Chapter 26: Classroom applications of corpus analysis

Alex Boulton, Université de Lorraine, France
Nina Vyatkina, University of Kansas, USA
Thomas Cobb, Université du Québec à Montréal, Canada

1. Introduction

Corpus linguistics is almost by definition *applied* linguistics, as was tacitly acknowledged when the American Association of Applied Corpus Linguistics (AAACL) dropped its third A in 2008. Its methodologies can be applied far beyond the discipline itself, not least in language teaching and learning, where its influence has been of three main types. The first lies in improved descriptions of language varieties and features which can inform aspects of the language to be taught (section 2.1 Upstream use); the second makes corpora and tools for analyzing them available to the teacher (section 1.2 Teacher use); the third puts them directly into the learner's hands (section 1.3 Learner use). We begin this chapter with an overview of all three types before concentrating mainly on the third type in the final sections, since other chapters in this volume deal in more detail with corpora and vocabulary (Chapter 26), lexicography and phraseology (Chapter 27), pedagogical materials (Chapter 29), and translation (Chapter 30).

2. Corpus use by materials designers, teachers, and students

2.1 Upstream use

Early instantiations of the first approach predate modern electronic corpora, with famous examples including Thorndike and Lorge's *Teacher's wordbook of 30,000 words* (1944) or West's *General service list* (1953) for English, and Gougenheim's *Dictionnaire fondamental de la langue française* (1958) for French. Work on frequency lists continues to this day derived from ever larger, electronic corpora, such as the British National Corpus (BNC, 100m words) and the Corpus of Contemporary American English (COCA, 1b words), and has spread beyond English, as seen in the *Routledge frequency dictionaries* series based on corpora of 13 other languages. Of course, frequency applies not only to words, but also to larger units like phrases and chunks, as in Martinez and Schmitt's (2012) BNC-based phrasal expressions list. While by no means the only criterion, frequency of form and meaning is a useful predictor of what can most usefully be taught at different points in the learning process, as argued by Cobb (2007) for the early stages, or Schmitt and Schmitt (2014) for later stages. This type of work can thus inform syllabus design and testing, as the choice and sequence of forms and meanings to teach and test becomes more empirically based, for example in the design of TOEFL tests (Biber *et al*., 2004) and frequency-based vocabulary tests (Nation & Beglar, 2007). Frequency analysis of learner corpora can also help to determine what learners of different backgrounds typically can and cannot do at different levels, again feeding into syllabus design more effectively than previous attempts at contrastive analysis based on qualitative structural differences (Granger, 2009). The English Profile project from Cambridge University is a major example of this type of work informed by both native-speaker and learner corpora.

One of the most influential indirect pedagogical applications of corpus research has been the Cobuild project at Birmingham University (see Sinclair, 1987) with the design of a large evolving corpus of English, including a radically new type of dictionary with the entries chosen and organized according to frequency and accompanied by uncompromisingly authentic examples taken from the corpus. All the large publishing houses have followed this lead, and today it would be inconceivable to produce a dictionary in a major language without substantial corpus input. The influence does not stop at lexis but can also be exploited in the production of usage manuals and grammar books, such as the *Grammar of spoken and written English* (GSWE: Biber et al., 1999/2021). Corpora have also been used in the construction of teaching materials, though in many cases (e.g., *Touchstone*; McCarthy, McCarten & Sandiford, 2006) the activities are indistinguishable from those in traditional books; the innovation is that the language taught is based on "real" usage and frequency data rather than depending on an author's (often fallible) intuitions or fortuitous occurrences in the language inputs selected for learners' attention.

But it is possible to go further still and make direct use of corpus material with learners. Reppen (2010: Ch. 2) and Bennett (2010: Ch. 3) discuss activities that make explicit use of corpus information in grammar books such as the LGSWE, sensitizing learners to issues of frequency, morphology, chunking, collocations, register, and so on. A small but growing quantity of recent published materials include corpus data, from grammar books (e.g., Reppen, Blass, Iannuzzi, Bunting & Diniz, 2020; Thornbury, 2004) to supplementary materials (e.g., Thurstun & Candlin, 1997) and even full courses (e.g., Karpenko-Seccombe, 2020; Mohamed & Acklam, 1995). In books like these, concordance lines and other corpus data are turned into activities that students can use to explore the language, either deductively (e.g., to test rules or categorize different uses), or inductively (i.e., to formulate their own hypotheses about usage and then test these against a corpus).

2.2 Teacher use

This brings us to the second major use of corpora in the language classroom, when teachers consult corpus data directly rather than relying on decision-makers upstream. First, corpus tools can be applied to individual texts, in helping decide whether a text is appropriate and what elements to focus on. Publicly available and free text profiling software such as VocabProfile online (www.lexutor.ca/vp) or AntWordProfiler offline (www.antlab.sci.waseda.ac.jp/) allows a teacher to input a text which is then returned with the lexis color-coded according to the frequency of each word in the BNC or COCA. Such information can help with decisions about which items to teach in a given text, for example ignoring or glossing less frequent items while using the highly visible multiple occurrences of others as an aid to teaching in context (Cobb, 2007).

From the teacher's perspective, corpora can help in deciding what to teach. Often the corpora used for this purpose are not large modern corpora like the BNC or COCA but rather smallish corpora like the Brown (1m words), or else purpose-built and sometimes level-appropriate text collections not necessarily meant to be representative of a language in its entirety. These corpora can be particularly useful in teaching languages for specific purposes where published materials are difficult to come by. Frequency of occurrence and typical usage can be a useful guide, though of course these need to be tempered by pedagogical considerations. Corpora can also provide a useful source of authentic language, as the teacher can select typical language *samples* to

complement or replace the invented language *examples* often found in teaching materials (Gavioli, 2005: 7). This applies not just to teaching, but also to testing: Stevens (1991) found the use of multiple authentic concordance lines especially beneficial in gap-fill tests, effectively allowing English for specific purposes (ESP) tests to be constructed from authentic rather than made-up language. In error correction, rather than providing "the answer", a teacher can conduct a search in corpora such as COCA or BAWE, where each query provides a unique URL, and copy that to the student's paper: clicking on the link conducts the exact same search and the learner has only to interpret the data (Gaskell & Cobb, 2004).

L1 and L2 native teachers can also turn to corpora when they have a language question, as intuition is notoriously unreliable in many cases (even textbook rules are at times quite inadequate descriptions of actual language use; e.g., Carter, Hughes & McCarthy, 1998). This can be helpful in correcting work outside the class but can also serve as an in-class "informant" when responding to unforeseen learner questions. Where no explanation comes readily to mind, it gives the teacher a way to test intuitions, and an alternative to inventing a spurious rule or simply replying "because" (cf. Johns, 1990). Finally, teachers can use corpora in similar ways to the manuals outlined above, selecting corpus data (concordance lines, distribution tables, collocation lists, cluster diagrams, and so on) to create focused activities. Several recently published open educational resources provide teachers with scaffolding as guides to utilizing corpora and other DDL tools in teaching and as repositories of ready-made activities (e.g., Cobb, 2021; Goulart & Veloso, 2023; Le Foll, 2021; Pinto *et al*., 2023; Vyatkina, 2020b).

2.3 Learner use

Here we come to the third and final major use of corpora by language learners themselves. Corpus-based learning tasks and activities can be designed along a wide spectrum from "hard" to "soft" (cf. Gabrielatos, 2005), beginning with totally controlled exercises and ending with genuine questions of interpretation. The teacher can decide on the question, query a relevant corpus and choose the appropriate information, which is then modelled into an activity with focused instructions and closed answers leading to predetermined outcomes. With time, any or all of these decisions and stages can however be taken over by learners themselves. The learner querying of corpora involves techniques that are essentially akin to the activities of corpus linguists: "Like a researcher, the learner has to form preliminary hypotheses on the basis of intuition or scanty evidence; those hypotheses then have to be tested and rejected or refined against further evidence, and finally integrated within an overall model" (Johns, 1988: 14). Corpus consultation in this manner may focus on learning *per se*, or it may involve using a corpus as a reference tool alongside dictionaries, GenAI and other resources in both comprehension and production, especially of written language. In reading, learners can quickly check specific patterns that may not be frequent enough to warrant a mention in dictionaries, or they can access all the occurrences of unknown words or uses in a given text, thus accessing more relevant and focused contexts than may be found in a dictionary (Cobb, Greaves & Horst, 2001). In drafting or revising texts or translations, learners can also check their tentative work against "normal" use in large or specialized corpora (O'Sullivan & Chambers, 2006) or use tools such as ColloCaid (https://collocaid.uk): begin typing a text and the tool suggests collocates and clusters from the corpus of academic writing. While some corpora include transcripts, a small number allow the user to query against videos. Among the best-known, as some queries can be conducted free, are www.PlayPhrase.me and www.YouGlish.com, although neither are really

"corpora". The first features films and TV series, the second videos from YouTube, each with a certain number of limitations, but the learner can type a query and, rather than seeing a written concordance to scan, go to several videos one after the other featuring the word and phrase, allowing work on intonation and pronunciation as well as lexicogrammar. Despite such innovations, though, it is perhaps surprising how similar DDL "looks" today and 10 or 15 years ago: there are more and larger corpora, faster and more efficient tools, and more esthetic interfaces, though most underlying functions have changed little.

Clearly in its most open-ended form, such activity can be quite demanding on the learner, who is likely to need intensive training or, perhaps preferably, scaffolding during extensive practice to reap the full benefits of corpus consultation. We therefore need sound theoretical reasons to introduce learners to work of this type (Chambers, Conacher & Littlemore, 2004). The basic idea is that massive but controlled exposure to authentic input is of major importance (input flood), as learners gradually respond to and reproduce the underlying lexical, grammatical, pragmatic and other patterns implicit in the languages they encounter. This can be through unconscious habit-formation from a behaviorist/emergentist perspective – see Hoey's (2005) theory of priming or Taylor's (2012) account of implicit accumulated memories in mental corpus theory – or through some element of conscious noticing from a language awareness perspective (Schmidt, 1990). In other words, rather than having a tool do all the work (as may be a risk with GenAI tools), the learner still needs to be actively involved in interpreting the data. Other proposed benefits include the motivation inherent in use of ICT for individualized purposes where the learners build their knowledge based on their own needs and interests; learner corpus work is thus a generally constructivist and inductive approach to language learning, the discovery and problem-solving procedures favoring cognitive and metacognitive development, critical thinking and noticing skills, language awareness and sensitivity in dealing with authentic text, as well as autonomy and life-long learning (see e.g., O'Sullivan, 2007: 277–278; Römer, 2006: 26).

All of these would appear to be desirable elements in current applied linguistics thinking. The question of course is whether corpus work really lives up to expectations, with benefits sufficient to justify the investment. For this, we need to look at research to date, which is the purpose of the rest of this chapter. The following section takes an overview of the research field as a whole, then focuses in on the findings from research syntheses and meta-analyses published to date in order to assess more broadly the benefits derived (or costs incurred) from the direct use of corpora by learners.

3. Empirical research in L2 corpus use

3.1 Chronology of DDL research

Getting learners to explore language is nothing new: they are frequently asked to compare example sentences on the blackboard or identify features of written or spoken texts. The use of corpora merely moves it up a level, increasing the quantity of authentic data available for examination, systematizing the querying procedures and output language, and potentially allowing learners a greater role in the process. According to McEnery and Wilson (1997: 12), the first such uses of corpora go back to the late 1960s at Aston University in Birmingham; other beginnings can be found in ESP courses at the University of Nottingham in the early 1970s (Butler, 1974). The first published paper to our knowledge is by McKay (1980), describing

learner use of printed corpus-based materials at San Francisco State University; the first description of hands-on concordancing at the University of Surrey can be found in Ahmad, Corbett and Rogers (1985). However, the approach, as well as the term "Data-Driven Learning" (DDL), is largely associated with Tim Johns at the University of Birmingham, where he and other colleagues gave their students access to Cobuild and other corpora and software in the 1980s for pedagogical purposes (see Johns & King, 1991).

Most of the early academic publications emanating from all this activity were descriptive and argumentative, following the typical evolutionary curve in applied linguistics research (Boulton & Pérez-Paredes, 2024). The first empirical evaluation was published by Baten, Cornu and Engels (1989), "followed by increasing numbers of ever more focused and sophisticated empirical evaluations (…) until such time as enough are available for various types of research syntheses, with theoretical underpinnings appearing at various points along the way" (Boulton, 2017: 485). This progression from descriptive to empirical to synthetic research roughly corresponds to three evolutionary stages of DDL research identified by Dong, Zhao, and Buckingham (2023: 339), namely "the conceptualizing stage (1980s–1998), the maturing stage (1998–2011), and the expansion stage (2011–now)". Although we would argue that the borders between stages are rather fuzzy, the fact remains that by now, the DDL field has grown into an established research discipline. The authors of this chapter have been conducting systematic searches for DDL publications over the last 15 years, trawling multiple research databases (LLBA, MLA, ERIC, Web of Science, etc.) and subsequently reference lists of all publications found therein (see Boulton and Vyatkina [2021: 69] for search procedures). By the time of writing, we have identified 800+ empirical studies and 20+ secondary research studies that synthesize empirical DDL research in various ways. Additionally, several recent syntheses of broader fields such as corpus linguistics and computer-assisted language learning (CALL) have included data on DDL research. We will draw upon all these synthetic studies in our overview of empirical DDL research in the following sections, which thus present a type of "meta-synthesis" (Plonsky, 2023: 1).

In their recent taxonomy, Chong and Plonsky (2024) identified 13 different types of secondary research studies in applied linguistics, and it is yet another sign of maturity of the DDL field that we were able to find characteristics of practically all of these in DDL syntheses published to date. Chong and Plonsky use four dimensions in their classification: purpose (research-focused or practice-focused), review process (how systematic), structure (how standardized), and text (monomodal or multimodal). Each dimension is considered a continuum, and each study may combine characteristics of different types. One important distinction is between meta-analyses that employ statistical methods to compare effect sizes across empirical studies vs. other types of syntheses that are purely qualitative or employ simple counting and descriptive, but not inferential, statistics. We review these two types of studies and their findings separately in sections 2.2 and 2.3. As many of these syntheses include some methodological review, we also summarize identified methodological issues in section 3.

3.2 Syntheses in DDL

3.2.1. DDL in corpus linguistics and CALL research syntheses

DDL's emergence as a separate research domain is confirmed in three recent bibliometric reviews of broader technology-related fields. Park and Nam (2017), to our knowledge, is the first such study that found a "data-driven learning" cluster (containing 29 articles) in their co-citation

analysis of corpus linguistics research published 1997–2016. Choubsaz, Jalilifar and Boulton (2024: 104) found that 18 DDL studies "have contributed to theorizing CALL papers in the last two decades", which constituted 2.8% of their target collection of 426 high-impact articles published in major CALL journals in 1983–2019. Crosthwaite, Ningrum and Schweinberger (2023) presented the most compelling evidence of the growing impact of DDL: they demonstrated that DDL researchers, along with those exploring academic discourse, were the most cited in Scopus-indexed corpus linguistics articles published in the last five years of the 2001–2020 time span. Other bibliometric studies of corpus-related research (e.g., Abduh *et al.*, 2023; He & Wei, 2019) also reference DDL studies while not singling them out as a cluster due to their focus on other topics such as language skills.

3.2.2. Types, scope, and chronology of DDL syntheses

Our search yielded 16 research syntheses devoted solely to DDL, not counting five meta-analyses which will be discussed in the next section. Their scope ranged from general (the whole DDL field) to specific (e.g., EAP writing). Regarding Chong and Plonsky's (2024) dimensions of secondary research, all these syntheses were research-focused (with an additional focus on practice in some). Specifically, they examined empirical studies which "subject some aspect of DDL to observation or experimentation with some kind of externally validated evaluation other than the researchers' own intuition" (Boulton, 2010: 130). The syntheses exhibited various degrees of three other dimensions: systematicity, standardized structure, and multimodality. It is noteworthy that all these syntheses also included elements of a "research agenda," usually in their concluding sections. Table 1 identifies the scope and additional focus of each synthetic study using the information from the studies themselves as well as Chong and Plonsky's (2024: 12) descriptors.

Table 1. Syntheses in DDL

	Study	**N prim.**	**Scope**	**Review Type**
Period A	Chambers (2007)	12	consultation of NS corpora	narrative
	Boulton (2007)	39	general	narrative
	Boulton (2010)	27	learning outcomes	narrative
	Yoon (2011)	12	concordancing for writing	narrative
	Boulton (2012)	20	ESP, consultation	narrative
Period B	Boulton & Tyne (2013)	116	general	critical
	Luo & Zhou (2017)	18	L2 English writing	narrative
	Boulton (2017)	46	general	historical
	Chen & Flowerdew (2018)	37	EAP, writing	critical
Period C	Boulton (2021)	351	general	systematic; scoping
	Boulton & Vyatkina (2021)	489	general	systematic; scoping
	Pérez-Paredes (2022)	32	normalization, 5 years	systematic; practice-focused
	Dong *et al*. (2023)	126	general	systematic; bibliometric
	Lusta *et al*. (2023)	81	general	systematic; practice-focused

Study	N prim.	Scope	Review Type
Sun & Park (2023)	32	English collocations	systematic
Boulton & Vyatkina (2024)	148	English, impact factor	systematic; methodological

Our synthesis collection can tentatively be divided into three periods (Table 1), the first one starting at the end of Dong *et al*.'s (2023) "maturing stage" of DDL (section 2.1). The five earliest syntheses (2007–2012) were written as basic narrative reviews of a nascent research field. The second period (2013–2018) was characterized by the appearance of two new review types (critical and historical review), and the third (2021–2024) was marked by both quantitative and qualitative changes in DDL syntheses in line with the recent "synthetic-mindedness" movement in applied linguistics research (Plonsky, 2023: 9). First, the pool of primary studies increased considerably. Second, state-of-the-art quantitative and qualitative review methods were introduced. Third, systematicity in reporting the methodology and procedures improved. Fourth, reviews became multimodal by heavily utilizing tables, charts, and figures to enhance clarity and impact. These studies also included additional foci, thus adding to the range of synthesis types in DDL research: scoping (i.e., providing an overview of the DDL field), bibliometric, practice-oriented, and methodological. Interestingly, Dong *et al*. (2023) confirmed the growing impact of synthetic studies by finding Boulton and Vyatkina's (2021) scoping review to have a "transformative" potential on DDL research.

3.2.3. Findings of DDL syntheses[1]

The narrative, critical, and historical reviews published in the first two periods (2007–2018) asserted the DDL research field, with several positive findings. First, corpora were shown to serve as an effective tool for reference and learning, enhancing linguistic accuracy, productive language use, and language awareness, especially in L2 writing. Second, learners generally showed positive attitudes towards DDL and a remarkable capacity for applying corpus techniques. Syntheses also pinpointed DDL limitations: that not all corpora are equally appropriate for all learners across all levels and language points, as well as the time-consuming nature of searching and interpreting concordance lines, which sometimes caused learner frustration and boredom if it became too mechanical. In their research agendas, the authors of synthetic studies called for increasing sample sizes, supporting learners in conducting independent searches and learning outside the classroom, and addressing the scarcity of appropriate DDL resources by developing and disseminating ready-made materials as well as simple, free, stable, and accessible tools for learners of various levels. Finally, they called for improved communication between researchers and teachers to ensure the development and adoption of DDL resources that were aligned with classroom realities.

Systematic DDL research syntheses published in the third period (2021–2024) have shown that the quantity of empirical studies has increased dramatically, as has the variety of corpora and related tools used for DDL. Nevertheless, some limitations of earlier research effectiveness persist, and new research gaps and desiderata have been put forward. Two syntheses with a

[1] The GPT-3.5 chatbot (https://chat.openai.com) was used to generate an initial list of the common themes discussed in section 2.2.3 based on discussion sections of the reviewed research syntheses. One of the authors then reviewed and revised this list.

practice-focused component (Lusta, Demirel & Mohammadzadeh, 2023; Pérez-Paredes, 2022) and Dong *et al*.'s (2023) bibliometric synthesis suggest specific research topics that could pave the way to the "normalization" of DDL:

a) syllabus integration
b) language teacher training
c) classroom implementation research
d) pedagogical approaches and learner differences
e) accessibility and usability of DDL
f) imitation vs. creativity in DDL
g) scaffolding techniques evaluation

Finally, all recent syntheses contain observations regarding the developments and persistent gaps related to research methodology in DDL studies. These issues, along with those gleaned from DDL meta-analyses (see 2.3), will be summarized in section 3.

3.3 Meta-analyses in DDL

3.3.1 Meta-analytic methodology

A meta-analysis combines quantitative findings from numerous similar studies into a single figure, based not on statistical significance between outcome differences but on the *size* of the effect of the treatment on a target language feature. Then in a second stage, overall effect size can be parceled out into the contributing effect sizes of specific moderator variables (MVs), e.g., hands-on (computer-based) vs hands-off (paper-based) DDL. Either type of effect size can be interpreted against Plonsky and Oswald's (2014) useful scheme of verbal equivalences (large, medium, or small) as distinguished for pre/post and control/experimental designs and tailored to the field of applied linguistics field (Table 2). These equivalences should not be thought of as dichotomous cut-offs (cf. significant or not significant, at an arbitrary level), but rather as "neighborhoods" (Plonsky & Oswald, 2014: 889). In an update, Plonsky, Hu, Sudina and Oswald (2023: 309) propose the alternative terms "small-ish," "medium-ish" and "large-ish." The differentiation between pre/post and control/experimental reflects the comparative ease of getting a strong result from the same group of participants (test–teach–test) compared to with different groups of participants.

Table 2. Plonsky and Oswald's (2014) field-specific interpretation of effect sizes

	Within groups (pre/post-test)	Between groups (control/experimental)
Large	1.4	0.9
Medium	1.0	0.6
Small	0.6	0.4

3.3.2 Cobb and Boulton (2015)

Cobb and Boulton contributed a chapter to the first edition of this handbook in 2015, in which they presented among other things the first preliminary meta-analysis of DDL. An initial set of 21 studies was investigated, with effect sizes calculated in standard deviation units (SDUs) via Cohen's *d*, the difference between means divided by pooled standard deviation. The differences for both pre/post-test (1.68 SDUs) and control/experimental (1.04) comparisons were higher than

expected: these are “very large” and “large” differences, respectively, by Plonsky and Oswald’s standards. However, as this was only a preliminary study with a small sample, individual MVs were not analyzed but pursued in a follow-up study (section 2.3.4).

The remainder of this section summarizes results from the four meta-analyses of DDL published since 2015. It will look at different approaches, similarities and differences in findings, and any issues associated with doing meta-analysis on this topic. We will start with research questions (RQs) posed by each meta-analysis and review them in chronological order on the assumption that the development of the methodology is cumulative.

3.3.3 Mizumoto and Chujo (2015)

RQ: “How effective, in terms of synthesized effect size, is the DDL approach in the Japanese classroom context?” (p. 3)

Like Cobb and Boulton’s (2015) meta-analysis of 21 primary studies, this study could also be called preliminary. It involved 14 published studies, all of which shared a common set of characteristics: a pre/post-test research design; Chujo as one of the authors; and low-proficiency Japanese English learners as the participants. Mizumoto and Chujo arrived at an overall effect size of 0.97 SDUs, in the region of “medium” for pre/post-test studies. Despite the small numbers and the homogeneity of both design and learners, the researchers proceeded to look at moderator variables, finding that lexicogrammar (vocabulary, collocations, and basic grammar) was the language feature most amenable to learning by DDL. The question here is whether these findings reflect any limitation in what DDL is good for beyond this particular learner population.

One point of interest in Mizumoto and Chujo’s overall finding of 0.97 SDUs is that it is stronger than Spada and Tomita’s (2010) partially comparable finding a few years earlier of 0.63 (“small’) for “simple grammar constructions” taught through explicit (though not corpus-based) instruction. Simple constructions are presumably comparable to “lexicogrammatical features”. However, the comparison is merely suggestive since Spada and Tomita’s analysis included both pre/post-test and control/experimental designs and was not limited to one group of learners.

3.3.4 Boulton and Cobb (2017)

RQ1. “How much DDL research is there?”
RQ2. “How effective is DDL, and how efficient is it?”
RQ3. “How can we best account for any variation observed?” (p. 354)

This study followed up on Cobb and Boulton (2015) as a full meta-analysis of all quantitative DDL studies then known. “Effective” refers to pre/post-test comparisons (i.e., DDL has an effect); “efficient” refers to a control/experimental group comparisons (i.e., whether DDL is more effective than the alternative). The researchers cast their net wider for source studies, capturing any full-text study written up through June 2014, including PhD dissertations and papers in less well-known journals, but not MA theses or conference presentations as these cannot be searched systematically. The initial trawl yielded 205 candidate publications, of which 64 were useable, for a total of 88 unique samples.

The main result was a confirmation of the preliminary study, with remarkably similar overall effect sizes of 1.5 SDUs for pre/post and 0.95 for control/experimental designs (both “large”). From the descriptive sections of the 64 publications, a hierarchy of candidate MVs was derived, some following the usual structure of meta-analyses in education (publication, population, treatment, design) and others adapted for the particularities of DDL (corpus size and type, paper or computer interface, used for reference or learning, etc.) yielding a broad set of 84 MVs in 25

groups. Individual effect sizes were calculated for each MV and, of the 84, about half (40) were found to be both sizeable and consistent enough for analysis. The larger effect sizes, in no particular order, were these: DDL seems more useful in foreign- over second-language environments; via computer interfaces over printed materials; using small local over large standard corpora; for intermediate learners over beginners or advanced learners; for learning vocabulary and collocation/lexicogrammar rather than other features of language; as a learning resource over reference resource; in more recent studies over older ones. The last point seemed an encouraging sign that research practices were improving and a culture of DDL emerging.

The small sample size for some MVs was addressed by separating out "the most robust results (understood as MVs with at least 10 unique samples in both P/P and C/E designs)", finding 70% had large effects, 25% medium, and only 5% small or negligible: "From this we reach the somewhat surprising and possibly encouraging conclusion that DDL works pretty well in almost any context where it has been extensively tried" (p. 386). But there remained a need for investigations of language beyond the word and phrase (like grammar, discourse, culture, and pragmatics) as well as delayed post-testing generally. These were identified as matters to be addressed in future meta-analyses, three of which have since become available.

3.3.5 Lee, Warschauer and Lee (2019)

RQ1. "How effective[2] is corpus use in improving L2 vocabulary learning?"
RQ2. "What are the moderators that influence the magnitude of the effectiveness of corpus use?" (p. 727)

Lee *et al*. focused solely on DDL for vocabulary learning. Rather than expanding the number of studies to include, their strict criteria excluded all studies that did not have a control or comparison group, leaving them with 29 studies and 38 separate effect sizes to work with, many the same that had already featured in Boulton and Cobb (2017). In addition to its focus on vocabulary, the originality of this study lies in its multilevel model of meta-analysis, namely a regression analysis in the calculation of MV effect sizes, which while uncommon is desirable in meta-analytic research (Plonsky *et al*., 2023: 322). The three types of vocabulary knowledge measured were referential/definitional, syntactic/collocational, and productive – a framework abundantly motivated in vocabulary research.

The regression analysis thus began with the overall effect size for vocabulary learning through corpus consultation, 0.74 SDUs; the authors interpreted this as a "medium" effect, though it is more accurately termed "between medium and large" unless we are using these descriptors as cut-offs (cf. Table 2 for "between groups"). The MV analysis produced effect sizes for the individual levels of vocabulary learning: 0.46 SDUs ("small") for referential/definitional knowledge, 0.92 ("large") for syntactic/collocational, and 0.53 ("small") for productive. This differentiation is, the authors propose, "one of the unique contributions of the study" (p. 745). They also included several studies with delayed or follow-up measures on each of the three dimensions, thus addressing Boulton and Cobb's (2017) wish-list. These showed that over time the two "small" effect sizes became negligible while the "large" one for syntactic/collocational learning remained large at 0.88 SDUs. This is a confirmation and specification of both Mizumoto and Chujo's and Boulton and Cobb's findings.

[2] Here the researchers use the term "effective" to include "efficient" without reference to study design, unlike Boulton and Cobb (2017) who separated the two terms.

The rest of the MV analysis confirmed some earlier points and added some interesting new ones. The authors found strong effects for Middle East studies; for studies of interventions with a duration of 10 weeks or more; for studies where concordance lines were pre-selected for clarity; for studies of interventions that included hands-on software use; and for studies involving intermediate and high-proficiency learners. They found equivalent strong effects for studies involving both language specialists and general students; both immediate and longer-term study periods; both with and without training in the use of corpus tools; and using both large public and small local corpora.

3.3.6 Ueno and Takeuchi (2023)

RQ1. "What are DDL's overall effects on short- and long-term retention of L2 learning?"
RQ2. "Which empirically motivated variables moderate DDL's effects on L2 learning?" (p. 2)

Much of the agenda here followed explicitly from Boulton and Cobb's (2017) wish-list: the number of studies was updated to include 2020; a growing number of delayed post-tests were worked into the analysis; the focus was broadened out from just vocabulary and lexicogrammar. The scope was also expanded to include conference presentations and other non-peer-reviewed studies.

The line-up included 80 studies performed after Boulton and Cobb's (2017) cut-off, a total of 532 documented (not necessarily published) DDL studies of different language features, which was reduced to 144 for the usual reasons with 422 individual samples/effect sizes; these were divided among MV variables that were largely those from Boulton and Cobb. Separate statistical analyses were provided for the four main data-sets, giving overall effect sizes of 1.12 SDUs for pre/post-test studies; 1.04 for pretest/delayed post-test studies; 0.74 for control/experimental studies, and 0.47 for control/experimental/delayed post-test studies. The authors called the first three of these "medium" and the last "small", though these terms may again be misleading if the reader comes away with the potentially erroneous interpretation that DDL is not very effective. Given our earlier point about "neighborhoods," and that Plonsky and Oswald's benchmarks were for immediate post-test (which sets the bar higher than delayed post-tests), a more accurate interpretation might be one medium and one medium-plus (immediate post-tests), and two possibly large effect sizes (delayed).

The MV findings were rich; the authors highlighted similarities and differences with Boulton and Cobb (2017) where possible, making both comparison and novelty clear. Among the most salient findings, the earlier strong effect size for Middle East studies was confirmed and expanded to include Africa, which in turn is part of a general pattern of stronger effects for foreign than second language settings. The number of delayed post-tests seems to have increased and the effect size for pre/delayed is described as "small to medium" – no problem with "medium" here, which refers to 1.04 SDUs. Within the environment MV, enough studies have been found to separate primary and secondary schools, with strong effect sizes for both. Perhaps most encouraging, studies were found that applied DDL to areas well beyond lexicogrammar, including grammar and pragmatics, with small but encouraging findings from both pre/post-test and pre/delayed post-tests. There were interesting "equally large" (or "equally medium") findings: for all types of interactions (concordancer, CALL program, paper concordance lines); for all study durations investigated; for all places where DDL activity might take place; for all learner proficiency levels; and for all publication types.

3.3.7 Ngo and Chen (2024)

RQ1. “What is the overall effect size of corpus use in ESL/EFL writing?”
RQ2. “What moderating factors contribute to the effects of corpus use in ESL/EFL writing?” (p. 3)

[429] This meta-analysis focuses on DDL studies of learning to write, 2000-2022 (i.e., subsequent to Boulton & Cobb’s, 2017, cull of studies up to 2014). The approach follows the latters’ methodology but uses a three-level analysis to counteract non-independence in studies with more than one effect size and uses Hedge’s *g* not Cohen’s *d* to calcuate effect sizes (believed to produce a slightly more accurate figure). The dimensions are 56 effect sizes in 30 studies, with an overall effect size of g=0.95, or “large” in the context of control/experimental designs. This is identical to Boulton and Cobb’s original finding for control/experimental DDL studies as a whole, d=0.95, which is encouraging, or perhaps more: on their MV of corpus use in writing, Boulton and Cobb (2014) had an SDU of just d=0.28 on the basis of 14 effect sizes, while in the present study writing is g=0.95 for 56 effect sizes. The stronger finding here thus reflects the MV “publication year,” which in the present study is a major source of both more and stronger effect sizes reflecting the growth of the DDL community, pedagogy, and technology.

Other MV results confirm earlier findings and add some new ones. Confirmations include the suitability of DDL for all proficiency levels; the abiding precedence of lexicogrammar, though reduced with strong results for discourse and structure; and equal strong effects for general and “DIY” (do-it-yourself) corpora over other types. New findings include the superiority of short and medium study periods over long (suggesting that pedagogies may fail to evolve in line with learners’ mastery of the technology); strong effects for both hands-on and -off DDL, though possibly for different learners stages of learning; and an unexpected twist involving the role of the teacher: large effects are found for both presence and absence of teacher assistance and feedback in corpus consultation, again explained as referencing different types of learners/stages of learning, but in either case supporting the attainability of independent learning within a DDL framework.

These results confirm several of Boulton and Cobb’s (2017) findings, add new ones, and plug some gaps. But they fail to plug an important gap, the lack of delayed post-testing. If all their 30 studies involve pre/post-testing, some must have involved delayed post-testing, but none are reported. On the bright side, the present researchers offer a model for handling large effects based on few studies: they state the number of effect sizes behind every MV, and where large effects come from few studies (e.g. for “high school” in “education level” based on two) they call for further investigation of this factor.

4. Methodological issues

4.1 Primary DDL research

4.1.1. Scope and sources

Broadly conceived, methodological issues encompass all information reported in the “Methods” sections of empirical studies, including description of the sampled population, activities, and research design. This section summarizes methodological characteristics of the empirical DDL research, drawing primarily on the most comprehensive and recent syntheses (Boulton, 2021; Boulton & Vyatkina, 2021; Boulton & Vyatkina, 2024) but also other DDL syntheses and meta-analyses (see 2.2 and 2.3).

4.1.2. Population characteristics

Boulton and Vyatkina (2021) found that most of the 489 empirical DDL studies published through 2019 were conducted at universities (85%) in language for general purposes (66%) with intermediate proficiency (68%) learners of L2 English (89%). While the same sampling contexts dominate the whole field of applied linguistics (Plonsky, 2023), their prevalence in DDL research is up to 30% greater. Boulton and Vyatkina (2024) did not find any considerable changes in these sampling patterns in the recent five years either. The geographic locations of DDL studies varied considerably by country but clustered in the three main regional hubs: Eastern Asia (a growing cluster), the Middle East, and Europe, with far fewer studies conducted in the Americas, Australasia, and Africa. It is therefore clear that DDL researchers need to diversify their sampling approaches, including more research into DDL for languages other than English (see Vyatkina, 2020a). One encouraging trend is that a growing number of recent studies focused on lower proficiency learners, although proficiency has been reported inconsistently, here as generally in applied linguistics research (Plonsky, 2023).

4.1.3. Activity characteristics

Regarding the linguistic focus, most DDL activities aimed at the development of lexical or lexicogrammatical L2 knowledge, with writing consistently being the favorite – and even growing in popularity among the language skills. The duration has been reported inconsistently (like L2 proficiency), in various combinations of time units, which complicates comparison across studies. In cases when duration was reported, it remained stable from 2008 at about 13 hours or 9 sessions over a semester (Boulton & Vyatkina, 2021). Researchers have consistently called for more longitudinal work, as being more ecologically valid than single-session laboratory studies, a call that has yet to be substantially addressed. All recent syntheses note a gradual increase in the size of corpora used for corpus exploration, a logical shift as larger corpora become available in open access. In contrast, the range of *types* of corpora has been decreasing, with a preference for large general corpora with built-in search and analysis tools (such as COCA and the BNC), followed by custom-designed corpora and academic corpora (such as MICASE and BAWE). Other types of corpora have been used infrequently, including web corpora, multimodal corpora, learner corpora, and corpora of literature or graded readers. Concordancers have been and remain the tool of choice for learner-corpus interaction, with 65% of the studies consistently using them throughout the timeline. Other CALL tools and prepared materials have been used less frequently, although studies addressing innovative tools with DDL affordances (e.g., YouGlish, PlayPhrase, AI tools) have recently started emerging.

4.1.4. Research design characteristics

Boulton and Vyatkina (2021: 78) summarize the following trends regarding research objectives and data collection instruments in DDL research published through 2019:

> Just over a quarter (26%) overall use a corpus as a reference resource, compared to 42% looking at learning outcomes [retention]. It is noteworthy that the [high-impact journal] papers are less focused on outcomes than the others (34% vs 44%), perhaps as pre/post-test designs are fairly obvious and easy to administer, and are relatively more interested in learners' behavior (32% vs 20%), which is more difficult to track. Other objects of

> study include the processes involved (23%) and, especially, the participants' attitudes and perceptions (…) (56%).

They also note that most studies used more than one instrument, two on average per study, typically combining tests and questionnaires.

Quantitative designs have been prevalent with roughly 50% of the studies employing inferential statistics and 35% reporting some descriptive statistics (raw frequencies and percentages), while qualitative studies account for about 15% (Boulton & Vyatkina, 2021). These proportions have remained relatively stable since 2005, although publications in high-impact journals tend to include a higher share of statistical analyses. Nevertheless, several syntheses point out insufficient rigor in both the justification for the method selection and the reporting of results.

The abovementioned distribution of objectives, instruments, and designs has not changed in the last five years. One value that has increased is the median group size, especially in articles published in high-impact journals (54 vs. 29 in earlier research; Boulton & Vyatkina, 2024). While this is a generally welcome development in quantitative research, the variation among studies has been very high (SD=54, Boulton & Vyatkina, 2021). Furthermore, most studies considered only DDL groups with barely a third including a control group for either non-DDL teaching or no teaching. Finally, only a few studies included more than one experimental group (i.e., groups conducting different DDL activities), although the number of such studies has recently been growing.

4.2 Secondary DDL research

Secondary research or synthesis studies have provided important insights into how DDL has been used for language learning and how effective it has been. There are, however, some methodological issues associated with secondary research, especially meta-analyses, that have occurred to us at least once while researching this chapter and so warrant a mention here. First, we find very small sample sizes behind some of the findings (sometimes a single study, especially for delayed post-tests). This is especially problematic with MV analyses (e.g., Lee *et al*., 2019), where it may be asked whether this area of research was ready to be meta-analyzed at this level of detail with the number of studies available, as highlighted by Plonsky *et al*. (2023: 316):

> Historically, social science researchers have mistakenly attributed observed variation in effect sizes to substantive factors (e.g., type of teaching method, level of school district resources) when in fact, a large part of that variation could be more parsimoniously attributed to small sample sizes.

Second, a danger lies in the calculations themselves: while it is generally preferable for meta-analysts to reanalyze the data themselves to ensure consistency (e.g., from *p*-values to effect sizes), this can lead to anomalies. Third, sometimes no distinction is made between positive and negative effect sizes (e.g., Ueno & Takeuchi, 2023), which might simply be reported as "large," even though a strong negative effect size should be evidence *against* the DDL intervention. It is highly unusual for scores to be higher at pre-test than at post-test, and when it happens, it apparently indicates either "unlearning" or confusion. Fourth, discussion sections are sometimes

dense with *p*-values for every effect size (e.g., Lee *et al*., 2019; Ueno & Takeuchi, 2023); however, discussion of statistical significance measured with the *p*-value is not recommended or even mentioned in Plonsky *et al*.'s (2023) practical guide to meta-analysis and is becoming increasingly controversial in social sciences in general (see Amrhein, Greenland & McShane, 2019; Trafimow & Marks, 2015). There may be a place for significance testing in this type of research, but it must be motivated (e.g., in hierarchical/regression analysis of MVs). Fifth is a minor quibble about the means of calculating effect size: though Cohen's *d* is the usual method for this (difference between means divided by pooled standard deviation), employed by Boulton ~~et al.~~and Cobb (2017) and Plonsky ~~et al.~~and Oswald (2014) and performable by any tyro researcher on an Excel spreadsheet, many researchers (e.g., Ngo & Chen, 2024) instead choose to employ Hedge's *g*, a far more complex calculation performed only with software like *R*, but with no mention of any advantage gained. All these issues raise questions about both learning from corpora and doing meta-analysis.

A final less technical issue is the extent to which meta-analysts need to know the research area they are working in. We have gone from a dearth to a plethora of syntheses and meta-analyses in applied linguistics, to the extent that it now seems necessary to remind synthesists to read and understand the studies their trawls pull up – as opposed to just "scraping" them for input to plug into the by now well-known formula for meta-analysis – and should probably work with an expert in that field if they are not experts themselves.

5 Innovative studies in DDL

This section examines three recent empirical DDL papers, outlining their significance, the methodologies (and problems with them), and the main results and implications.

5.1 Liu and Gablasova (2023)

This first paper compares the effectiveness of two corpus-based approaches (hands-on DDL and work with a corpus-informed dictionary) and a non-corpus-based approach for teaching L2 collocations, finding an advantage for DDL. This study is representative due to its typical DDL context (intact university classes with advanced learners of English) and focus (L2 collocations). It was conducted in China, a region with growing popularity of DDL teaching and research. However, the study also has several novel characteristics. First, it was longitudinal, both in terms of the duration of the DDL intervention (11 weeks) and the inclusion of a three-month delayed posttest, and featured two experimental groups. Second, it utilized #LancsBox, a novel suite of open-access DDL resources, which, in addition to the ubiquitous concordancer, included a "graphical collocations" tool that visualized collocation strength with lines of different thickness and different colors. Third, in addition to commonly used vocabulary recognition tests and learner attitude questionnaires, the study explored learner confidence in their test responses.

The study employed a sound quasi-experimental design with a relatively large sample size: 100 participants were randomly divided into three groups of equivalent mean age (20 years old) and gender distribution (predominantly female), length of English study, and L2 proficiency (B2–C1 CEFR). Learning gains were measured with a receptive vocabulary knowledge test that contained 100 verb-noun, adverb-verb, and adverb-adjective collocations taken from and in proportions reflective of the Academic Collocation List (Ackermann & Chen, 2013). Participants

chose which of the two given pairs of words were collocates (e.g., *strongly agree* or *greatly agree*) and indicated their level of confidence in each choice (0%, 25%, 50%, 75%, or 100%). Instruction in all groups consisted of 1.5-hour writing sessions. Learners wrote an academic essay during the first session and revised it during the second, this procedure being repeated four times for eight weeks total. They were encouraged to focus on collocations in writing and revising. The groups differed in the materials learners used during the revision: the "DDL" group used #LancsBox, the "OCD" group used the corpus-informed *Oxford Collocations Dictionary*, and the "control" group used online tools of their choice (e.g., bilingual dictionaries, translators). Learner performance was monitored to make sure they only used their assigned tools. The two experimental groups had training sessions in tool use before and halfway through the instructional period.

The study found slightly higher collocation knowledge gains for the DDL group but this advantage over other groups was very modest, as were the absolute gains (about 3 out of 100 collocations). Nevertheless, the DDL group further improved on the three-month delayed posttest while the OCD group declined, which lets the authors conclude that "longitudinal DDL intervention indeed can foster a high degree of autonomy and encourage self-directed learning" (p. 21). This result was triangulated with post-course questionnaires: while most learners in both experimental groups expressed intent to continue using their tools (DDL: 97%, OCD: 81%), only the DDL group reported actually doing so post-intervention (DDL: 68%, OCD: 0%). Interestingly, the study also found that the confidence of the DDL group participants in both their accurate and inaccurate test responses increased over time, showing that "self-directed and independent learning as part of DDL may result in incorrect conclusions being reached and held by learners" (p. 20), indicating the need for more teacher guidance. It must also be noted that the "control" group showed an improvement pattern similar to that of the DDL group, indicating that learners' use of the tools of their own choosing was also effective, but the study did not explore what tools were used. Finally, if a multivariate method (e.g., a linear mixed-level model) had been employed for data analysis (instead of a series of ANOVAs), it might have revealed more nuanced results.

5.2 Crosthwaite, Storch and Schweinberger (2020)

This study looked at learners' use of corpus information to correct writing errors. Gaskell and Cobb (2004) was the first study involving teacher-created concordance hyperlinks placed in learner writing atop their errors (see section 2.1), but with no written message and no requirement to devise their own searches (see also Vincent & Nesi, 2018, for a similar approach). Since then, different forms of this basic feedback strategy have been used in several studies, with generally positive results, particularly for collocational and lexical errors (as shown in the meta-analysis by Boulton & Cobb, 2017). Crosthwaite *et al.*'s goal here was to dig deeper into the many variables involved in corpus-based feedback, particularly written corrective feedback (WCF), i.e., the language used by instructors to indicate where an error has occurred and what sort of corpus search might supply the information to correct it.

A taxonomy of WCF types was proposed along a continuum from explicit to implicit, with theoretical rationale and literature review for each and an advance acknowledgment that there is probably a fine line between too explicit and not explicit enough which will vary across learners and error types. The four types of WCF, from least to most explicit, were (1) the error zone (not the exact error) is indicated by highlighting; (2) the error zone is highlighted and the exact error

underlined; (3) the error zone is highlighted and an error code provided; and (4) the error zone is highlighted and an indirect (non-determinative) comment or hint provided. Advanced academic writers with 6–8 hours training in corpus correction were divided into four groups by WCF condition and asked to correct four types of lexical and six types of grammatical errors in their own writing.

The findings are a complex matrix of outcomes: revision success or failure x two main error types x four WCF types x corpus consultation (yes or no). The descriptive statistics seem to show that lexical errors were largely revised successfully with corpus consultation whatever the WCF, with a slight advantage for the highlight and highlight+underline conditions, while grammar errors were largely revised successfully *without* corpus consultation, whatever the WCF, but that when a corpus was used, the revisions were actually more successful. To this picture the inferential statistics seem to add that "WCF condition [is] a significant factor across all errors" (p. 13) with highlight-only the winning condition (hence the title, "less is more").

It is a laudable endeavor to mine more deeply with finer implements into the variables of corpus-based learning. In this case, some useful nuggets have come to light, e.g., that verbal instructions make little difference to a correction (so can be eliminated from the teacher's chores, thus validating Gaskell and Cobb's approach) and that grammar errors are conditionally amenable to corpus influence (a novel finding of wide implication).

On the down side, there lingers a hangover of overwrought *p*-value reasoning in the results that both brings to mind the many arguments over recent years for effect size as the appropriate statistical model for applied linguistics (see section 3.2) and presents an outcome somewhat contradictory to that of the descriptive statistics (e.g., WCF is "statistically significant" despite being designated "a confound" in the abstract, and then later in the learners' apparent ability to correct errors whatever the WCF). There are no effect sizes in this study nor any obvious way to calculate them, meaning that the findings cannot be incorporated into any future meta-analysis of DDL or WCF research.

5.3 Charles and Hadley (2022)

One criticism of DDL is its time-consuming nature, which can be countered if it is shown that it leads to long-term changes in learning behavior: greater language awareness, noticing abilities, autonomy, etc. – steps towards becoming a better learner. Direct research on this is however still scarce. An alternative counter is that querying becomes quicker and more effective over time; few papers adopted delayed post-tests even a few weeks after instruction (11% according to Boulton & Vyatkina, 2021), and Charles and Hadley provided a yet longer view in this paper. The course itself lasted only 6 two-hour sessions but was repeated 50 times over 9 years for a total of 544 participants and included a follow-up questionnaire one year after the end of the course to assess take-up of both corpus and concordancing.

The course was reported in several papers by Charles, culminating in the overview presented here of the entire period 2009–2017. The elective class targeted graduate students needing English for academic writing. The many different disciplines and language backgrounds remained varied and stable over time. Small groups (M=9 per class) were shown how to compile their own corpora of research articles (RAs); most corpora were fairly small (fewer than 20 articles for less than 200k tokens) but allowed comparison against a collection of their own writings (see Charles, 2018, on this). The software was AntConc; Charles (2018) found that the KWIC concordance tool was rated "very useful" by 81% of the 90 respondents, followed by

Clusters at 49%; combining "useful" and "very useful", all tools rated over 75%. The courses were taught by four different tutors, initially in computer rooms but later mostly on learners' own devices, conducive to adoption out of class.

Pre-course questionnaires showed 24% already had some knowledge of corpora; by the end of the course, autonomous corpus use had risen to an average 87% (SD=2.7%), dropping again to 62% (SD=11.3%) one year later for at least occasional use. All these results are relatively stable over time, suggesting that corpora had not become "normalized" during that period. Regular autonomous use (i.e., at least once a week) rose from 11% (SD=4%) to 61% (SD=6.8%) post-course and 37% (SD=10.3%) a year later. The main reason for discontinued use was simply that the participants no longer needed to produce academic writing. The risk of course is that, once they stop, they may not return; recommendations included a "refresher" course for doctoral students, or an introduction to large general corpora at master's level. Other reasons may include technophobia, antithetic learning styles, time constraints, problems with the tools or corpora, and preference for other resources (see Charles, 2022, for further analysis).

Drawing conclusions from questionnaire data is problematic. Questionnaires are increasingly popular in gathering information about use and attitudes towards corpus use (Boulton & Vyatkina, 2021), but have several limitations. First, respondents tend to be a self-selecting group which may not be representative of the whole, though it is notable that the number returned here is relatively high: 544 (pre-course), 343 (post-course), 221 (after one year) – i.e., 41% of those enrolled, and 64% of those who completed. Attrition during the course may account for much of the initial drop; difficulties in contacting students explain much (although not all) of the rest. Actual corpus size and content were analyzable as they were stored on university servers, but triangulation from other sources would allow the verification of actual use. Other emic sources such as learner interviews would have increased depth and nuance, especially for what they used the corpus for (writing or revision, types of searches and items queried), and no indication was given of whether corpus use did in fact improve the users' academic writing. It would have been possible to cross these results with demographic data to see if, for example, uptake correlated with particular disciplines, academic level, L1 background, and so on. Additional instruments (e.g., for motivation or learning styles) might have helped explain some of the variation.

The underlying question is whether the rate of uptake makes investment in the course worthwhile. It was clearly extremely useful for the 61% of regular users at the time of writing, and 37% in the long term. We would argue that this is more than enough justification, as no tool, technique or approach will be 100% effective for 100% of learners, but with no obvious point of comparison, the assessment will be largely a question of perspective. The teacher's role is not discussed in detail but likely to be crucial in promoting enthusiasm. Personal motivation is paramount: an immediate need to write. The actual building of the research article corpus – choosing and cleaning of the texts – seems in itself to promote mastery of the domain and genre, with results more specific to the learners' own needs than any pre-prepared corpus could aspire to. Still, general corpora may be more appropriate for some needs (for example, sheer size is needed to expose many mid-frequency collocations), and other corpus-assisted academic writing tools may be preferred by some (e.g., ColloCaid); it is too early to judge the impact of the recent generation of AI tools, with initial assessments varied (e.g., Lin, 2023, on ChatGPT as a substitute "concordancer").

6 Conclusion

Corpora have found many uses in the field of language teaching and learning in the hands of decision-makers, teachers, and learners. Published research covers classroom applications for a wide variety of learner profiles and for widely varied uses, from highly controlled to entirely autonomous work, from paper-based materials to hands-on concordancing, from reference resource to learning tool. As Boulton and Cobb (2017: 386) put it, "DDL works pretty well in almost any context where it has been extensively tried", where "extensively tried" then meant at least ten studies. This underlines the highly flexible role of corpora – there is no single "right" way to use them. From a research perspective, this may lead to a perceived fragmentation of the field, hence the review of syntheses here to gain a clearer oversight of data-driven learning in particular. Our own syntheses have highlighted a number of areas in need of further research. Among other things, more longitudinal, ecological, open-ended studies are needed, especially addressing the main benefits attributed to corpus work in promoting autonomy, learning to learn, language awareness, noticing, etc. – and, consequently, in helping become "better learners". A platitude, but more research is needed to show us how to take best advantage of corpus tools and techniques for language learning, to "bridge the research-practice gap" (Chambers, 2019). How much training is needed? How much ongoing scaffolding and at what points to reduce it? Are certain learning or personal styles favored or disfavored? How is the success of such learning best measured? What is the ideal complementarity between search-based and other forms of instruction? What are the effects on learning behaviors and development of (meta)cognitive skills? These questions are central rather than peripheral to language learning in general, and to DDL in particular. We also note highly varying rigor in design and reporting for both qualitative and quantitative studies.

A final word. Traditional corpus consultation is in some ways a relatively marginal activity, to be found in few classrooms around the world. However, it is in many ways analogous with internet searches and use of other technologies for querying the vast stores of data available, which has arguably become the dominant learning mode in our culture. Language learners regularly use Google as a substitute "concordancer" to browse the web as a substitute "corpus" to help with usage in context, particularly in their writing (e.g., Han & Shin, 2017). Indeed, "Googling" is largely an invention of corpus linguists (Crystal, 2012) and the majority of internet users are busy becoming knowledge co-constructors from corpus data. This, of course, is definitely not to say that all search-based learning is accurate, permanent or worthwhile – far from it – in language learning or any other area. This is even more true of recent generative artificial intelligence based on large language models (LLMs), essentially an extended application of corpus linguistics. Unsurprisingly, researchers were quick to consider whether AI would make DDL redundant, with half a dozen papers published in the year following the appearance of ChatGPT. Some (e.g., Lin, 2023) have shown that it is possible to produce output similar to what one would get from a corpus, though her proposed instructions are long and complex and the results less than perfect. Others (e.g., Crosthwaite & Baisa, 2023) have argued that we should play to the strengths of each tool and approach, rather than forcing AI to produce corpus-like output, or abandoning corpora altogether. In particular, a tool that does everything for you makes learning redundant, while DDL puts the learner at the center by requiring careful thought to interpret the data output.

References

Abduh, A., Mulianah, A., Darmawati, B., Zabadi, F., Sidik, U., Handoko, W., Jayadi, K., & Rosmaladewi, R. 2023. The Compleat Lextutor application tool for academic and technological lexical learning: Review and bibliometric approach. *Indonesian Journal of Science & Technology*, 8(3), 539–560. https://doi.org/10.17509/ijost.v8i3.63539

Ackermann, K. & Chen, Y. H. 2013. Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Ahmad, K., Corbett, G., & Rogers, M. 1985. Using computers with advanced language learners: An example. *Language Teacher (Tokyo)*, 9(3), 4–7.

Amrhein, V., Greenland, S., & McShane, B. 2019. Scientists rise up against statistical significance. *Nature*, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Baten, L., Cornu, A.-M., & Engels, L. 1989. The use of concordances in vocabulary acquisition. In C. Laurent & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines*, 452–467. Multilingual Matters.

Bennett, G. 2010. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press. https://doi.org/10.3998/mpub.371534

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortez, V., Csomay, E., & Urzua, A. 2004. *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. ETS/TOEFL.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 2021. *Grammar of spoken and written English*. Amsterdam: John Benjamins. [Previously published in 1999 as *The Longman grammar of spoken and written English*]

Boulton A. 2007. But where's the proof ? The need for empirical evidence for data-driven learning. In M. Edwardes (Ed.), *Proceedings of the BAAL Annual Conference 2007*, 13–16. Scitsiugnil Press.

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*, 129–144. Equinox.

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*, 261–291. John Benjamins. https://doi.org/10.1075/scl.52.11bou

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

Boulton, A. 2021. Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education*, 9–34. John Benjamins. https://doi.org/10.1075/scl.102.01bou

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. & Pérez-Paredes, P. 2024. Data-driven learning: Pedagogy and technology. In R. Hampel & U. Stickler (Eds.), *Bloomsbury handbook of language learning and technology*. Bloomsbury Publishing.

Boulton, A. & Tyne, H. 2013. Corpus linguistics and data-driven learning: A critical overview. *Bulletin Suisse de Linguistique Appliquée*, 97, 97 118. https://doc.rero.ch/record/11876/files/bulletin_vals_asla_2013_097.pdf

Boulton, A. & Vyatkina, N. 2021. Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning & Technology*, 25(3), 66–89. https://doi.org/10125/73450
Boulton, A. & Vyatkina, N. 2024. Expanding methodological approaches in DDL research. *TESOL Quarterly*, 58(3), 1193–1204. https://doi.org/10.1002/tesq.3269
Butler, C. 1974. German for chemists: Teaching languages to adults for special purposes. *CILT Reports and Papers*, 11, 50–53.
Carter, R., Hughes, R., & McCarthy, M. 1998. Telling tails: Grammar, the spoken language and materials development. In B. Tomlinson (Ed.), *Materials development in language teaching*, 67–89. Cambridge University Press.
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*, 3–16. Rodopi. https://doi.org/10.1163/9789401203906_002
Chambers, A. 2019. Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, 52, 460–475. https://doi.org/10.1017/S0261444819000089i
Chambers, A., Conacher, J., & Littlemore, J. (Eds.). 2004. *ICT and language learning: Integrating pedagogy and practice*. University of Birmingham Press.
Charles, M. 2018. Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes*, 36, 15–25. https://doi.org/10.1016/j.jeap.2018.08.003
Charles, M. 2022. The gap between intentions and reality: Reasons for EAP writers' non-use of corpora. *Applied Corpus Linguistics*, 2(3), 100032. https://doi.org/10.1016/j.acorp.2022.100032
Charles, M. & Hadley, G. 2022. Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, 56, 101095. https://doi.org/10.1016/j.jeap.2022.101095
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
Chong, S.W. & Plonsky, L. 2024. A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, 15(4), 1569-1594. https://doi.org/10.1515/applirev-2022-0189
Choubsaz, Y., Jalilifar, A., & Boulton, A. 2024. A longitudinal analysis of highly cited papers in four CALL journals. *ReCALL*, 36(1). https://doi.org/10.1017/S0958344023000137
Cobb, T. 2007. Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, 11(3), 38–63. http://dx.doi.org/10125/44117
Cobb, T. 2021. Corpus for courses: Data-driven course design. *Bulletin Suisse de Linguistique Appliquée*, numéro spécial(2), 11–30.
Cobb, T. & Boulton, A. 2015. Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of English corpus linguistics* (1st ed.), 478–497. Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.027
Cobb, T., Greaves, C., & Horst, M. 2001. Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la didactique des langues secondes*, 133–153. Editions Logique.

Crosthwaite, P. & Baisa, V. 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Crosthwaite, P., Ningrum, S., & Schweinberger, M. 2023. Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics*, 28(3), 344–377. https://doi.org/10.1075/ijcl.21072.cro

Crosthwaite, P., Storch, N., & Schweinberger, M. 2020. Less is more? The impact of written corrective feedback on corpus-assisted L2 error resolution. *Journal of Second Language Writing*, 49, 100729. https://doi-org.bases-doc.univ-lorraine.fr/10.1016/j.jslw.2020.100729

Crystal, D. 2012. Searchlinguistics. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley. https://doi.org/10.1002/9781405198431.wbeal1047

Dong, J., Zhao, Y., & Buckingham, L. 2023. Charting the landscape of data-driven learning using a bibliometric analysis. *ReCALL*, 35(3), 339–355. https://doi.org/10.1017/S0958344022000222

Gabrielatos, C. 2005. Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ*, 8(4), 1–35. http://tesl-ej.org/ej32/a1.html

Gaskell, D. & Cobb, T. 2004. Can learners use concordance feedback for writing errors? *System*, 32(3), 301–319. https://doi.org/10.1016/j.system.2004.04.001

Gavioli, L. 2005. *Exploring corpora for ESP learning*. John Benjamins. https://doi.org/10.1075/scl.21

Gougenheim, G. 1958. *Dictionnaire fondamental de la langue française*. Didier.

Goulart, L. & Veloso, I. (Eds). 2023. *Corpora in English language teaching: Classroom activities for teachers new to corpus linguistics*. Montclair State University. https://pressbooks.pub/testbook123

Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching*, 13–32. John Benjamins. https://doi.org/10.1075/scl.33.04gra

Han, S. & Shin, J.-A. 2017. Teaching Google search techniques in an L2 academic writing context. *Language Learning & Technology*, 21(3), 172–194. https://doi.org/10125/44626

He, C. & Wei, X. 2019. Study of corpus' influences in EAP research (2009–2018): A bibliometric analysis in CiteSpace. *English Language Teaching*, 12(12), 59–66. https://doi.org/10.5539/elt.v12n12p59

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. Routledge.

Johns, T. 1988. Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe & H. Wekker (Eds.), *Computer applications in language learning*, 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. & King, P. (Eds.). 1991. *Classroom concordancing. English Language Research Journal*, 4.

Karpenko-Seccombe, T. 2020. Academic writing with corpora: A resource book for data-driven learning. London and New York: Taylor & Francis. https://doi.org/10.4324/9780429059926

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*. Pressbooks. https://pressbooks.pub/elenlefoll/

Lee, H., Warschauer, M., & Lee, J. H. 2019. The effects of corpus use on second language vocabulary learning: A multi-level analysis. *Applied Linguistics*, 40(5), 721–753. https://doi.org/10.1093/applin/amy012
Lin, P. 2023. ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), 100065. https://doi.org/10.1016/j.acorp.2023.100065
Liu, T. & Gablasova, D. 2023. Data-driven learning of collocations by Chinese learners of English: A longitudinal perspective. *Computer Assisted Language Learning*. Advance Online Publication. https://doi.org/10.1080/09588221.2023.2214605
Luo, Q. & Zhou, J. 2017. Data-driven learning in second language writing class: A survey of empirical studies. *International Journal of Emerging Technologies in Learning*, 12(3), 182–196. https://doi.org/10.3991/ijet.v12i03.6523
Lusta, A., Demirel, Ö., & Mohammadzadeh, B. 2023. Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Helyon*, 9(12), E22731. https://doi.org/10.1016/j.heliyon.2023.e22731
Martinez, R., & Schmitt, N. 2012. A phrasal expressions list. *Applied Linguistics*, 33(3); 299–320. https://doi.org/10.1093/applin/ams010
McCarthy, M., McCarten, J., & Sandiford, H. 2006. *Touchstone 4* (Teacher's Edition). Cambridge University Press.
McEnery, T. & Wilson, A. 1997. Teaching and language corpora. *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. http://mizumot.com/files/ecs2015.pdf
Mohamed, S. & Acklam, R. 1995. *Intermediate choice* (Students' Book). Longman.
Nation, I. S. P. & Beglar, D. 2007. A vocabulary size test. *The Language Teacher*, 31(7), 9–13.
Ngo, T. T.-N., & Chen, H. H.-J. 2024. The effectiveness of corpus use in ESL/EFL writing: A meta-analysis. *Language Teaching Research*. Advance Online Publication. https://doi.org/10.1177/13621688241260183
O'Sullivan, Í. 2007. Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3), 269–286. https://doi.org/10.1017/S095834400700033X
O'Sullivan, Í. & Chambers, A. 2006. Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15(1), 49–68. https://doi.org/10.1016/j.jslw.2006.01.002
Park, H. & Nam, D. (2017). Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. *Linguistic Research*, 34(3), 427–457. https://doi.org/10.17250/khisli.34.3.201712.008
Pérez-Paredes, P. 2022. A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1–2), 36–61. https://doi.org/10.1080/09588221.2019.1667832
Pinto, P. T., Crosthwaite, P., de Carvalho, C. T., Spinelli, F., Serpa, T., Garcia, W., & Ottaiano, A. O. (Eds.). 2023. *Using language data to learn about language: A teachers' guide to classroom corpus use.* University of Queensland. https://uq.pressbooks.pub/using-language-data

Plonsky, L. 2023. Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, 8(1), 75. https://doi.org/10.3390/languages8010075
Plonsky, L., Hu, Y., Sudina, E., & Oswald, F. 2023. Advancing meta-analytic methods in L2 research. In A. Mackey & S. Gass (Eds.), *Current approaches in second language acquisition research: A practical guide*, 304–333. Wiley. https://doi.org/10.1002/9781394259670.ch14
Plonsky, L. & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. https://doi.org/10.1111/lang.12079
Reppen, R. 2010. *Using corpora in the classroom*. Cambridge University Press. https://doi.org/10.1017/9781139042789.003
Reppen, R., Blass, L., Iannuzzi, S., Bunting, J. D., & Diniz, L. 2020. *Grammar and Beyond, with academic writing* (2nd edition). Cambridge: Cambridge University Press.
Römer, U. 2006. Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 121–134. https://doi.org/10.1515/zaa-2006-0204
Schmidt, R. 1990. The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158.
Schmitt, N. & Schmitt, D. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484–503. https://doi.org/10.1017/S0261444812000018
Sinclair, J. (Ed.). 1987. *Looking up: An account of the COBUILD project in lexical computing*. Collins.
Spada, N. & Tomita, Y. 2010. Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x
Stevens, V. 1991. Concordance-based vocabulary exercises: A viable alternative to gap-filling. In T. Johns and P. King (Eds.), *Classroom concordancing. ELR Journal*, 4, 47–61.
Sun, W. & Park, E. 2023. EFL learners' collocation acquisition and learning in corpus-based instruction: A systematic review. *Sustainability*, 15, 13242. https://doi.org/10.3390/su151713242
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
Thornbury, S. 2004. *Natural grammar: The keywords of English and how they work*. Oxford University Press.
Thorndike, E. & Lorge, I. 1944. *The teacher's word book of 30,000 words*. Columbia University.
Thurstun, J. & Candlin, C. 1997. *Exploring academic English: A workbook for student essay writing*. CELTR.
Trafimow, D. & Marks, M. 2015. Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991
Ueno, S. & Takeuchi, O. 2023. Effective corpus use in second language learning: A meta-analytic approach. *Applied Corpus Linguistics*, 3(3), 100076. https://doi.org/10.1016/j.acorp.2023.100076
Vincent, B. & Nesi, H. 2018. The BAWE Quicklinks project: A new DDL resource for university students. *LIDIL*, 58, 1–17. https://doi.org/10.4000/lidil.5306
Vyatkina, N. 2020a. Corpora as open educational resources for language teaching. *Foreign Language Annals*, 53(2), 359–370. https://doi.org/10.1111/flan.12464

Vyatkina, N. (Ed.). 2020b. *Incorporating corpora: Using corpora to teach German to English-speaking learners*. University of Kansas Open Language Resource Center. https://corpora.ku.edu

West, M. 1953. *A general service list of English words*. Longman.

Yoon, C. 2011. Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10, 130–139. http://dx.doi.org/10.1016/j.jeap.2011.03.003