practices post-DDL have been realised by the same learners in other language-related or even non-language-related domains.

More generally, it seems intuitive that DDL *ought* to lead to more effective learning (or even 'better learners'), but the relative lack of longitudinal research designs leaves this as yet empirically unverified. This is not a trivial point, given the time-consuming nature of DDL in the early stages: if there are no substantial long-term benefits, then the entire enterprise is based on very shaky foundations. Mostly, studies with delayed post-tests show the expected result – that scores decrease subsequent to the immediate post-test but are still higher than the pre-test – but the

delay in these cases is short, typically a week or two. This is understandable where the test items are few in number, as one would not necessarily expect learners to remember individual words, structures or usage directly based on a short-term encounter from DDL or any other approach. What is needed are studies that look at more general trends, such as improved critical thinking or language awareness, among others. The first (accidental) indication that this may be the case comes from Johns et al. (2008), where the DDL group not only scored higher on the target items but also fared significantly better in their end-of-term exams on different content. The first attempt to examine such knock-on effects directly (Boulton, 2011) involved comparing DDL and control groups working with an unseen text at the end of a semester simply to see if they had noticed certain features that were unrelated to their teaching. The results were encouraging, but this type of study is in desperate need of follow up.

Focus on practice: What is happening in DDL training?

In many of the reported studies to date, information on the form, approach and content of the DDL training conducted with students and/or trainee teachers is still lacking. One issue is that many studies feature small participant samples and short durations (Boulton & Vyatkina, 2021), although this is typical of CALL as a whole (Gillespie, 2020). Another issue is that many studies also often fail to appropriately describe the actual DDL training regimen that took place, with vague descriptions such as "a training session was given to the participants, with a video tutorial demonstrating how to consult COCA" (Tsai, 2019, p. 812). Information on exactly what students were asked to do, how they engaged with the selected resources, how they interacted with others to achieve lesson objectives, and what the teachers' role in all this was, is often incomplete or, worse, completely unavailable. Similarly, information on the duration of DDL treatment may be reported in sessions, minutes, hours, weeks, months, semesters or years, making comparison difficult. For example, a semester's course may feature two three-hour classes over 15 weeks given over entirely to DDL, or just a few minutes in occasional classes over a handful of weeks – a tremendous difference. Without knowing what took place, we are therefore no closer to finding out what really 'works', and what doesn't.

Another issue is the difficulty involved in tracking what learners are actually doing during DDL, with most attempts limited to observation or self-report via questionnaires or interviews. A very small number of more ambitious studies (e.g. Pérez-Paredes et al., 2011; Crosthwaite et al., 2019), have tracked offline measures including corpus query syntax, though DDL-specific studies featuring online real-time measures such as eye-tracking, keystroke-logging or fMRI procedures are currently unavailable.

From the students' perspective, Johns was among the earliest to suggest specific procedures with "identify, classify, hypothesise" (1991, p. 4), or, later, "research, practice, improvise" (1997, p. 101). These may not sound particularly radical today, but they contrasted at the time with the traditional "three 'P's" of presentation, practice, production. Perhaps better known within DDL are McEnery et al.'s (2006, p. 99) "three 'I's" of "illustration" (observing real data), "interaction" (discussing) and "induction" (coming up with one's own rule or generalisation, in line with constructivism). L. Flowerdew (2009, p. 407) proposes an optional fourth 'I' before the final stage: "intervention" where the teacher helps the students along. However, these schema seem to assume

that practice will happen if and when necessary. Other notable studies providing detailed information on classroom practices for DDL include Kennedy and Miceli (2001, 2010, 2017), whose description of 'pattern-hunting' and 'pattern-refining' strategies for DDL instruction have resulted in a wealth of later studies adopting these techniques (e.g. Y.-J. A.Wu, 2021). Quinn (2015) provides a detailed 6-stage approach to EFL students' corpus training for improving L2 writing. This involves introducing corpora as a concept at stage 1, explaining the reasoning why a corpus should be used at stage 2, paper-based concordancing at stage 3, online corpus practice at stage 4, responding to written corrective feedback via corpora at stage 5, and making revisions to student texts at stage 6. Full descriptions of how this plan was operationalised are provided in the paper, including the steps leading to corpus consultation and a description of how knowledge was consolidated through DDL from an erroneous sentence in a student's production.

Moreover, while many studies compare DDL against a 'control' group (often corresponding to induction vs deduction), Tsai (2019) compared the two with both groups using DDL: the inductive group consulted a corpus before moving to a dictionary, while the deductive group followed the inverse procedure – in other words, using a dictionary to confirm corpus findings or vice versa. Both were found to be productive, though the corpus-first group were more attentive to usage and collocation, while the dictionary-first group favoured definitional meaning. This type of development is important not just for its immediate results but in terms of design, as the 'DDL vs traditional teaching' construct has already provided such quantities of studies that further comparison is barely warranted today: more relevant is to compare different types of DDL, or different tools or corpora for different purposes, or different populations, or different degrees of scaffolding, etc.

From the teachers' perspective, information on how DDL was introduced to them, and how it was integrated (or not) into their classroom practices is also crucial if the field is to develop expertise (and popularity) in this area, although this data is often missing from DDL studies. That said, it seems likely that millions of learners around the world are using Google to search the web in ways not entirely dissimilar to concordancer and corpus; surveys would not be difficult to envisage, but, unfortunately, have yet to materialise, and this remains to be explored in more detail. Han and Shin (2017) produced one of the few research attempts to put this on a formal footing: the students were receptive but produced mixed results, the recommendation being for more guidance, especially at lower levels. That said, a number of recent studies are beginning to reveal new insights into how DDL can be incorporated into teacher training. Examples include Schaeffer-Lacroix (2019), who provides a detailed description of the successes and failures regarding the training of French teachers of L2 German to develop a corpus-based language learning activity. The study notes 'first-order' barriers related to trainees' technical abilities, accompanied by 'second-order' (and more serious) barriers related to trainees' conceptions of corpora and DDL as a viable pedagogical approach. In particular, her study discusses the various in-class negotiations between trainees that lead some to success and others to failure. Leńko-Szymańska (2017) analysed 53 corpus-based projects submitted by teacher trainees following DDL training, determining whether the trainees had acquired the technical skills required to perform their own corpus analyses, whether these skills were sufficient to inform their teaching, and whether they had acquired the pedagogical skills to enable to exploit corpora in the classroom. Despite demonstrating mastery of the corpus software for their own use, the trainees were reported as still lacking the skills required to integrate

their use within classroom practice. Ma et al. (2021) is another useful study in this area, presenting a DDL-focused teacher training intervention documenting changes in teachers' pedagogical content knowledge (PCK – Shulman, 1986) from *comprehension* of subject matter and corpora/corpus tools, to *transformation* of pedagogical practice including preparation of materials, selection of tools, and adaptation of these to meet students' characteristics and needs. It is important to frame DDL training within established theories of teacher knowledge in education such as PCK as well as Technological, Pedagogical and Content Knowledge (TPACK – Kohler and Mishra, 2009; see Meunier, 2019 for a description of its potential for DDL; Crosthwaite et al., forthcoming, for a practical application). Doing so should help to increase 'buy-in' from those within the education field who have the means and leverage to engage with schools and other institutions to embed such training within a wider range of professional development contexts within and outside of applied linguistics/CALL.

An additional and equally important problem is that in each of the positive teacher training interventions reported above, the corpus linguist could be said to be the main stakeholder in the introduction (and the eventual success) of DDL within a given teaching and learning context, in line with Pérez-Paredes' (2019) and Chambers' (2019) criticism of the field. In particular, Chambers' (2019) study reveals a 'research-practice gap' wherein language teachers who are not researchers may find it difficult to replicate a given corpus intervention when the applied linguist has left the building. Viana and Lu (2021, p. 1485) also note the value of engagement with corpora through continuing professional development projects in "democratizing access" to DDL for both language-oriented and non-language oriented academics/professionals alike. Essentially, we need to expand the target audience for DDL beyond tertiary (language) learners and their academics to include a more diverse audience, investigating DDL for professional practice, private tutoring, language schools, business and more, as and when the actual need arises. We could go even further by encouraging teachers of other disciplines to use corpus tools when dealing with any type of electronic text for any purpose (Adolphs, 2006). In this way, the tools might have a chance of becoming a reflex go-to resource rather than a specialist tool that is returned to the metaphorical dusty shelf when the immediate need is overcome, never to be thought of again. So far though there have been vanishingly few follow-up studies looking at continued uptake months or years after a course. One exception is Charles (2014) who waited a year to contact former participants from a post-graduate course who had compiled their own discipline-specific corpora for academic writing. 40 of 103 responded: 70% continued to use their corpus, with more intending to when the need arose, and others at least used published corpora such as COCA. In total, 86% still used corpora in one way or another (38% regularly) for writing and/or revising their academic writing – an encouraging finding, but one that needs replicating for other populations, tools, purposes, etc.

One very promising direction lies in the use of Open Educational Resources (OERs) for DDL, where structured activities built around specific corpus applications for specific corpora, and suitable for non-experts in corpus linguistics, can be shared freely online for teacher/student DDL training. As Vyatkina (2020, p.364) notes, training materials "usually take the form of stand-alone .pdf files [meaning that teachers] must go back and forth between reading these materials and searching corpora online"; more tailored, specific resources must therefore be made available to professionalise any training that is to take place. Useful examples include Crosthwaite's (2020)

short private online course *Improving Writing through Corpora*[1] taking participants through the basics of corpus queries and interpreting corpus data using the SKELL[2] and SketchEngine[3] online tools. Over 800 participants have taken the course to date across 65 countries, while Portuguese and Mandarin-language translations of the course content have already been completed. The course has already been used for teacher training purposes in Australia and Indonesia (Crosthwaite et al., 2021). The project *Incorporating corpora*[4] mentioned in Vyatkina (2020) boasts a suite of OERs dedicated to DDL for L2 German. Vincent and Nesi's (2018) *Quicklinks*[5] platform provides a suite of language learning activities targeting common issues with L2 academic writing with accompanying pre-selected URL links to concordances within the BAWE corpus. Le Foll (2021) has created an OER e-book *Creating corpus-informed materials for the English as a foreign language classroom*[6] which contains materials developed by and for teacher trainees, again using freely available corpus resources, with materials for primary, secondary and tertiary language learners. In each case, these materials aim to reduce the need for a corpus linguist to deliver DDL training, although of course it will take greater promotion of these resources within and outside of the corpus linguistics community if teachers and institutions are to realise their potential.

Despite improved software, is the field taking advantage?

One area where there has been undoubted improvement over the last decade is the availability of accessible corpora, user-friendly corpus query tools, and even online tools allowing for DIY corpora to be uploaded for processing, accompanied by a range of query and visualisation options (e.g. *Voyant Tools* – Sinclair & Rockwell, 2016[7]). However, the majority of DDL studies so far still report using tools that, while excellent *research* applications, may be poor choices for language teaching and learning, or for 'ordinary' users. This may be due to one or a combination of the following:

- The level of technical knowledge needed to use the tool (for DDL).
- Complex log-in procedures or an expensive license to use (see Chen & J. Flowerdew, 2018, for a detailed critique).
- Complicated or unintuitive user interfaces at odds with how modern learners typically access digital information through resources such as Google.
- Unsuitable corpus data for the target learners (e.g. COCA for younger or less proficient L2 learners), although this problem is often attributable to the researcher's choice rather than the data itself.
- An overwhelming focus on concordances as the sole mode of input and data visualisation.

[1] https://edge.edx.org/courses/course-v1:UQx+SLATx+2019/about
[2] https://skell.sketchengine.eu/
[3] https://app.sketchengine.eu/#open
[4] https://corpora.ku.edu/
[5] https://bawequicklinks.coventry.domains/
[6] https://elenlefoll.pressbooks.com/
[7] https://voyant-tools.org/

By way of example, Boulton and Vyatkina (2021) found that 31% of all 489 DDL studies they found up to and including 2019 used the BYU suite (now English-Corpora.org[8]) especially for COCA and the BNC; the site claims 130,000+ non-researchers using these corpora every month. While excellent tools in themselves, they could be said to require a degree of technical knowledge to use efficiently, while also requiring a log-in beyond a certain number of queries, with regular freezes to encourage payment. Attempts have been made to introduce new options that are relevant for learners, with each search result providing external links to dictionaries, translations, Google searches, images, pronunciations, etc. The site also allows users to input their own text, with different types of information grouped on the same page for any word that is then clicked on: relative frequency and distribution in different registers, synonyms, typical topics, collocates of different parts of speech, clusters and a KWIC presentation – again, with external links for the target item directly in *YouGlish*[9] (video extracts from YouTube aligned against subtitles) and *Linguee*[10] (parallel texts resulting from mainly human translations). Also in Boulton and Vyatkina's survey, 41% of all studies used at least one self-compiled or local corpus, and some their own software (though this tendency is decreasing as more resources become publicly available). The advantage is specificity, but their lack of availability outside the class or institution may prevent continued use or replication elsewhere. Importantly, students often report that while they feel such DDL tools are useful within the study period, they may not necessarily be willing (or even able) to replace their current digital language learning tools (e.g. online dictionaries or translation websites) with corpus consultation outside of the classroom (Chen & J. Flowerdew, 2018).

And why should they? Today's learners exist within a digital world, and while the concept of 'digital natives' (Prensky, 2009) has been criticised for its simplistic assumptions that modern learners come equipped with the requisite digital literacy and management skills required to navigate this world (Gatto, 2019), what we do know is that accessing digital language content is something that virtually all learners are now used to, with language input in digital forms from a wide range of sources including online games, social media, chat software and multimodal subtitles. Contrast this with the early days of DDL when many students had never even used a keyboard (e.g. Gan et al., 1996). One can only speculate as to how DDL might have developed had the internet been massively available in the 1980s and 1990s, and whether 'corpora' (and all that implies) would have been the most relevant data source. At that time, barring study trips or stays abroad, access to language was mainly in the classroom via teacher and course book, possibly in a resource centre or through occasional newspapers, films and other informal contact. DDL was intended to "cut out the middleman" (i.e. the teacher) between the learner and the language (Johns, 1990, p. 18), but in many classroom contexts this *has already happened*, sparking a change from learners being taught to having the tools at their disposal to teach themselves, with the teacher guiding this process. But they are only likely to transfer these practices if DDL tools evolve to accommodate the ways that *they* are used to sourcing information, rather than DDL practitioners trying to force learners to adopt KWIC concordancing with limited left/right context provided and/or with an overwhelming range of query options to choose from. In essence, we feel that DDL approaches where concordances are the start and finish of DDL pedagogy are unnecessarily limited

[8] https://www.english-corpora.org
[9] https://youglish.com
[10] https://www.linguee.com

in value, and should be supplemented, augmented – or even replaced entirely – with complementary methods of linguistic analysis if DDL is ever going to gain a foothold in mainstream education practice.

Useful studies expanding the boundaries of DDL in this way include Meunier (2019), who discusses the use of a website with film and TV clips aligned between subtitles and video (*PlayPhrase.me*[11]) and another with song lyrics (*LyricsTraining*[12]) for DDL. In both cases, the textual data available from learners' queries are supported by multimedia clips which are available in a wide range of languages, allowing DDL-like learning to occur via tools that were not designed with language learning in mind. On the other hand, dedicated online language learning applications such as *Linggle*[13] (Lai & Chang, 2020) and *FLAX*[14] (S. Wu et al., 2021) are changing the nature of corpus queries and query output to make them easier, more intuitive, and more informational. Corpus applications embedded into word processors such as *Collocaid*[15] (Frankenberg-Garcia et al., 2019) are making promising in-roads into classroom practice in a number of contexts, in this case for help with collocations and phrasing in (academic) writing. Corpus applications are also being combined with natural language processing technology, such as the (soon to be) OER software *G-Rubric*[16] which can provide automated written corrective feedback for revision of learner texts, extracting specialised lexis that the student may or may not have used through comparison with corpus-based model answers (Lancho et al., 2018). This has recently been used with students of English for specific professional purposes, incorporating a gamification element with *G-Rubric* with promising results (Díez-Arcón et al., 2021).

As important as the tools, the nature of the corpus data itself should also be subject to change in response to the needs of its learners. One way is to make corpora more accessible through the use of graded readers or simplified texts, a possibility explored by Hadley and Charles (2017) to encourage extensive reading at lower levels of proficiency. The results were not overwhelming in comparison with the control group, but the groundwork has been laid for future studies to pay greater attention to student needs, preferences, and openness to DDL. Another data type that relates directly to the learners' experience is language produced by them or their peers, again underexplored as a potential source of input for DDL. Promising results were obtained by Moon and Oh (2018), who had learners comparing their own writing against a reference corpus, showing that negative evidence can have its uses and receive positive reactions too. Technology can also play a part in this, through proprietary web-crawling software such as *BootCat*[17] (Baroni & Bernardini, 2004). Such software generates corpora for analysis on the basis of several user-defined keywords, retrieving them from open source data from the web to Wikipedia or even user-defined websites, which have since been used in studies such as Smith (2020). Moreover, certain popular corpora used in DDL studies may often contain pedagogically inappropriate material or offensive/inappropriate content which can impact the uptake of DDL with pre-tertiary learners.

[11] www.playphrase.me
[12] https://fr.lyricstraining.com
[13] https://linggle.com/
[14] http://flax.nzdl.org/greenstone3/flax
[15] https://www.collocaid.uk/
[16] https://www.grubric.com/
[17] https://bootcat.dipintra.it/

Addressing this issue, a recent innovation is crowdsourced corpora, where participants can flag text within a corpus as problematic so that it can be filtered if necessary (Zingano Kuhl et al., 2021). Additionally, pedagogic corpora such as *SACODEYL*[18] and *BACKBONE*[19] are specifically designed with particular learners and learning goals in mind, in these cases for the vastly underappreciated applications to spoken language. One problem here though is that funded projects need continued support and maintenance to keep them up to date. While few and far between (Timmis, 2015), such corpora can help DDL practitioners engage with new, previously unavailable/inaccessible or even formerly unwilling audiences.

Conclusion

The title of this chapter is deliberately provocative: DDL is certainly not 'dead', as witnessed by the increasing numbers of research publications over time and the positive results of various syntheses and meta-analyses in recent years. Nonetheless, there is a feeling that we could be doing more to bring it into the mainstream – *where that is warranted* - but that clearly depends on local conditions and immediate and long-term goals for teaching and individuals, and nobody should expect DDL to be a panacea across the board. The present (and admittedly highly personal) overview argues that DDL research may be treading water, but that there is plenty of scope for innovation if we only take a step back and look at new possibilities without the blinkers of dogmatic reflexes such as "that isn't DDL!"

To begin with, since many learners are already using various tools in various ways that are not entirely dissimilar to DDL, we could encourage that and help them to do it better (or at least, since they're doing it anyway, to avoid the obvious pitfalls and problems). These general-purpose tools (*Google*, *Linguee*, *PlayPhrase*, etc.) have been designed for ease of use and may represent a way into more sophisticated DDL for some learners, but could also be an end in themselves for even larger numbers. One might object that 'these aren't corpora so it's not DDL'; then again, one might wonder if we do actually need corpora for DDL. Any electronic document can be searched rapidly to identify and highlight repeated occurrences of words and phrases; e.g. the approach could be extended bilingually by comparing *Wikipedia* pages in the mother tongue and target language. In other words, in the initial stages at least, we could be thinking about bringing a DDL *mindset* to our learners, rather than expecting them to come to corpus linguistics; a similar point could be made about the relevance of the corpora for the learners themselves, from topics covered to graded readers and learner corpora. The tiny numbers of papers on such topics suggest a wide open space for future research.

Secondly, many empirical studies in applied linguistics tend to focus on the target as somehow 'apart' from other teaching; this makes it easier to research, but it retains a level of artificiality. In the case of DDL, we need to see how it can be integrated more fluidly with other activities and resources in a regular teaching programme. How can links be made between DDL and other materials (textbooks, authentic documents, the internet), skills (especially speaking, but also receptive skills), and of course translation via parallel or comparable corpora? Additionally, how

[18] https://www.um.es/sacodeyl
[19] http://webapps.ael.uni-tuebingen.de/backbone-search/faces/initialize.jsp;jsessionid=6ED116DFEDD14067FED5A425B1FC5533

can we keep DDL 'fun' and avoid its stigmatisation as dry and mechanical? Many corpus software analysis tools have a wealth of other ways to sort and visualise data, but these rarely feature overtly in research. For the language itself, there is more than just vocabulary and lexicogrammar to a world of discourse, switching between corpus and text, and even for pronunciation, intercultural awareness, and so on. We could also look beyond the usual populations, which are overwhelmingly university students needing English (because that's where publishing researchers have the most obvious pool of participants); what about younger learners, or DDL for professional needs in the workplace, and DDL for languages other than English? There are some excellent and innovative studies out there, some of which are reported in this chapter, with others including Leray and Tyne (2016) whose study deals with very young learners in their mother tongue (French) in collaboration with their regular teachers, but these are largely outnumbered by more pedestrian work covering the same ground. Formal replication studies are always useful, but very rare; rather, the repetition seems to stem from an ignorance of research conducted to date.

Thirdly, much DDL research has been reductionist, looking for increased scores in vocabulary learning or lexicogrammar use. Again, this may make for a more satisfying set of findings, but it ignores the real strengths of the approach – at least, such as they are alleged. We know that DDL is time-consuming, especially in early stages as learners get to grips with the new tools, techniques, language types, etc.; this needs to be off-set against longer-term benefits. It is difficult of course to know if an item learned today from any teaching approach, method or activity will still be remembered in a month, a year or a decade from now, but that's missing the point. DDL has often been argued to lead to deeper cognitive processing, greater language awareness, more sophisticated linguistic reasoning, greater autonomy in dealing with language and continued progress – in sum, becoming 'better learners' in the long run. Intuitively, one feels these things should be the case; but there comes a time when this needs putting to the test directly rather than being inferred from successful studies that assume this from the start. These knock-on effects – i.e. that DDL for language points A, B and C may also improve the ability to deal with X, Y and Z at some future point – desperately need direct exploration. This might partly come through a more solid theoretical foundation, drawing on a wide range of theories: noticing in SLA, constructivism in education, intercultural communication in sociology, pattern-detection in DST, etc. Psychology is an especially promising source here, from general processes such as cognitive depth and memory, to chunking and usage-based models of exemplar-driven learning. Admittedly, such things will not be easy to test (which is presumably one reason why they have been little examined to date), but that does not mean the task is impossible.

Similarly, a reductionist methodology limits the ecology of the research. There is a need for more long-term designs, or at least delayed post-test data, as well as research into what learners do with corpora outside class and after the course has finished. Besides seeing how DDL works with 'regular' teachers, we also need larger samples from more groups, and greater diversity of profiles in terms of proficiency, age, profile, needs and motivations, among other things. This would also open up new avenues for comparing two or more experimental conditions, thus avoiding the 'DDL vs control' mindset which, as argued above, has largely been done to death. These could include different languages or years or disciplines, within an institution or between institutions, and even between countries. Finally, we do need improved methodology and reporting. While this criticism is not exclusive to DDL, it's little consolation to think that we are no worse than anyone else – we

should be better! In addition to the fairly mundane recommendations above and in many survey papers in applied linguistics, there is a need for more original designs that get away from the usual: (a) pre/post-test or control/experimental group comparisons, especially relying exclusively on statistical significance; (b) questionnaires and interviews, especially where extracts are used to 'prove' a point with no indication of their representativity or generalisability – they're just interesting factoids. Both have their uses, and can be combined in a truly mixed methods approach rather than simply juxtaposed, but the former (e.g. gap-fills and multiple-choice questions) tend to be fairly limited in terms of ecological validity, while the latter are often subjective and unreliable. We could also be looking at other types of output from learners, including extended writing, speaking and translation, as well as global comprehension; and different ways to access their actual behaviour and attitudes, notably eye-tracking and think-aloud protocols.

To conclude: DDL has a varied and distinguished history over the past 30 years. It's up to us all to make sure that continues.

References

Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies*. Routledge.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004* (n.p.). https://home.sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf

Boulton, A. (2011). Language awareness and medium-term benefits of corpus consultation. In A. Gimeno Sanz (Ed.), *New trends in computer-assisted language learning: Working together* (pp. 39–46). Macmillan ELT.

Boulton, A. (in press). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education*. John Benjamins.

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

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3), XX–XX. https://doi.org/10125/XXXXX

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

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. English for Specific Purposes, 35, 30–40. https://doi.org/10.1016/j.esp.2013.11.004

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. (1999). Applying constructivism: A test for the learner as scientist. *Educational Technology Research & Development, 47*(3), 15–31. https://doi.org/10.1007/BF02299631

Crosthwaite, P. (2020). Taking DDL online: Designing, implementing and evaluating a SPOC on data-driven learning for tertiary L2 writing. *Australian Review of Applied Linguistics, 43*(2), 169–195. https://doi.org/10.1075/aral.00031.cro

Crosthwaite, P., Luciana, & Schweinberger, M. (2021). Voices from the periphery: Perceptions of Indonesian primary vs secondary pre-service teacher trainees about corpora and data-driven learning in the L2 English classroom. *Applied Corpus Linguistics, 1*(1), 100003. https://doi.org/10.1016/j.acorp.2021.100003

Crosthwaite, P., Wong, L. L., & Cheung, J. (2019). Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning. *ReCALL, 31*(3), 255–275. https://doi.org/10.1017/S0958344019000077

Crosthwaite, P. et al. (forthcoming). ...

Díez-Arcón, P., & Martín-Monje, E. (2021). G-Rubric: The use of open technologies to provide personalised feedback in languages for specific purposes. *EDULEARN21 Proceedings* (pp. 2635–2643). https://doi.org/10.21125/edulearn.2021.0574

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Wiley. https://doi.org/10.1002/9781118346136.ch7

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: a critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Flowerdew, L. (2015). Data-driven learning and language learning theories: Wither 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

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23–39. https://doi.org/10.1017/S0958344018000150

Gan, S-L., Low, F., & Yaakub, N. (1996). Modeling teaching with a computer-based concordancer in a TESL preservice teacher education program. *Journal of Computing in Teacher Education, 12*(4), 28–32. https://doi.org/10.1080/10402454.1996.10784301

Gatto, M. (2019). Query complexity and query refinement: Using web search from a corpus perspective with digital natives. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for younger learners* (pp. 106–129). Routledge. https://doi.org/10.4324/9780429425899-7

Gillespie, J. (2020). CALL research: Where are we now? *ReCALL, 32*(2), 127–144. https://doi.org/10.1017/S0958344020000051

Hadley, G., & Charles, M. (2017). Enhancing extensive reading with data-driven learning. *Language Learning & Technology, 21*(3), 131-152. https://doi.org/10125/44624

Han, S., & Shin, J.-A. (2021). Teaching Google search techniques in an L2 academic writing context. *Language Learning & Technology, 21*(3), 172–194. https://doi.org/10125/44626

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. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). Addison Wesley Longman.

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

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology, 5*(3), 77–90. https://doi.org/10125/44567

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology, 14*(1), 28–44. https://doi.org/10125/44201

Kennedy, C., & Miceli, T. (2017). Cultivating effective corpus use by language learners. *Computer Assisted Language Learning, 30*(1-2), 91–114. https://doi.org/10.1080/09588221.2016.1264427

Koehler, M., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*(1), 60–70.

Lai, S.-L., & Chang, J. S. (2020). Toward a pattern-based referencing tool: Learner interactions and perceptions. *ReCALL, 32*(3), 272–290. https://doi.org/10.1017/S0958344020000105

Lancho, M. S., Hernández, M., Paniagua, Á. S. E., Encabo, J. M. L., & De Jorge-Botana, G. (2018). Using semantic technologies for formative assessment and scoring in large courses and MOOCs. *Journal of Interactive Media in Education, 2018*(1), 1–10. https://doi.org/10.5334/jime.468

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

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

Lee, P., & Lin, H. (2019). The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention. *System, 81*, 14–25. https://doi.org/10.1016/j.system.2018.12.011

Leńko-Szymańska, A. (2017). Training teachers in data driven learning: Tackling the challenge. Language Learning & Technology, 21(3), 217-241.

Leray, M., & Tyne, H. (2016). Homophonie et maîtrise du français écrit: apport de l'apprentissage sur corpus. *Linguistik online*, *78*(4).

Le Foll, E. (Ed.) (2021). *Creating corpus-informed materials for the English as a foreign language classroom: A step-by-step guide for (trainee) teachers using online resources* (3rd edn). Pressbooks. http://dx.doi.org/10.5281/zenodo.4992504

Ma, Q., Tang, J., & Lin, S. (2021). The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning*, advance access. https://doi.org/10.1080/09588221.2021.1895225

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meunier, F. (2019). A case for constructive alignment in DDL: Rethinking outcomes, practices and assessment in (data-driven) language learning. In P. Crosthwaite (ed.) *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 13-31). Routledge / Taylor & Francis. https://doi.org/10.4324/9780429425899-2

Moon, S., & Oh, S.-Y. (2018). Unlearning overgenerated 'be' through data-driven learning in the secondary EFL classroom. *ReCALL, 30*(1), 48–67. https://doi.org/10.1017/S0958344017000246

O'Keeffe, A. (2020). Data-driven learning: A call for a broader research gaze. *Language Teaching, 54*(2), 259–272. https://doi.org/10.1017/S0261444820000245

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*, advance access. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., Mark, G., & O´Keeffe, A. (2020). *The impact of usage-based approaches on second language learning and teaching*. Cambridge University Press.

Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J. M., & Jiménez, P. A. (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, *24*(3), 233–253. https://doi.org/10.1080/09588221.2010.539978

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*(1), 440–460. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: Journal of Online Education, 5*(3), n.p. https://www.learntechlib.org/p/104264/

Quinn, C. (2015). Training L2 writers to reference corpora as a self-correction tool. *ELT Journal*, *69*(2), 165–177. https://doi.org/10.1093/elt/ccu062

Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb-preposition collocations. *System*, *91*, 1–14. https://doi.org/10.1016/j.system.2020.102268

Schaeffer-Lacroix, E. (2019). Barriers to trainee teachers' corpus use. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for younger learners* (pp. 47-64). Routledge. https://doi.org/10.4324/9780429425899-4

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Sinclair, S. & Rockwell, G. (2016). *Voyant Tools*. http://voyant-tools.org/

Smith, S. (2019). DIY corpora for accounting & finance vocabulary learning. *English for Specific Purposes, 57*, 1–12. https://doi.org/10.1016/j.esp.2019.08.002

Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice*. Routledge. https://doi.org/10.4324/9781315715537

Tsai, K. J. (2019). Corpora and dictionaries as learning aids: inductive versus deductive approaches to constructing vocabulary knowledge. *Computer Assisted Language Learning*, *32*(8), 805–826. https://doi.org/10.1080/09588221.2018.1527366

VanPatten, B., & Benati, A. (2010). *Key terms in second language acquisition.* Bloomsbury.

Viana, V., & Lu, L. (2021). Corpus linguistics and continuous professional development: Participants' prior knowledge, motivations and appraisals. *Revista de Estudos da Linguagem*, *29*(2), 1485–1527. https://doi.org/10.17851/2237-2083.29.2.1485-1527

Vincent, B., & Nesi, H. (2018). The BAWE quicklinks project: A new DDL resource for university students. *Lidil. Revue de Linguistique et de Didactique des Langues, 58,* 1–17. https://doi.org/10.4000/lidil.5306

Vyatkina, N. (2020). Corpora as open educational resources for language teaching. *Foreign Language Annals*, *53*(2), 359–370. https://doi.org/10.1111/flan.12464

Wu, S., Fitzgerald, A., Yu, A., & Chen, Z. (2021). What are language learners looking for in a collocation consultation system? Identifying collocation look-up patterns with user query data. *ReCALL, 33*(3), 229–247. https://doi.org/10.1017/S0958344021000057

Wu, Y.-J. A. (2021). Discovering collocations via data-driven learning in L2 writing. *Language Learning & Technology, 25*(2), 192–214. https://doi.org/10125/73440

Zingano Kuhn, T., Šandrih Todorović, B., Arhar Holdt, Š., Zviel-Girshin, R., Koppel, K., Luís, A. R., & Kosem, I. (2021). Crowdsourcing pedagogical corpora for lexicographical purposes. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion* (Vol. 2, pp. 771–779). EURALEX Proceedings.