LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Corpus Use in Language Learning: A Meta-Analysis

Alex Boulton and Tom Cobb

Université de Lorraine and Université du Québec à Montréal

This study applied systematic meta-analytic procedures to summarize findings from experimental and quasi-experimental investigations into the effectiveness of using the tools and techniques of corpus linguistics for second language learning or use, here referred to as data-driven learning (DDL). Analysis of 64 separate studies representing 88 unique samples reporting sufficient data indicated that DDL approaches result in large overall effects for both control/experimental group comparisons ($d = 0.95$) and for pre/posttest designs ($d = 1.50$). Further investigation of moderator variables revealed that small effect sizes were generally tied to small sample sizes. Research has barely begun in some key areas, and durability/transfer of learning through delayed posttesting remains an area in need of further investigation. Although DDL research demonstrably improved over the period investigated, further changes in practice and reporting are recommended.

We would like to thank the participants at the Teaching and Language Corpora conferences where earlier versions of this paper were presented and especially Luke Plonsky, Lourdes Ortega, and John Norris for their invitation to a symposium on meta-analysis at the International Association for Applied Linguistics in 2014 in Brisbane kindly sponsored by *Language Learning*. Our thanks to Luke Plonsky again for his input on an earlier draft of this paper as well as to the anonymous reviewers. We are also grateful to the authors and coauthors who responded to our e-mails and in some cases managed to provide papers or further information on their studies: Kiyomi Chujo, Susan Conrad, Averil Coxhead, Ewa Donesch-Jezo, Laura Gavioli, Zeping Huang, Ali Akbar Jafarpour, Betsy Kerr, Hsien-Chin Liou, Gillian Mansfield, Daehyeon Nam, Yasunori Nishina, Kathryn Oghigian, Simon Smith, and Serge Verlinde (whether their papers could finally be included or not).

This article has been awarded Open Materials and Open Data badges. All materials and data are publicly accessible via the Open Science Framework at https://osf.io/jkktw. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Alex Boulton: Atilf – CNRS/Université de Lorraine, 44, avenue de la Libération, BP 30687, 54063 Nancy Cedex, France. E-mail: alex.boulton@univ-lorraine.fr

DOI: 10.1111/lang.12224

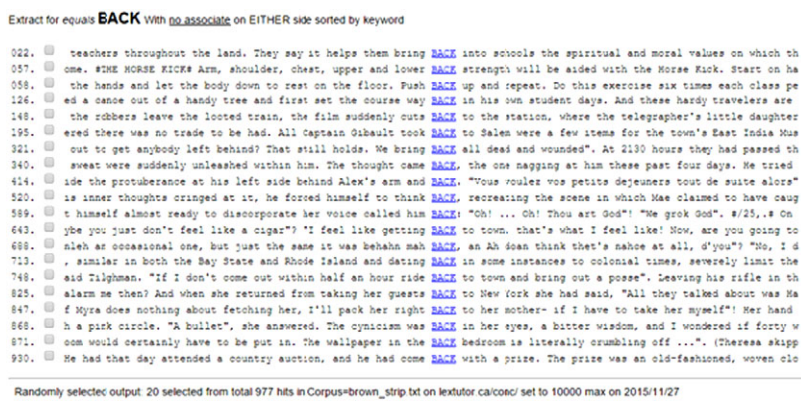
Keywords corpus-based language learning; data-driven learning; DDL; meta-analysis; research synthesis

Introduction

The purpose of the present synthesis was to answer the question of whether there are, on the whole, positive learning outcomes resulting from language learners' use of the tools and techniques of corpus linguistics in what has come to be known as *data-driven learning* (DDL; Johns, 1990). This approach typically involves getting foreign or second language (L2) learners to work with written or spoken corpus data. Figure 1 shows a typical concordance output for the word *back*, the learner's task being to identify idiomatic versus literal uses, assess their relative frequencies, verify these frequencies in 20 random lines from a different corpus, and then compare the concordance data with the entry for this word in a range of dictionaries.[1] Although the approach itself may seem rather time consuming for a single word, it is the processes that are important, and DDL may help learner development if it leads, for example, to increased language sensitivity, noticing, induction, and ability to work with authentic data.

The Case in Principle: Arguments For and Against

The DDL approach is geared to making sense of language input but has several potential advantages that other input approaches do not. Core among these is that input assembly replaces input simplification, thus maintaining authenticity


Extract for equals **BACK** With no associate on EITHER side sorted by keyword

022. teachers throughout the land. They say it helps them bring BACK into schools the spiritual and moral values on which th
057. ome. #THE HORSE KICK# Arm, shoulder, chest, upper and lower BACK strength will be aided with the Horse Kick. Start on ha
058. the hands and let the body down to rest on the floor. Push BACK up and repeat. Do this exercise six times each class pe
126. ed a canoe out of a handy tree and first set the course way BACK in his own student days. And these hardy travelers are
148. the robbers leave the looted train, the film suddenly cuts BACK to the station, where the telegrapher's little daughter
195. ered there was no trade to be had. All Captain Gibault took BACK to Salem were a few items for the town's East India Mus
321. out to get anybody left behind? That still holds. We bring BACK all dead and wounded". At 2130 hours they had passed th
340. sweat were suddenly unleashed within him. The thought came BACK, the one nagging at him these past four days. He tried
414. ide the protuberance at his left side behind Alex's arm and BACK. "Vous voulez vos petits dejeuners tout de suite alors"
520. is inner thoughts cringed at it, he forced himself to think BACK, recreating the scene in which Mae claimed to have caug
589. t himself almost ready to discorporate her voice called him BACK: "Oh! ... Oh! Thou art God"! "We grok God". #/25,.# On
643. ybe you just don't feel like a cigar"? "I feel like getting BACK to town. that's what I feel like! Now, are you going to
688. nleh ar occasional one, but just the same it was behahn mah BACK, an Ah doan think thet's nahce at all, d'you"? "No, I d
713. , similar in both the Bay State and Rhode Island and dating BACK in some instances to colonial times, severely limit the
748. aid Tilghman. "If I don't come out within half an hour ride BACK to town and bring out a posse". Leaving his rifle in th
825. alarm me then? And when she returned from taking her guests BACK to New York she had said, "All they talked about was Ha
847. f Myra does nothing about fetching her, I'll pack her right BACK to her mother- if I have to take her myself"! Her hand
868. h a pink circle. "A bullet", she answered. The cynicism was BACK in her eyes, a bitter wisdom, and I wondered if forty w
871. oom would certainly have to be put in. The wallpaper in the BACK bedroom is literally crumbling off ...". (Theresa Skipp
930. He had that day attended a country auction, and he had come BACK with a prize. The prize was an old-fashioned, woven clo

Randomly selectec output 20 selected from total 977 hits in Corpus=brown_strip.txt on lextutor.ca/cosc/ set to 10000 max on 2015/11/27

Figure 1 Concordance output from the Brown corpus for the word *back* (http://www.lextutor.ca/conc).

of language. Another advantage lies in identifying which forms and meanings in a language (whether words, structures, pragmatic patterns, etc.) are most frequent and thus probably most worth knowing. DDL consists of the consultation of language data by learners themselves and thus incorporates the notions of learner autonomy, induction, exemplar-based learning, and constructivism, in the sense of letting learners discover linguistic patterns for themselves (with varying degrees of guidance) rather than being spoon-fed predigested rules. In addition to these mainly pedagogical principles that have been extensively discussed in the literature, DDL has a number of potential theoretical advantages as a method of making input comprehensible.

1. DDL reflects current linguistic theory. Language is increasingly seen as dynamic, complex, probabilistic, interactive, and patterned rather than rule governed, as highlighted by current usage-based theories of language (e.g., Tomasello, 2003). Reflecting this, linguistic knowledge can be conceived of as a mental corpus (Taylor, 2012) of combined experiences of language use. Corpus linguistics has given rise to many insights to this patterning, including the idiom principle (Sinclair, 1991), lexical priming (Hoey, 2005), and norms and expectations (Hanks, 2013). DDL helps learners to recognize the fuzzy nature of authentic language use in context, essential if they are to deal with it.
2. DDL reflects current learning theory. Rules are hard, patterns are easy; or rather, rules are an artificial intellectual abstraction, whereas the human brain is programmed to detect patterns in the world around us (e.g., Barrett, Dunbar, & Lycett, 2002), including while learning and using language. This coincides with constructivist principles (Cobb, 1999) and allows the learner to proceed toward the target norm by progressive approximations (Aston, 1998, p. 13). DDL arguably promotes such skills, which should be transferable to new contexts and thus produce better learning outside the classroom, increasing learner autonomy and lifelong learning for them.
3. DDL reflects current psycholinguistic theory. Because pattern induction is a natural process, it reduces the cognitive load of processing (Sweller, Ayers, & Kalyuga, 2011), freeing up resources for the still considerable effort of meaning construction required on the part of the learner—effort being the often missing component in instructed language learning and a reliable predictor of both depth of processing and retention (Hulstijn & Laufer, 2001). DDL further provides access to the massive amounts of authentic language needed (input flood) but, crucially, it organizes it to

make patterns salient, as is necessary for noticing (Schmidt, 1990). There is also increasing psycholinguistic evidence for the importance of chunking (Millar, 2011), which DDL helps to highlight.

4. DDL reflects current second language acquisition (SLA) research findings. A rebalancing of language instruction from overemphasis on meaning and top-down processing to inclusion of form-focus and bottom-up processes has been recommended for years (Doughty & Williams, 1998), but in practice the result has often seemed to lead back to vocabulary lists and grammar exercises. DDL offers a way forward on this front.
5. DDL reflects and may inform existing learner practice. Learners are already involved in using information and communication technology (ICT) to search for answers to their language questions, especially via the use of Google as a "concordancer" for the Web as "corpus" (Chinnery, 2008; Kilgarriff & Grefenstette, 2003). Properly conceived DDL activities can build on these existing behaviors, refining them and using them as a way in to corpus work (Boulton, 2015).

Although advocates of DDL are quick to seize on such alleged advantages, others are equally quick to point out potential counterarguments. These are not articulated often in the DDL research literature, which is typically cast from a positive or even enthusiastic perspective, and hence suffers from a version of publication bias. Among the most frequently cited criticisms: many language teachers and students remain uncomfortable with computer work generally (even if they use ICT for other purposes on a daily basis); chopped-off concordance lines may help expose patterns yet be off-putting to some and are not designed for gaining meaning as traditionally conceived via linear reading; most corpora are composed of authentic native language well beyond the comfort level of many learners; and DDL work requires substantial training, and the processes are time consuming when learners could simply be told or use pedagogically derived resources such as dictionaries. Even many of our colleagues who are provisionally supportive of concordance work would nonetheless prefer to constrain its use to specific situations such as teaching advanced learners only or restrict the implementation to simplified corpora only, to paper concordance lines only, to use as a reference tool rather than a learning aid, or to vocabulary/collocation acquisition or awareness raising. These misgivings, although vaguely defined and forming an informal, fugitive literature, definitely exist, deserve to be taken seriously, and may find either support or rebuttal in what we uncovered in this meta-analysis.

The Need for Evidence: From Primary Research to Meta-Analysis

The in-principle case for or against DDL might seem irresolvable. An alternative is to treat it as an empirical matter, and this is the motivation for this effects-oriented meta-analysis. Our definition of the field as the use of the tools and techniques of corpus linguistics for L2 learning or use gives a deliberately broad remit. This is not an original definition, mirroring as it does that offered by Gilquin and Granger (2010, p. 359), though it clearly goes beyond the narrow view of DDL as entirely autonomous, serendipitous corpus browsing—a strawman view that has been seized upon perhaps more by critics than advocates because it is "doubtful . . . whether this can be fruitfully put into practice in the reality of ELT classrooms" (Mukherjee, 2006, p. 14). The broader definition proposed here thus makes sense from a pedagogical and theoretical perspective (cf. Boulton, 2011). It is also the goal of this type of synthesis "to make generalizations . . . across a range of populations and scenarios" (Plonsky & Ziegler, 2016, p. 19). For present purposes, this definition allowed us to cover studies that researchers have offered as examples of DDL rather than inventing some arbitrary or subjective cutoff point. For example, although most DDL has been inductive, it may include some deductive practices (e.g., Oghigian & Chujo, 2010), which have thus been included here. Conversely, the vast majority of inductive language work has made no use of corpus tools or techniques and was thus excluded as it did not represent DDL in the commonly accepted sense.

The studies have variously reported both hands-on concordancing and use of printed concordances or worksheets, large general corpora and small tailor-made ones, corpora as learning aids or reference resources, and so forth. A potential drawback to a broad collection of studies lies in the possibility of comparing apples and oranges. A meta-analysis only works insofar as the studies included can be treated as approximate replications of the same phenomenon (Lipsey & Wilson, 2001) among the same basic population. The sheer variety of research questions and designs in DDL initially led Boulton (2012) to wonder if a meta-analysis was possible at all. In fact, the studies included in a meta-analysis can rarely be considered identical (Borenstein, Hedges, Higgins, & Rothstein, 2009), and studies that appear similar to one researcher may be different to another (Lipsey & Wilson, 2001). In the end, it is a question of granularity because meta-analyses have been conducted on far wider bases, such as whether computer-assisted language learning (CALL) is effective (e.g., Grgurović, Chapelle, & Shelley, 2013; Felix, 2008), whether language instruction makes a difference (Norris & Ortega, 2000, 2001), or what factors affect educational achievement (Hattie, 2009). To launch this meta-analysis, we only

needed studies that were "similar enough" (Norris & Ortega, 2006, p. 216). A broader domain for inclusion better reflects the diversity of practices in real-world contexts and increases generalizability, and a larger data set increases power and accuracy (Plonsky & Brown, 2014). In line with most meta-analyses, assuming a random-effects model then allowed us to proceed to the more interesting part of the analysis: whether variation in effect sizes can be attributed to differences in study designs, populations, language focus, implementation or other features, all of which can be treated as moderator variables (Cumming, 2012). In other words, our meta-analysis began with a collection of fruit, with a view to separating apples and oranges later on (Ortega, 2010).

A meta-analysis synthesizes quantitative results in the form of effect sizes. These are simple to calculate and understand in terms of the basic formulae used and can be averaged across studies. A meta-analysis nevertheless offers a daunting array of choice-points throughout its execution. In preparing this study, we drew on a number of existing meta-analyses in applied linguistics beginning with Norris and Ortega (2000), as well as various manuals and textbooks (e.g., Cumming, 2012; Lipsey & Wilson, 2001) and notable articles examining good practice (e.g., Plonsky & Oswald, 2014; Plonsky & Ziegler, 2016). Due to its nature, a meta-analysis cannot accommodate qualitative studies. Their findings may nonetheless help to inform and validate all types of studies in future. For a comparison of meta-analyses and narrative surveys, see Norris and Ortega (2006) and the special issue of *Applied Linguistics* edited by R. Ellis (2015).

An earlier attempt at meta-analyzing DDL research (Cobb & Boulton, 2015) reported effect sizes of $d = 1.68$ for within-groups (pre/posttests) and $d = 1.04$ for between-groups (control/experimental) designs. We stressed however that this was only a preliminary study and that the high d values were worth a second, critical look in a more principled meta-analysis. The most obvious change here is the increase from 21 to 88 unique samples drawn from a wider, more exhaustive and up-to-date trawl of papers (e.g., Ph.D. dissertations, use of more databases, more recent cutoff point). Other differences include more rigorous extraction of effect sizes (including values derived from t and F tests and missing data solicited directly from the authors) as well as the formulae used and their interpretation (unbiased d, winsorizing), inclusion of a separate effect size for every unique sample in each study, and a detailed coding manual with calculation of subeffect sizes allowing a post hoc moderator analysis in an attempt to identify what may be responsible for variation between studies (entirely absent in the previous article). Although there have been various other attempts at synthesis of DDL (e.g., Boulton, in press; Chambers, 2007), to our

knowledge the only other meta-analysis to date is by Mizumoto and Chujo (2015). Their survey of 14 studies arrived at an overall effect size of $d = 0.97$ and identified lexicogrammar as the most promising area for DDL work ($d = 2.93$). It should be noted that their study drew only on within-groups designs in papers coauthored by Chujo and was exclusive to the Japanese context among lower-proficiency learners of English. The present meta-analysis sought to broaden this scope.

We therefore addressed three main research questions:

1. How much DDL research is there?
2. How effective is DDL, and how efficient is it?
3. How can we best account for any variation observed?

Method

Data Collection

The first stage in the data collection process was to assemble as many studies as possible conforming to the definition given above. Ideally the result would be an exhaustive collection, though practical considerations meant that a number of further criteria needed to be applied. In the present case, we included only full-text descriptions that were publicly available and text types where we were relatively likely to access the majority of the work. We did not institute start or finish cutoff points, accepting any study up to and including the first 6 months of 2014, though it is likely that some prior publications will be referenced online at a later date.

Because our goal was to gain a comprehensive view of the entire field, warts and all, the solution was to include all studies initially, then to compare factors possibly related to quality as moderator variables. This allowed us to avoid quality judgments that are highly subjective (Lipsey & Wilson, 2001); for example, it is common for double-blind reviews of the same paper to come to quite opposite recommendations. It also avoided indirect quality indicators such as restricting the perimeter to internationally recognized peer-reviewed journals, thus excluding much that may be of value and introducing publication bias. Burston (2013) reported that nearly 60% of all MALL (mobile-assisted language learning) studies would be excluded under such a criterion. Nonetheless, publication bias is still likely insofar as studies that do not produce significant differences or other desired outcomes are more likely to remain unpublished altogether (N. Ellis, 2006), the famous *file-drawer problem* (Rosenthal, 1979).

Extending the trawl to fugitive or grey literature rather than attempting to judge the quality of the papers of course leaves a meta-analysis open to

charges of garbage in, garbage out. Our intention here was to begin inclusive and to let the analysis itself show whether the different source types influenced the outcomes. In other words, quality can be treated as an empirical part of the investigation (Norris & Ortega, 2006, pp. 18–19; Ortega, 2010, p. 121; Plonsky, 2014, p. 466) and may give rise to recommendations for future research. We therefore included studies from less well-known or regional outlets along with Ph.D. dissertations and any other form of full-text write-up. Excluded from the study were master's theses, conference posters and presentations because collection would have been highly serendipitous and fragmentary and reporting partial.

Keywords in all searches included various combinations of *corpus/corpora*, *data-driven*, *DDL*, *Johns*, and *concordance/concordancer/concordancing* with contextualizers *language*, *learning*. Some searches produced tens of thousands of potential hits; in such cases, either the query was narrowed by introducing more search terms or the list was browsed until 100 consecutive entries yielded no new hits (cf. Plonsky, 2011). Once a seemingly relevant title was detected, the abstract was read and candidate papers were downloaded or obtained from other sources in order to constitute as large an initial pool as possible.

The trawls began with a search of various online databases: *Linguistics and Language Behavior Abstracts*, the *Modern Language Association International Bibliography*, *ProQuest Dissertations and Theses*, Education Resources Information Center, JSTOR, the *Directory of Open Access Journals* online, and Google Scholar (with the additional filter of pdf files only to reduce vast numbers of hits, often of slide presentations, Web sites, and notes). The bibliographies of all papers were then scoured for possible further leads. Next, Google was used to locate other papers that referred to those already found by searching for exact-word titles (within quotation marks). Though not common practice, *ancestry chasing* (Li, Shintani, & Ellis, 2012) or *forward citations* (Plonsky, 2011) did produce some publications that otherwise would have slipped through the net. Further searches were conducted on the Web sites of all sources that had more than one paper on the list at this stage or that seemed particularly promising for other reasons. This included journals, publishers, and Web sites for conference proceedings or known series of publications. Where a Web site did not have an adequate search engine built in, an advanced Google search was conducted specifying the site.

The trawls thus conducted brought up a total of 205 publications that we both agreed corresponded to our broad definition of empirical evaluations of DDL. The full list is provided in Appendix S1 in the Supporting Information online. Among those not pursued were seven papers in languages other than

English. Because none were suitable for meta-analysis, this justified the decision to limit our cull to English only. However, we have inevitably missed some relevant and interesting data reported in other languages; for example, 5 of the 14 DDL papers surveyed by Mizumoto and Chujo (2015) were in Japanese. Two papers were excluded for plagiarism, confirmed after consultation with the original author. Twelve studies were reported in more than one paper (typically a paper derived from a Ph.D. dissertation or expanding a short publication in conference proceedings). This is not uncommon (cf. Burston, 2015; Shintani, Li, & Ellis, 2013; Spada & Tomita, 2010), and these cases were counted as a single study, based on one main publication backed up where necessary with extra information from the secondary source(s). The pool of papers was further reduced for one of three main reasons:

1. Some did not focus on outcomes but on some other aspect such as learners' behavior or feedback on the treatment, often collected via observations, logs, questionnaires, or interviews.
2. Some used designs or instruments unsuitable for the current study, such as posttests for two different experimental DDL groups but no pretest or control group, description of error types, or relative frequencies of target item use.
3. Some did not provide the data necessary to calculate effect sizes, that is, the number of participants, the mean scores, and standard deviations, or the raw scores or statistics from which these could be derived. E-mails were sent to all authors of papers that fell into this category, which allowed us to add several studies that otherwise would have been excluded (see Author Note above for all authors who responded to such requests, whether fruitfully or not).

In this way, the general pool of 205 publications presenting empirical DDL studies was reduced to those that conformed to the criteria for inclusion in the meta-analysis (see Table 1).[2]

Coding and Data Extraction

Once the papers had been selected for inclusion, a coding manual was drawn up for descriptive and potential moderator variables. The various categories were divided into the same sections used by many other meta-analyses: publication, population, treatment, and design. To ensure harmonization, we applied our own criteria rather than adopting labels used in the original studies (cf. Jeon & Kaya, 2006), but still, like H. Lin (2014, p. 135), we often had to rely on "best guesses due to insufficient information given in the primary studies."

Table 1 Inclusion/exclusion criteria for studies in the meta-analysis

Set	Criterion	Included	Excluded
Original pool of empirical DDL studies	Domain	L2 use of corpus linguistic tools and techniques	L1 use
	Research type	Empirical evaluations	Descriptive, argumentative, position papers, etc.
	Publication type	Journal articles, book chapters, Ph.D. dissertations, conference proceedings, occasional/working papers	Spoken presentations, slides, notes, MA theses
Inclusion criteria for the meta-analysis	Language	English	Other
	Research questions	Outcomes of corpus use for learning or reference purposes	Behaviors, attitudes
	Research instruments	Etic (e.g., tests, written productions)	Emic (e.g., questionnaires, reflections)
	Research design	Quantifiable comparisons (pre/posttests and/or control/experimental groups)	Other
	Data provided	*N*, *M*, *SD*, or a way of extracting equivalent	Other
	Quality	All	Ø

The complete coding table can be found in Appendix S2 in the Supporting Information online, with a brief description of the criteria applied. It should be noted that the coding scheme refers to the study as a whole rather than just to the meta-analysis; in particular, the number of participants may vary where not all could be used for calculating effect sizes.

Once the coding manual had been agreed on, three papers were independently coded and the results compared. A further set of 10 papers was then randomly selected and again independently coded to check interrater reliability. Given the complexity of the coding sheet, this was assessed simply by counting the number of discrepancies between the two authors' coding rather than using calculations such as Cohen's kappa, following standard procedure in meta-analyses (e.g., Plonsky, 2011). In each of the four sections, agreement was rated at 90% to 96%; all disagreements were satisfactorily resolved, and the coding sheet refined where necessary. To complete the spreadsheet, the first author input an initial coding that was then thoroughly checked by the second author, and any further issues encountered were again resolved through discussion. Coding all papers by two separate, experienced researchers improves reliability but is relatively exceptional. In many meta-analyses, coding is performed by a single researcher or assistant.

Collecting, selecting, and coding the studies were time-consuming but relatively unproblematic procedures, with issues arising easily resolved. Far more complex in the end was extracting the data needed for the meta-analysis and the calculation of the effect sizes themselves. Some studies had extremely convoluted designs with a plethora of possible comparisons. Though this is often glossed over in survey papers or treated as purely mechanical, it does present the meta-analyst with numerous choices in extracting the essential data. The first author provided a preliminary summary in a similar manner to the coding. Because major issues had been resolved in advance, this was accepted by the second author following minor revisions such that no interrater reliability calculation was deemed necessary. The main choices are outlined below. It must be remembered that the overall objective was to seek maximum granularity by extracting as many data points as were available while preserving data that might be used in understanding moderators of the overall effects (Research Question 3). The full table can again be found in the Supporting Information online.

Some studies had two or more entirely independent experimental groups. These were each treated separately and not combined. Following Norris and Ortega (2000), where two control or comparison groups were involved, the one closest to the experimental design was chosen (e.g., traditional treatment

rather than no treatment, dictionaries rather than no reference tools). The result was consequently both more conservative and more ecologically valid because most other between-groups studies have involved comparison groups rather than true no-treatment controls. A small number of studies were longitudinal in nature with several intermediary tests. Although these are interesting and to be encouraged (e.g., Larsen-Freeman, 2006), for present purposes we included only pretest and posttest data. The data for delayed test results are included in the supplementary materials (https://osf.io/jkktw) but not discussed further here. This is common in meta-analyses and simplifies what is already a complicated picture, especially where the delay period varies substantially from one study to another. A small number of studies used the same students in both experimental and control treatments—either where the test included items that had been subjected to different treatments (e.g., some via DDL, some via traditional methods) or where students alternated between treatments in different weeks or switched treatments part way through. In such cases, the same participants functioned as both experimental and control subjects and were counted as such—an example of good practice according to Felix (2008).

Many studies provided results with quite a high degree of granularity derived from different instruments (e.g., multiple choice vs. cloze) or for different language items (e.g., collocations vs. lexicogrammar) or skills or functions (e.g., error correction vs. translation). These were recorded individually, each with its own effect size, so that subsequent analysis might identify relevant moderator variables (cf. Norris & Ortega, 2000). But for the main study, we followed standard practice by averaging the various elements to arrive at a single effect size for each unique sample—what Li et al. (2012, p. 9) called "shifting units of analysis."

In most cases, the data needed for effect sizes could be collected from the tables or text in the papers themselves. For the sake of consistency, effect sizes were calculated for all papers, even where they were originally given in the studies themselves. We repeatedly found it necessary to take a step back from the data with a common-sense eye on items that stood out as being statistically unlikely or impossible (cf. Larson-Hall & Plonsky, 2015). Where the summary data were incomplete for the main elements (*N*, *M*, *SD*), various options were available. Some studies were reported in more than one paper and the data could be completed from the sister publication. Where the primary data were given in an appendix, these could be used to calculate missing elements. In four cases, we were able to use data from *t* or *F* tests, but only to extract a single effect size in otherwise rich studies. For the remaining 13 studies, we e-mailed the authors. Five (38%) responded with the appropriate data; three responded

but were unable to help. For comparison, Plonsky, Egbert, and Laflair (2015) had a positive return from 14% of authors contacted.

Effect Sizes

Some studies were conducted between groups (control vs. experimental, henceforth C/E), others within groups (pretest vs. posttest, henceforth P/P); some featured both designs. The two were meta-analyzed here but kept separate for a number of reasons (Lipsey & Wilson, 2001). In particular, one would expect the mean to be higher in P/P designs, partly as use of the same participants reduces extraneous variance (Plonsky & Oswald, 2014) but also because the only difference between the two tests is an intervention: even poor teaching normally leads to some improvement. In the C/E design, however, very few studies use true controls with no intervention whatsoever. Where medicine compares treatment against a placebo, in SLA it is arguably more desirable to compare against whatever instruction learners would otherwise have. The comparison is thus between two interventions (i.e., experimental vs. traditional), each of which will have some effect, and in theory either could be more effective than the other. In other words, DDL versus no teaching should lead to a higher effect size than DDL versus traditional teaching.

Cohen's *d* (Cohen, 1988; see Figure 2) is the most common formula in meta-analyses in language teaching. Adopting familiar statistics increases transparency, comparability, and replicability (Plonsky, 2014), and we were anxious to avoid "technicism, or the overemphasis on manipulating data via novel meta-analytic techniques to the detriment of theoretical and conceptual depth" (Norris & Ortega, 2007, p. 810). Essentially, this measure compares the difference between the mean scores (P/P for within-groups, or C/E posttests for between-groups), divided by the pooled standard deviation of both groups (more highly dispersed scores reduce the reliability and power of the results) and measured in *SD* units. Because the component data are simply descriptive statistics, they can be extracted from *t* or *F* tests using equations given in Lipsey and Wilson (2001, p. 198–199; see Figures 3 and 4).

A further consideration involved effect-size weighting in the case of small samples (of which there were many in this analysis), which may have high sampling error (Lipsey & Wilson, 2001). Weighting is a controversial issue. A common method used in meta-analyses is the formula for "unbiased *d*" (Hedges, 1981), which Cumming (2012, p. 294) called a "very good approximation adjustment" (see Figure 5). All values for *d* given in these analyses used the unbiased *d* formula. To calculate this, Cumming used $df - 1$ where others have suggested $df - 9$. We retained Cumming's formula because $df - 1$ can

$$d = \frac{M_1 - M_2}{\sqrt{\frac{{SD_1}^2 + {SD_2}^2}{2}}}$$

Figure 2 Formula for calculating Cohen's *d* from *M* and *SD*.

$$d = \sqrt{\frac{F(N_1 + Nn_2)}{N_1 N_2}}$$

Figure 3 Formula for calculating *d* from *F*-test information.

$$d = t\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

Figure 4 Formula for calculating *d* from *t*-test information.

$$d_{unb} = \left(1 - \frac{3}{4df - 1}\right) \times d$$

Figure 5 Formula for calculating unbiased *d*.

accommodate very small sample sizes, while $df - 9$ does not work where $N = 10$ and actually increases effect size where $N < 10$.

Once effect sizes had been calculated, outliers could easily be observed through a funnel plot. Rather than excluding these, we adopted the common practice of winsorizing effect sizes at the 5th and 95th percentiles to provide more robust results (cf. Li et al., 2012; Lipsey & Wilson, 2001). This turned out to be $d = 3.0$, within the range recommended by Lipsey and Wilson.

The interpretation of effect sizes is also an issue. Cohen's (1988) original rule of thumb suggested that a *d* value of 0.2 could be considered a small effect, 0.5 medium, and 0.8 large. For the field of education, Hattie (2009) proposed levels of 0.2, 0.4, and 0.6, noting that anything below 0.4 should be abandoned as effectively useless (p. 18). Plonsky and Oswald (2014) argued convincingly for field-specific benchmarks derived empirically from typical findings in that field. Taking the 25th, 50th, and 75th percentiles as indicators for small, medium, and large effects, their study of 91 meta-analyses in SLA suggested 0.6, 1.0, and 1.4,

respectively, for within-groups designs and 0.4, 0.6, and 0.9 for between-groups comparisons. We endorse this approach and adhered to these benchmarks in our study. Also typically included in reports of grouped effect sizes are the confidence intervals (CI), traditionally set at 95%—an arbitrary figure that parallels choices for *p* values in the null hypothesis significance testing (NHST) statistical model (Oswald & Plonsky, 2010). A CI is a measure of heterogeneity within a collection of effect sizes and thus a prediction of how likely it is that the same mean would be found for a different sample of studies from the same population. A large CI shows greater heterogeneity, suggesting that the same mean might not be found in a different sampling from the same population. If the CI includes zero, the mean is considered not to indicate a reliable effect. The procedure for calculating CIs is provided in Appendix S3 in the Supporting Information online.

Overall Results

Application of the various inclusion/exclusion criteria reduced the original pool of 205 publications to 64, or just under one third, for a total of over 3,000 participants. This might seem a low inclusion rate, but N. Ellis (2006) noted that other meta-analyses have reported similar figures, and Yun (2011) included only 10 from an original trawl of 200 publications. Reassuringly, the rate of includable papers seems to be increasing over time, a sign of growing rigor and better reporting, in line with Larson-Hall and Plonsky (2015). Our number of studies also compared favorably with Oswald and Plonsky's (2010) survey of 27 meta-analyses in SLA, with a median of only 16 studies. In total, 88 unique samples were ultimately harvested from our 64 studies, some yielding only a single effect size, others two or more. For any part of the analysis, only a single effect size is reported for any given sample.

Figure 6 shows the overall evolution of DDL research from the first empirical study in 1989, in which *DDL* refers to the entire pool of publications collected for the present study (205), many of which could not be included, *MA* to the number of empirical studies meta-analyzed here (64), some of which contained more than one unique sample, and *k* to the number of unique samples (88). Following a typical pattern (Shintani et al., 2013), empirical studies lagged behind early theoretical or descriptive studies but have since taken off and were still increasing as of June 2014.

Having selected the studies and extracted and sorted the relevant data, we put everything together to provide overall effect sizes for the two main study types and then broke the whole back down again into subgroups according to our coded variables (the potential moderators).

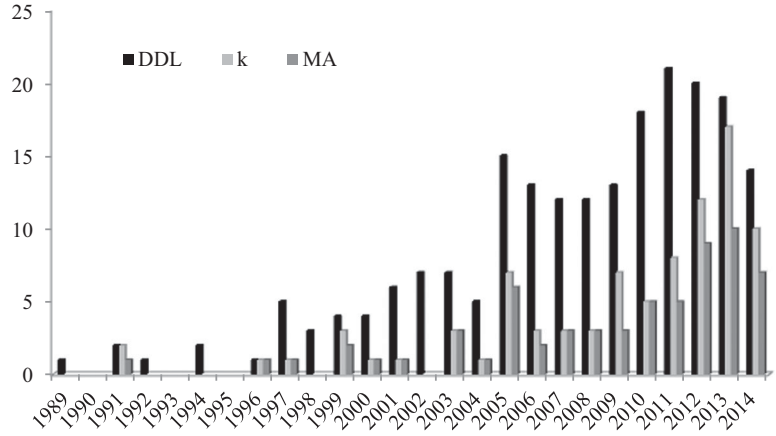


Figure 6 Evolution of empirical data-driven learning studies (DDL) as a whole in relation to meta-analyzable studies (MA) and the number of unique samples (k).

Overall Effect Sizes

Of the 64 studies meta-analyzed, some had more than one experimental group, giving a total of 88 unique samples that we separated into C/E and P/P designs. As some studies compared the experimental group posttest to both a pretest and a control group, the results could be counted in both categories, so that in the end we were able to work with data from 50 C/E and 71 P/P samples.

Main effect sizes are given in Table 2, where each row represents a unique sample, along with the number of participants in the control (where applicable) and experimental groups. Effect sizes are given for the main comparisons: P/P (i.e., where the same participants were tested before and after a DDL intervention) and C/E (i.e., where control and experimental groups were given the same posttest). More complete data can be found in Appendices S4 and S5 in the Supporting Information online, which take the form of a standard spreadsheet. Though many statistical packages can help with meta-analyses, especially in producing visual representations, all the essential information can be stored and calculations performed transparently via a spreadsheet.

Within Groups (P/P) and Between Groups (C/E)

For P/P designs, a total of 71 unique samples showed d ranging from –0.09 (the only negative score) to 3.0 following winsorizing in 10 cases. The mean gain effect was $d = 1.50$ ($SD = 0.91$), higher than Mizumoto and Chujo's (2015) $d = 0.97$ in Japan. This placed the result in the top quartile of Plonsky and Oswald's (2014) pooling of meta-analyses in SLA outlined earlier and

Table 2 Main effect sizes of all studies included in the meta-analysis

		N		Effect size (d_{unb})	
ID	Reference	CG	EG	EG pre/EG post	CG post/EG post
1	Abu Alshaar & Abuseileek (2013)	16	16	[a]3.00	0.87
2	Ashouri et al. (2014)	30	30	[a]3.00	[a]3.00
3a	Bale (2013a, 2013b)		8	2.40	
3b	Bale (2013a, 2013b)		9	1.35	
3c	Bale (2013a, 2013b)		4	2.03	
3d	Bale (2013a, 2013b)		6	1.67	
4	Boulton (2007)	51	53	0.19	–0.06
5	Boulton (2008a)		113	0.64	
6	Boulton (2009)	32	34	0.87	0.46
7	Boulton (2010a, 2008b)	62	62	0.70	0.34
8	Boulton (2011)	25	34		0.37
9	Braun (2007)	12	13		[a]3.00
10a	Buyse & Verlinde (2013)	17	17		0.58
10b	Buyse & Verlinde (2013)	17	18		–0.14
11	Çelik (2011)	34	32	2.49	0.26
12	Chan & Liou (2005)		32	2.41	
13	Chang (2010, 2012)		7	1.40	
14a	Chang & Sun (2009)		13	[a]3.00	
14b	Chang & Sun (2009)		13	[a]3.00	
15a	Chatpunnarangsee (2013)		9	1.11	
15b	Chatpunnarangsee (2013)		10	1.41	
15c	Chatpunnarangsee (2013)		5	2.70	
16a	Chen (2011)		22	1.38	

(*Continued*)

Table 2 Continued

		N		Effect size (d_{unb})	
ID	Reference	CG	EG	EG pre/EG post	CG post/EG post
16b	Chen (2011)		29	2.64	
17	Chujo et al. (2013)		22	0.65	
18a	Chujo & Oghigian (2012)	23	25	1.51	1.98
18b	Chujo & Oghigian (2012)	23	14	1.12	0.73
19	Cobb (1997a, 1997b)	11	11	2.42	[a]3.00
20a	Cobb (1999a, 1997b, 1999b)	17	18	0.70	0.44
20b	Cobb (1999a, 1997b, 1999b)	9	12	0.85	0.49
21a	Cotos (2010, 2014)		16	1.58	
21b	Cotos (2010, 2014)		15	1.94	
22	Curado Fuentes (2007)	20	20		[a]3.00
23	Daskalovska (2014)	25	21	1.85	1.51
24	Frankenberg-Garcia (2012)	12	12		2.38
25	Frankenberg-Garcia (2014)	12	13		0.31
26	Gan et al. (1996)	48	48	1.60	1.27
27	Gao (2011)		21	0.67	
28	Gaskell & Cobb (2004)	13	19	0.04	0.50
29	Gordani (2013)	35	35	[a]3.00	0.87
30	Hadi (2013)	25	25	[a]3.00	2.00
31	Hadi & Alibakhshi (2012)	32	32		1.63
32	Horst (2005)		14	1.05	
33	Horst & Cobb (2001)		30	0.16	
34	Horst et al. (2005)	14	19	1.27	0.65
35	Huang (2012, 2014)	20	20	0.45	0.15
36	Johns et al. (2008)	11	11	0.33	[a]3.00

(*Continued*)

Table 2 Continued

		N		Effect size (d_{unb})	
ID	Reference	CG	EG	EG pre/EG post	CG post/EG post
37	Kaur & Hegelheimer (2005)	9	9		0.36
38	Kayaoğlu (2013)		23	2.20	
39	Koosha & Jafarpour (2006), Jafarpour & Koosha (2005)	100	100		0.79
40	Lee & Liou (2003)	46	46		1.09
41	Lewandowska (2013)	15	14	1.15	0.17
42	Lin & Liou (2009), Lin (2008)		25	1.57	
43a	Liou et al. (2006)		38	0.63	
43b	Liou et al. (2006)		32	2.41	
44	Lu (2008)	30	30		0.57
45a	Miangah (2011)		16	0.51	
45b	Miangah (2011)		17	1.23	
45c	Miangah (2011)		17	1.50	
46	Moreno Jaén (2010)		21	1.06	
47	Nam (2010a, 2010b)	11	10	0.71	–0.39
48	Oghigian & Chujo (2010), Chujo et al. (2009)	25	22	1.20	2.41
49a	Oghigian & Chujo (2012a)		5	0.01	
49b	Oghigian & Chujo (2012a)		5	0.69	
50a	Oghigian & Chujo (2012b)		6	0.47	
50b	Oghigian & Chujo (2012b)		9	0.16	
51	Pirmoradian & Tabatabaei (2012)	15	15	[a]3.00	[a]3.00
52	Poole (2012)	9	9	0.82	0.43
53	Rapti (2010)	14	14	–0.09	0.47

(*Continued*)

Table 2 Continued

		N		Effect size (d_{unb})	
ID	Reference	CG	EG	EG pre/EG post	CG post/EG post
54	Smart (2012, 2014)	18	16	0.56	0.32
55	Someya (2000)	20	20	0.85	0.66
56	Sripicharn (2003)	18	22		0.08
57a	Stevens (1991)	22	22		1.00
57b	Stevens (1991)	22	22		0.68
58	Sun & Wang (2003)	40	41		0.57
59	Supatranont (2005)	26	26		1.58
60a	Tian (2014)		20	2.33	
60b	Tian (2014)		20	1.48	
60c	Tian (2014)		20	1.48	
61a	Tian (2005b, 2005a)	25	27	[a]3.00	0.26
61b	Tian (2005a, 2005b)	23	23	[a]3.00	0.43
62a	Tongpoon (2009)	8	14	1.47	0.02
62b	Tongpoon (2009)	16	19	1.72	–0.34
62c	Tongpoon (2009)	8	28	1.81	0.73
62d	Tongpoon (2009)	16	20	1.52	–0.12
63a	Yang et al. (2013)		35	2.05	
63b	Yang et al. (2013)		28	[a]3.00	
64	Yoon & Jo (2014)		4	1.08	
	k			71	50
	M	23.64	22.41	1.50	0.95
	SD	16.14	17.04	0.91	0.99
	Upper 95% CI			1.71	1.22
	Lower 95% CI			1.28	0.67

Note. CG = control group; EG = experimental group. [a]Effect sizes winsorized to 3.0 standard deviations. References to all meta-analyzed studies are included in Appendix S1 in the Supporting Information online; all citations in this table (e.g., Cobb 1999a, 1999b) correspond to the reference list in Appendix S1, not the reference list at the end of this article.

above their benchmark of 1.4 for a large effect. The 95% CI was between 1.28 and 1.71, well above zero and in a relatively narrow range that showed the comparatively low variance in the studies.

For C/E designs (50 unique samples), the mean difference effect ranged from –0.39 (there were five negative scores) to 3.0 following winsorizing for three studies. The mean difference effect was $d = 0.95$ ($SD = .99$), where, as mentioned, Plonsky and Oswald found 0.6 for a medium effect and 0.9 for a large effect.[3] The CIs were slightly wider at 0.67 and 1.22, indicating a greater

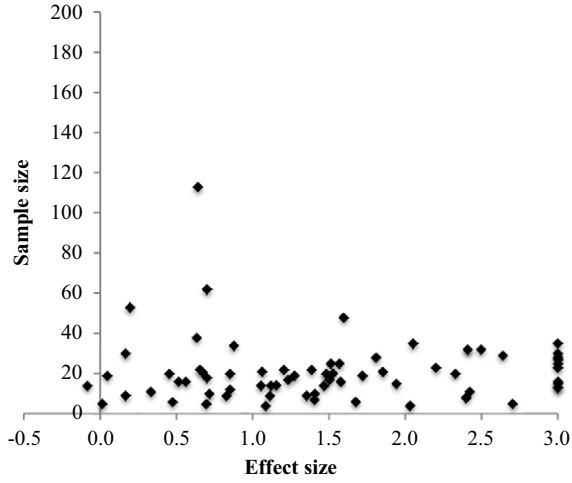


Figure 7 Within-group (P/P) design ($M = 1.51$).

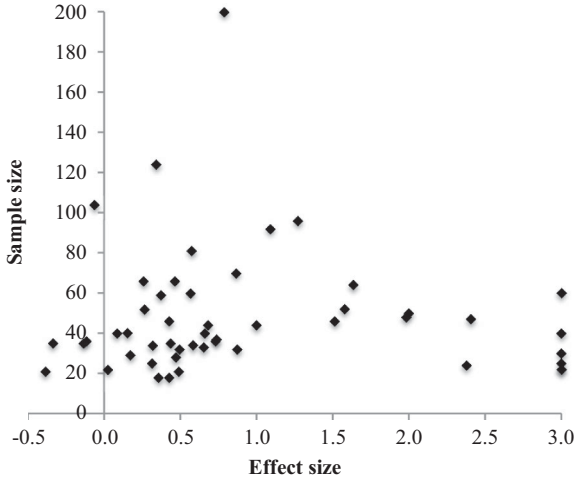


Figure 8 Between-group (C/E) design ($M = 0.95$).

range in the results, but still well above zero, suggesting confidence that the true mean for these studies lay in this range.

The two sets of results can be visualized for dispersion in the funnel plots in Figures 7 and 8. Apart from the obviously smaller sample sizes in the P/P design (because only one group was involved), the horizontal spread also

seemed rather different. P/P scores were spread fairly evenly across the scale, while C/E scores clustered around 0.5 and tailed off above 1.0 (not counting the 10 winsorized scores). This suggested that simply using unbiased *d* might not be sufficient to allow the two different designs to be satisfactorily combined.

In addition to being a convenient way to visualize the results (cf. Cumming, 2012), funnel plots also provide an indication of potential publication bias through any asymmetries (fewer studies to the left of the mean suggesting that small samples with low effects go unpublished) or systematic small-sample effects (because the means are plotted against *k* size). Figure 7 shows that the P/P design mostly relied on quite small sample sizes, but these are relatively evenly distributed either side of the mean. Larger samples were used in C/E designs (involving two groups) and did indeed tend to reveal a funnel shape in Figure 8. Both are what we would expect and did not provide much evidence of publication bias (Norris & Ortega, 2000; Oswald & Plonsky, 2010). However, it is clear that the means were pulled upward by the studies that were winsorized at 3.0; simply eliminating these as outliers would reduce the mean P/P effect size to 1.25 ($SD = 0.72$; CI = [1.07, 1.43]), and the mean C/E effect to 0.67 ($SD = 0.67$; CI = [0.47, 0.86]), which are still substantial results.

Results for Moderator Variables

Having pooled the results of these studies for the general picture, we then broke them down again by moderator variables (MVs) to investigate variation within the general picture (Lipsey & Wilson, 2001), with all original analyses presented in Appendices S6 and S7 in the Supporting Information online. This assumed a random-effects model, which several researchers have claimed should be the default setting for meta-analysis (Borenstein et al., 2009, chap. 21; Cumming, 2012, p. 213; Oswald & Plonsky, 2010). On this basis, Q tests were not required because heterogeneity is implicit in this assumption. However, the choice of model is largely theoretical and unlikely to affect the values substantially (Oswald & Plonsky, 2010). In total, 84 variables in 25 groups were examined in the four coding categories (publication, population, treatment, and design).

Publication Variables

There might be a tendency for effect sizes to increase over time as researchers focus on the most promising questions (Lee, Jang, & Plonsky, 2015) or conversely to decrease in a *Proteus effect* (Plonsky & Oswald, 2014) whereby they regress toward the mean as research becomes more nuanced concerning specific features. The small number of early studies in particular made this difficult to assess here, whether looking at individual years or grouping over different

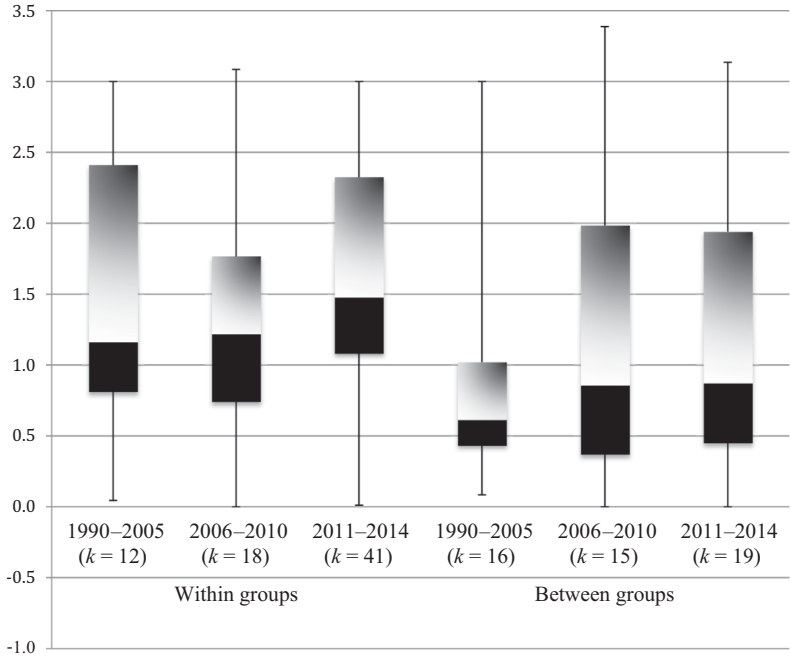


Figure 9 Evolution in effect sizes over time.

time periods. Figure 9 shows the quartiles over three time periods for P/P and C/E designs. Both seemed to show an increase in effect size over time. Exact values for means and standard deviations are given in Table 3 along with CIs.

The question of study quality is difficult to assess objectively, but the publication source might provide some indication. The second section of Table 3 compares journal articles against Ph.D. dissertations and other publication types (mainly book chapters). For P/P designs, it seems that journal articles gave the highest effect sizes, followed by Ph.D. dissertations; this order changed for C/E comparisons, though few studies were involved. As journals were the largest category, we separated out those ranked in the top 50 of the Scientific Journal Rankings and Journal Citation Reports language and linguistics categories: *Computer Assisted Language Learning*, *Language Learning*, *Language Learning & Technology*, *ReCALL*, and *System*.[4] This gave a total of 25 studies in ranked journals, compared to 39 in unranked ones. The results showed higher effect sizes for ranked than unranked journals for both P/P and C/E designs, which may suggest submission bias (if researchers keep their best results for prestigious journals) or acceptance bias (if those journals only accept

Table 3 Breakdown of studies by publication date and type, journal rank, and page length

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Feature	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Publication date										
1991–2005	12	1.45	1.04	0.86	2.03	16	0.82	0.70	0.47	1.16
2006–2010	18	1.26	0.89	0.85	1.68	15	0.93	1.26	0.29	1.56
2011–2014	41	1.61	0.88	1.34	1.88	19	1.08	0.99	0.63	1.52
Publication type										
Journals	50	1.60	0.97	1.33	1.87	36	1.05	1.00	0.72	1.38
PhDs	14	1.49	0.70	1.12	1.85	8	0.35	0.60	–0.06	0.77
Other	7	0.79	0.50	0.42	1.16	6	1.11	1.15	0.19	2.03
Journal prestige										
Ranked	21	1.67	0.96	1.26	2.08	13	1.13	1.12	0.53	1.74
Unranked	29	1.54	0.98	1.19	1.90	23	1.01	0.96	0.61	1.40
Length										
1–10 pages	6	1.81	1.00	1.01	2.61	6	1.21	1.14	0.31	2.12
11–20 pages	29	1.77	1.08	1.38	2.16	21	1.00	0.85	0.64	1.36
20+ pages	15	1.18	0.58	0.88	1.47	9	1.06	1.34	0.19	1.94

papers with substantial findings). Alternatively, studies that make it into the top journals could be carefully conducted to eliminate extraneous variables. A final indirect measure of quality may lie just in the number of pages. Researchers convinced that their study is important may go to more trouble writing it up while those with doubts commit themselves only to a short paper. Or long papers might indicate tergiversation to "contextualize" small or ambiguous findings while short papers indicate confidence in strong or clear findings. Our shorter papers produced higher effect sizes than longer ones, again with minor differences for P/P and C/E designs.

Design Variables

An important component of study quality in DDL work appeared to be population. As shown in Table 4, more participants generally meant larger effect sizes in P/P designs and the opposite in C/E designs. The same patterns held for overall sample sizes. Only one study used an apparently true control group, with most using comparison groups that followed a different treatment (such as explicit teaching or use of dictionaries) or on occasion providing the same

Table 4 Design variables found in the meta-analysis studies

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Design variable	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
EG sample size										
< 20	13	1.17	1.02	0.61	1.72	9	1.81	1.45	0.86	2.75
20–49	34	1.59	0.89	1.29	1.89	28	0.95	0.81	0.66	1.25
50+	24	1.54	0.86	1.19	1.88	13	0.34	0.40	0.12	0.55
Control										
Comparison	–	–	–	–	–	32	1.06	1.06	0.69	1.43
Identical	–	–	–	–	–	8	0.95	0.90	0.33	1.57
Constitution										
Intact groups	–	–	–	–	–	19	0.81	0.91	0.40	1.22
Random assignment	–	–	–	–	–	14	0.84	0.94	0.35	1.34
Instruments										
Selected response	13	1.44	0.89	0.95	1.92	5	1.14	1.21	0.09	2.20
Constrained response	16	1.89	0.93	1.43	2.35	11	0.75	0.57	0.41	1.09
Free response	15	0.86	0.65	0.53	1.19	12	1.00	1.26	0.29	1.71
Mixed	24	1.60	0.84	–0.63	3.84	20	0.89	0.93	–2.00	3.79
Statistical tests										
0	10	0.60	0.53	0.28	0.93	5	0.33	0.31	0.05	0.60
1	35	1.60	0.91	1.30	1.90	24	1.19	1.03	0.78	1.60
2+	26	1.70	0.84	1.37	2.02	21	0.81	0.98	0.39	1.23
Other instruments										
0	15	1.62	1.03	1.10	2.14	17	0.91	0.90	0.48	1.34
1	31	1.27	0.92	0.94	1.59	21	1.00	1.02	0.56	1.44
2+	25	1.71	0.78	1.40	2.01	12	0.91	1.12	0.28	1.54

students with different treatments on different items so that they formed their own comparison group. The former seemed to provide slightly larger effect sizes, though both were near or over the benchmark for large, and there were only eight studies using the second design. Unfortunately, 42 of our 88 unique samples gave no indication of how the groups were constituted. For those that did, intact groups outnumbered random assignments, but there was little difference in the resulting effect sizes (medium to large).

In terms of data collection instruments, Table 4 indicates that selected response tests (multiple choice) showed quite large effects but featured in the fewest number of studies overall. Constrained constructed responses (focusing

on specific items with limited response options) gave the largest effect sizes in P/P but the lowest in C/E designs. The freest types, such as writing or translation, showed a large effect in C/E and a medium effect in P/P studies. The most common category of more than one type of instrument yielded large (P/P) and medium (C/E) effects. But it should not be inferred that DDL is more or less appropriate in line with this criterion. The instruments used did not consistently or accurately reflect the pedagogical or linguistic objectives of each study and were ultimately a feature of the study design.

Also of possible interest is the type of analysis conducted. Table 4 shows that this was usually in the form of NHST, especially with *t* tests (in 47 of the 88 unique samples overall), followed by analyses of variance or covariance in 35 cases. Effect sizes were provided in some form for nine samples, though as mentioned earlier these were recalculated here. In 13 cases descriptive statistics were provided with no statistical analysis as such. It might be tempting to infer that this group involved less sophisticated research overall—not just in the analysis but also in the design. It was indeed this group that produced the smallest effect sizes overall. In P/P studies, use of more tests was associated with larger effect sizes, but in C/E designs it seemed to complicate things—perhaps because several derived from Ph.D. dissertations, which tended to produce lower mean difference effect sizes. Finally, we looked at any tools used in these studies other than those explicitly used to derive the effect sizes, again on the assumption that the total number of instruments might provide an indication of sophistication if not quality per se, but effect sizes showed no discernible relationship to number or type of instruments used (see Table 4).

Population Variables

It is often objected that DDL may not be appropriate for all types of learner, though few studies have addressed this question directly. Individual differences are difficult to assess in meta-analyses, though we did attempt to glean some studies including information about age and sex. Average age was given explicitly in only a small minority of cases and was not pursued because it overlapped with the particular year in a program of study, where more information was provided. Sex was not analyzed either, as the few studies that provided the information did not separate their results by sex.

There may however be larger cultural differences that can be analyzed here. It is clear from Table 5 that DDL has been more widely researched in Asia than in other parts of the world with large (P/P) or medium (C/E) effect sizes, though the largest effects have been achieved in the Middle East. It should be noted that a fifth of the studies in Asia and nearly half of those in the

Table 5 Population samples used in the studies included in the meta-analysis

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Sample type	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Region										
Asia	36	1.55	0.89	1.26	1.85	18	0.84	0.91	0.42	1.26
Middle East	13	2.07	0.98	1.54	2.60	13	1.39	1.03	0.83	1.95
Europe	12	1.15	0.75	0.73	1.58	13	0.95	1.13	0.34	1.57
North America	10	0.95	0.61	0.58	1.33	6	0.31	0.36	0.02	0.60
Context										
Foreign language	57	1.56	0.95	1.32	1.81	44	1.03	1.02	0.73	1.33
Second language	10	0.95	0.61	0.58	1.33	6	0.31	0.36	0.02	0.60
L1										
Chinese	20	1.81	1.01	1.37	2.25	8	0.82	0.92	0.18	1.46
Romance	5	0.69	0.32	0.41	.98	7	0.97	1.20	0.08	1.86
Japanese	9	0.74	0.49	0.42	1.06	4	1.44	0.88	0.58	2.31
Persian (Farsi)	7	2.18	1.07	1.39	2.97	6	1.88	0.98	1.10	2.67
Thai	7	1.68	0.51	1.30	2.05	6	0.33	0.71	–0.24	0.89
Arabic	4	1.74	1.14	0.62	2.86	6	1.08	0.97	0.31	1.85
Other	8	1.38	0.84	0.80	1.96	9	0.75	1.04	0.07	1.43
Mixed	11	1.35	0.67	0.95	1.74	4	0.44	0.15	0.29	0.59

Middle East produced very large effects, winsorized down to 3.0, suggesting that design and analysis may have a role to play. Taken at face value, these results may seem counterintuitive because many Asian and Middle Eastern cultures favor education that is teacher fronted, deductive, and strong on rote learning—the antithesis of DDL. This has been observed by several researchers in Taiwan in particular (e.g., Yeh, Liou, & Li, 2007), although Smith (2011, p. 294) noted that his undergraduate students there "expect to be taught in a way that is markedly different from their high school experience" and that "the last thing one would expect them to want is more gap-fills and error correction exercises." Conversely, it was in Europe and North America that effect sizes were rather lower (though still reasonably robust), two regions where inductive, problem-solving approaches would seem more in line with prevailing cultures. One obvious possibility is that DDL was not different enough from traditional teaching in these parts of the world, and this was somewhat borne out by C/E designs producing the lowest effect sizes. Extending the analysis to individual first language (L1) backgrounds, these correlated to a large extent with the

above findings, with speakers of Asian and Middle Eastern languages showing larger effect sizes on P/P designs (with the exception of Japanese), though the picture was again more mixed for C/E studies and the number of unique samples for each language relatively small.

Sophistication Variables

Target language was not a realistic variable because all but two studies had English as the L2 (one Spanish, one mixed). Proficiency in the L2 was particularly difficult to assess and involved a lot of informed guesswork from the descriptions available with differing terminology, sometimes derived from standardized tests, sometimes from in-house tests, sometimes from the teachers'/researchers' individual perceptions, which are likely to be culture bound. In their study of "advanced" learners in major CALL journals, for example, Burston and Arispe (2016) found that 50% were at B1 and 34% at B2 levels on the Common European Framework of Reference for Languages. In our case, the intermediate category in particular seemed to include individuals who elsewhere would be ranked as lower- or upper-intermediate or even false beginner or advanced. Nonetheless, Table 6 shows that we generally had moderately or even very large effect sizes in most cases, the exception being lower-intermediate learners in P/P studies. One might think that students specializing in the L2 (including in linguistics, translation, or teacher training) would have more sophisticated linguistic reflexes and thus do relatively better with the techniques involved in DDL. This certainly seemed to be the case for P/P designs, although the effect sizes for students in nonlinguistic disciplines were also medium to large. C/E studies produced the largest effect in the social sciences (but based on few unique samples) and lower effect sizes in the hard sciences (here engineering, maths, science, medicine, and architecture). Among mixed groups, effect size was very low and the CI included zero. Another possible indication of sophistication might be educational level. Unfortunately, there is little research with high school learners: six C/E samples (producing a large effect size) and only four P/P samples (with a respectable mean effect size, though the CI descended below zero, indicating heterogeneity). At the university level, effect sizes were again mixed for different years (where this information was available) and effect sizes for postgraduate students seemed particularly promising, though again from few samples.

Treatment Variables

Some studies involved highly experimental treatments designed to limit variables and isolate the DDL factor, and others were more ecological and integrated

Table 6 Breakdown of studies by proficiency level of subjects, their degree specialization and type of institution where they studied

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Sample type	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Proficiency										
Advanced	14	1.58	0.88	1.12	2.04	6	1.09	1.09	0.21	1.96
Intermediate+	12	1.34	0.65	0.97	1.71	11	0.71	0.88	0.19	1.23
Intermediate	23	1.72	1.06	1.29	2.15	12	1.27	1.11	0.64	1.89
Intermediate–	14	1.40	0.98	0.89	1.92	12	0.32	0.39	0.10	0.55
Lower	8	1.10	0.65	0.65	1.55	7	1.72	1.15	0.87	2.57
Speciality										
Languages	18	1.84	0.77	1.49	2.20	12	1.23	1.05	0.64	1.83
Social sciences	6	1.11	0.71	0.54	1.68	7	1.44	1.29	0.49	2.40
Other sciences	20	1.24	0.92	0.83	1.64	13	0.86	0.72	0.47	1.25
Mixed	17	1.61	0.81	1.23	2.00	10	0.19	0.39	–0.05	0.43
Institution										
School	4	1.56	1.67	–0.08	3.20	6	1.41	1.26	0.40	2.42
Uni 1	24	1.41	0.81	1.08	1.73	13	0.96	0.86	0.49	1.43
Uni 2–3	9	1.27	0.47	0.97	1.58	13	0.45	0.76	0.04	0.86
PG	6	2.09	0.73	1.50	2.67	3	1.72	1.13	0.44	2.99

DDL into regular courses (see Table 7). To examine this, we separated studies conducted under laboratory-like conditions from those in regular classrooms; both contexts showed large effect sizes, except for C/E designs in lab conditions. Unfortunately, only two studies were conducted in other contexts, so it was not possible to see how DDL resources might be used out of class or after a course has finished. We also looked at the length of the intervention, though again this involved a degree of informed guesswork because some measured duration in hours and minutes, while others measured in class sessions, weeks, months, terms, semesters, or years. Those of 2 hours or less, usually in one class period, were considered short-term experimental studies. Somewhat longer ones covering three to eight classes were considered medium duration (M = 10 hours 20 minutes). The remainder, consisting of at least 10 classes, were considered long term, typically approximating a semester's work for 25 to 30 hours, though some introduced DDL for just a few minutes each class, while others were up to 1 year (three studies) or even 2 years (one study). Again, differences depended on design. P/P studies yielded the largest effect

Table 7 Studies included in the meta-analysis by the treatment used

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Treatment	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Ecology										
Class	52	1.55	0.88	1.31	1.79	32	1.06	1.11	0.68	1.45
Lab	13	1.65	0.98	1.12	2.19	10	0.86	0.83	0.34	1.37
Duration										
Short	18	1.54	1.02	1.07	2.01	17	0.89	0.90	0.47	1.32
Medium	18	1.89	0.56	1.63	2.15	10	0.85	1.22	0.10	1.61
Long	34	1.31	0.93	1.00	1.62	22	1.05	1.00	0.64	1.47
Interaction										
Concordancer	36	1.80	0.79	1.54	2.05	22	0.93	0.99	0.52	1.34
CALL program	12	1.41	0.88	0.92	1.91	6	1.33	1.29	0.30	2.37
Paper	13	1.06	0.96	0.53	1.58	15	0.52	0.58	0.22	0.81
Mixed	9	0.88	0.78	0.37	1.39	6	1.35	1.03	0.52	2.18
Corpus size										
< 1 m words	7	1.36	0.67	0.86	1.86	6	1.92	1.25	0.91	2.92
1–99 m words	15	1.53	0.84	1.11	1.96	9	0.50	0.68	0.06	0.94
> 100 m words	17	1.66	1.00	1.18	2.14	11	1.09	0.89	0.56	1.61
Corpus type										
Public (mono)	34	1.42	0.97	1.09	1.74	22	0.62	0.78	0.29	0.95
Local (mono)	20	1.67	0.91	1.28	2.07	18	1.16	0.99	0.70	1.62
Parallel	9	1.35	0.69	0.90	1.81	5	1.11	1.05	0.19	2.03

sizes in medium- or short-term contexts, and C/E studies gave the advantage to long-term work—but always with medium to large effects.

DDL can be implemented in many different ways, notably in terms of the type of interaction learners have with corpus data. It seems that those that actually used technology, that is, where the learners used a concordancer themselves or via some kind of CALL program, tended to show large effect sizes, but learners using concordance printouts or other paper-based materials did less well. Some studies used a combination of the two, typically leading from work on paper to hands-on concordancing. The results in these cases were mixed, with such designs producing large effects sizes in C/E studies and lower effect sizes in P/P studies. Where learners had access to corpus software, a few used small monolingual corpora (7,000 to 500,000 words),

usually compiled with precise goals in mind from local sources (news Web sites, research articles, coursebooks, novels, learners' own productions). These nonetheless provided respectable to strong effect sizes, particularly in C/E designs. Large monolingual corpora of 100 to 500 million words or more also produced large effects. Intermediate size corpora (1 to 29 million words) were more mixed—a large P/P effect size and a small C/E one. However, not all studies provided information about corpus size. In the case of parallel corpora, the L2 corpus tokens only were taken into consideration. Reworking these calculations in terms of the corpus type rather than size, the most frequently used publicly available monolingual corpora were the Brown, British National Corpus, and Corpus of Contemporary American English. Using calculations based on corpus type, the C/E designs produced more substantial effect sizes than P/P. The use of parallel corpora gave large effect sizes, most commonly where the L1 was Chinese or Japanese. Unfortunately, multimodal corpora cannot be considered here as they featured in only two studies, one of which was winsorized.

As for software, of the 51 studies that involved hands-on concordancing, 15 used more than one software program. Among the remainder, only the BYU interface and LexTutor produced effect sizes based on five or more unique samples, three of which in each case derived from the same study. Even grouping the others by type (e.g., AntConc, WordSmith Tools, and the Longman MiniConcordancer as instances of downloadable programs for use with any monolingual corpus) only covered four studies. This made any interpretation delicate and difficult to explore in more detail from the data available at the time of our study.

Objectives Variables

Surprisingly, perhaps, most DDL research to date has targeted language for general purposes, resulting in reassuringly large effect sizes (see Table 8). Language for specific purposes also came out fairly well, though the results were mixed in the case of language for academic purposes: a very small effect size for C/E designs with a lower confidence limit approaching zero. Most research in this area has attempted to evaluate corpus use for learning rather than reference purposes, yielding large effect sizes. Where corpora were used as a reference resource (typically during the data collection phase of the study), effects were still large in P/P but medium in C/E designs.

Two final aspects for consideration were the language skills and language forms targeted. These required more painstaking treatment, returning

Table 8 Objectives of the studies included in the meta-analysis

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Target	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Objective										
LGP[a]	42	1.54	0.96	1.25	1.83	31	1.16	1.08	0.78	1.54
LSP[b]	9	2.15	0.78	1.64	2.66	8	1.02	0.89	0.40	1.64
LAP[c]	19	1.14	0.67	0.84	1.44	10	0.29	0.34	0.08	0.50
Use										
Learning	48	1.56	0.85	1.31	1.80	39	0.98	1.04	0.66	1.31
Reference	21	1.36	1.03	0.92	1.80	11	0.82	0.83	0.33	1.31

Note. [a]Language for general purposes. [b]Language for specific purposes. [c]Language for academic purposes.

to the individual subeffect sizes calculated for each study and the methodologies and instruments used to obtain them, ensuring that unique samples were not counted more than once for any MV. Language skills were counted when they constituted a genuine teaching focus in a study, as opposed to a spinoff of the data collection instrument. To the traditional four skills (listening, speaking, reading, and writing), we added translation as a third code. Further distinctions could be made for subskills, especially using corpora for error correction or revision, but this overlapped with the reference function discussed above. The most revealing finding for skills (see Table 9) was the paucity of studies for speaking and listening. This did not mean that the corpora used did not include (transcripts of) spoken language but that the studies themselves did not directly teach or evaluate speaking or listening. Writing clearly dominated, though with mixed results—a medium effect size for P/P designs, negligible in C/E designs. Translation might be more promising, though it only featured in seven P/P studies, where it had a very large effect and has yet to be explored in C/E designs (at least in terms of meta-analyzable data). The other skills similarly remain largely underresearched, making difficult any strong pronouncement between productive and receptive skills.

Looking at language forms was even more complicated and involved revisiting each study to see if it had a specific linguistic focus, even if this was not necessarily the stated aim. Vocabulary covered those studies that mainly concentrated on learning quantities of new items (including phrasal verbs and

Table 9 Language skills and forms examined in the studies included in the meta-analysis

	Pretest/posttest studies					Control/experimental studies				
				95% CI					95% CI	
Language target	*k*	$M(d_{unb})$	*SD*	lower	upper	*k*	$M(d_{unb})$	*SD*	lower	upper
Language skill										
Listening	4	0.42	0.11	0.31	0.53	4	0.59	0.87	–0.26	1.44
Speaking	0	–	–	–	–	0	–	–	–	–
Reading	0	–	–	–	–	3	1.80	1.47	0.14	3.47
Writing	20	1.12	0.79	0.78	1.47	14	0.28	0.80	–0.14	0.70
Translation	7	2.04	0.79	1.46	2.63	0	–	–	–	–
Language aspect										
Vocabulary	29	1.54	0.95	1.19	1.88	22	0.68	0.96	0.28	1.08
Lexicogrammar	49	1.54	0.91	1.28	1.79	40	0.75	0.88	0.48	1.03
Grammar	18	1.24	1.08	0.74	1.74	9	0.62	1.34	–0.25	1.50
Discourse	5	1.78	1.29	0.65	2.92	3	0.31	0.25	0.03	0.59

other multiword units), often measured using multiple-choice questions to test meaning in context. This inevitably segued into the second category of lexicogrammar, though here the emphasis was less on meaning than on how the target word fits into its surroundings. The instruments often tested knowledge of collocations and colligations. This category in turn faded into grammar (e.g., tenses or articles), which itself gave way to discourse and textual awareness. Our distinctions undoubtedly reflect a substantial degree of subjective judgment, because language is itself fuzzy and notoriously resists discrete categorisation (Hunston, Francis, & Manning, 1997). Nonetheless, we felt that some grouping could be useful for several reasons. First, learners are more likely to be aware of vocabulary and grammar issues than of usage and discourse. Related to this, dictionaries and grammar books are obvious resources, with usage manuals underused, perhaps as they sit uncomfortably between the other tools, unable to provide as many entries as dictionaries or cover as much generalizable content as grammars. This means that even with an extensive command of vocabulary and grammar, learners may still have great difficulty in producing effective, natural-sounding language (cf. Hoey, 2005). This is the gap where DDL can arguably make its greatest contribution. Johns (2002, p. 109) claimed that it is "on the 'collocational border' between syntax and lexis . . . that DDL methods seem to be most effective," a claim supported by Mizumoto and Chujo's (2015) meta-analytical finding of the largest effects for

DDL in lexicogrammar. But a border only derives meaning from its neighbors. DDL's focus on context may mean that it is difficult to acquire the large numbers of words that spring to mind when thinking of vocabulary, while at the other end of the scale DDL may be "difficult to reconcile with the 'big themes' of language teaching" such as grammar (Hunston, 2002, p. 184). Many further subcategorizations would be possible, but the distinctions are delicate enough as it is, and for the moment might be jeopardized by finer granularity into, for example, collocation versus colligation. From the results obtained (see Table 9), it seems that DDL was a strong methodology for learning language per se, including lexicogrammar, especially in P/P comparisons. The differing effect sizes between the two designs for discourse again highlighted the need for caution when interpreting data from small numbers of studies.

Discussion

Our analysis uncovered a few weak results, but these were more than offset by many medium to strong ones. With a large data set such as this, however, it would be easy not to see the wood for the trees. In this section, we attempt to sketch an overall picture by returning to our three research questions, looking at the most salient patterns derived from areas with the greatest number of samples.

How Much DDL Research Is There?

It is often claimed that there is little DDL research in the sense of empirical, results-oriented investigations, but this is clearly not true. To mid-2014, we identified 205 publications reporting empirical evaluation of DDL, with output generally increasing year after year. This yielded 64 meta-analyzable studies over a 25-year period, with 88 unique samples—compare this to Norris and Ortega's (2000) 49 unique samples over 18 years for effectiveness of L2 instruction as a whole. Further, our meta-analysis did not include the many excellent qualitative DDL studies from the same period, the full extent of the fugitive literature, nor DDL studies published in languages other than English. DDL research is a flourishing field.

How Effective and Efficient Is DDL?

Effectiveness studies look at DDL's ability to increase learners' skills or knowledge through a P/P design, efficiency studies through a C/E group comparison of different ways of covering the same content. In our results, a focus on effectiveness yielded an average *d* of 1.5, efficiency 0.95. These figures both occur in the top quartile of meta-analyses in SLA covered by Plonsky and

Oswald (2014) and can thus be considered large in our field. It should also be remembered that, unlike many of the meta-analyses reviewed by Plonsky and Oswald, our unbiased *d* was weighted for sample size and winsorized to 3.0 in cases of higher values. Our results are particularly strong in relation to CALL, where meta-analyses typically have yielded small to medium effect sizes. For 12 direct C/E comparisons, Plonsky and Ziegler (2016) reported a mean effect size of 0.68. However, effect sizes in CALL correlated negatively with the number of studies sampled (–0.44), casting doubt on the robustness of the findings overall. Our outcomes paint an optimistic picture of the value of big language data that can be entrusted to learners themselves.

How Can We Best Account for Any Variation Observed?

Tempting though it may be (cf. Eysenck, 1978) to seize on a single statistic in a meta-analysis and conclude that it works, it is more interesting to look at "what works for whom in what circumstances and in what respects, and how" (Pawson & Tilley, 2004, p. 151), in other words, to look at which MVs contributed more or less to overall effects. Our intent here has been to present effect sizes for MVs as completely and transparently as possible, with the supplementary materials (as part of the Supporting Information online, available alongside this article, and through data and materials available through the Open Science Framework at https://osf.io/jkktw) providing the raw data for others to analyze in their own way.

The resulting wealth of data can make it difficult to see patterns, however, especially as we have reported effects for both P/P and C/E designs, whereas most meta-analyses in SLA have concentrated only on a single design and may have presented a simplified picture. In particular here, it is notable that the two designs did not systematically provide equivalent effect sizes, highlighting again the danger of fixating on a single value. From a possible 84 moderator variables in 25 different categories similar to those commonly tested in other meta-analyses, within-groups designs yielded 49 large effect sizes, 20 medium, 8 small, and only 1 lower than small. Between-groups designs yielded 47 large, 19 medium, 5 small, and 11 lower than small. In total, this equates to 60% large effect sizes, 24.5% medium, with only 15.5% small or negligible.

In line with most meta-analyses in applied linguistics (a notable exception being Goldschneider & DeKeyser, 2001), we did not attempt a multiple regression analysis of our MVs. Such a procedure is mainly suited to continuous MVs rather than the categorical variables emerging from most meta-analyses including ours and where every cell is filled. One way to simplify things is to concentrate only on the most robust findings, which we have done by including

only those MVs where: (a) effect sizes could be calculated for both P/P and C/E designs; (b) the CI did not include zero; and (c) the results were based on at least 10 samples since, as we saw in the case of CALL, smaller samples may be less reliable.

For the 40 MVs where both designs produced an effect size under these criteria, the correlation coefficient was .35. In 22 of the 40 cases, the P/P and C/E *d* values ranked in the same band (19 large, 3 medium). A further 15 differed by one level (14 large/medium, 1 medium/small); only 3 pairs were out by more than this, supporting the reliability of this approach. These values are given in Table 10, separated into groups of MVs that tell us something about: (a) the quality of the studies involved, (b) the situations where DDL may be more or less applicable, and (c) the effectiveness of different uses.

Quality

The date of publication seemed to have little role to play in effect size (see Table 10), though it is in recent years that the P/P and C/E designs both give large effect sizes. Although insufficient data were available from other sources, journal articles gave large effect sizes, whether from ranked or unranked journals. Most papers were between 11 and 20 pages long, again giving large effects. Duration seemed not to be a major factor either, giving one large and one medium effect size in each design, although ecological studies conducted in regular class conditions did both give large effect sizes. The remaining features all gave rise to medium or large results with two exceptions. It is unclear why large experimental group samples should give rise to such disparate results, or why instruments requiring free responses should give smaller effect sizes in P/P than in C/E designs. These exceptions notwithstanding, insofar as these features provided an indication of study quality, the large effect sizes obtained cannot be attributed to an abundance of poor-quality research.

Situations

DDL has been more extensively explored in foreign than in second language environments, achieving large effect sizes in the Middle East, and large to medium in Asia and Europe (see Table 10). This may indicate, again speculatively, that DDL is strongest where the likelihood of native speaker (English) instructors is lowest, thus offering some measure of authenticity and learner independence, possibly married to the traditional focus on language form at the expense of communication (compared to North America, for example, where learners are surrounded by authentic input, independence is imposed, communication prevails, and DDL has not been as extensively trialled). Contrary to popular

Table 10 Effect sizes for both within-group (P/P) and between-group (C/E) designs and moderator variables (MV)

MVs relating to quality

Quality	MV	P/P	C/E
Publication date	1991–2005	L	M
	2006–2010	M	L
	2011–2014	L	L
Publication type	Journals	L	L
Journal prestige	Ranked	L	L
	Unranked	L	L
Paper length	11–20 pages	L	L
Duration	Short	L	M
	Medium	L	M
	Long	M	L
Ecology	Class	L	L
EG sample size	20–49	L	L
	50+	L	0
Target instruments	Constrained	L	M
	Free	S	L
Other instruments	0	L	L
	1	M	L
	2+	L	L
Statistical tests	1	L	L
	2+	L	M

MVs relating to situation

Situation	MV	P/P	C/E
Region	Asia	L	M
	Middle East	L	L
	Europe	M	L
Context	FL	L	L
Proficiency	Int+	M	M
	Int	L	L
	Int–	L	0
Institution	Uni 1	L	L
Speciality	Languages	L	L
	Other Sciences	M	M

MVs relating to corpus use

Practice	MV	P/P	C/E
Interaction	Concordancer	L	L
	Paper	M	S
Corpus size	>100 m words	L	M
Corpus type	Public	L	M
	Local	L	L
Objective	LGP	L	L
Language use	Learning	L	L
	Reference	M	M
Language aspect	Vocabulary	L	M
	Lexicogrammar	L	M

Note. L = large (black), M = medium (dark grey), S = small (light grey), 0 = negligible (white).

opinion, it is not among postgraduates with advanced levels of proficiency that DDL has been most widely researched, and large effect sizes have been found in first-year university courses for intermediate levels, although, as noted earlier, definitions are loose in this area. The reason for the difference between P/P and C/E designs at lower-intermediate level warrants further exploration. Although large effects have been found for learners specializing in languages, medium effects have also been reported in other sciences—remembering always that medium still means in the top 50% of meta-analyses in SLA surveyed by Plonsky and Oswald (2014).

Corpus Use

Contrary to some claims, the DDL approach seems to be most effective when using a concordancer hands-on rather than through printed materials (large vs. medium/small effects; see Table 10). Tailor-made local corpora may be somewhat more effective than large public corpora, though in all cases the effect sizes were large or medium. This trend was not limited to corpora for specific purposes because large effects for general purposes were found as well, and large effects for learning as opposed to medium when corpora were used as a reference resource (e.g., in writing, though, insufficient studies could be included for different skills). When it comes to the language objectives themselves, we found large or medium effects for vocabulary and lexicogrammar, though there were again insufficient studies to warrant strong claims for or against their use in grammar or discourse.

Summary and Critical Evaluation

The aim of this study was to quantify outcomes relating to the use of the tools and techniques of corpus linguistics for L2 learning or use, which we labeled DDL. The point of a meta-analysis is not to promote or defend a given field but to provide a clearer overview, thus providing context for all. This is not definitive, as others may make different choices at any stage (defining the field, selecting papers, coding, extracting effect sizes, pooling, interpreting, etc.) and take a different perspective, even with identical research questions in the exact same field. The entire data set is available in the supplementary materials (https://osf.io/jkktw) for just this reason and can be used to explore intriguing or counterintuitive results such as how proficiency interacts with other moderator variables.

An initial pool of 205 studies that had attempted empirical evaluation of some aspect of DDL showed output increasing over time. Of these, 64 were meta-analyzable giving 88 unique samples. This suggested that there is more

empirical research than is sometimes claimed, and no evidence of publication bias was found. The average effect sizes (unbiased Cohen's *d*) were 1.50 for P/P designs, 0.95 for C/E designs, both of which count as large inasmuch as they placed them in the top quartile of meta-analyses in L2 research as a whole. Average effects seemed to be increasing somewhat over time, and large effects cannot be attributed to standard indicators of study quality. In particular, journal articles gave higher effect sizes than other publications types, and ranked journals only slightly higher than unranked journals.

In terms of the 84 MVs analyzed, 60% produced large and 24.5% medium effect sizes in the two designs—an encouraging finding overall. However, though the MVs we examined are typical of meta-analyses in applied linguistics, the current set is not exhaustive enough to account for all variation across the studies, and it may well be that other factors are responsible in the complex field that is language learning (e.g., Larsen-Freeman, 2006). Focusing only on the most robust results (i.e., MVs with at least 10 unique samples in both P/P and C/E designs), 70% had large effects, 25% medium, and only 5% small or negligible. The most consistent large effects showed that DDL is perhaps most appropriate in foreign language contexts for undergraduates as much as graduates, for intermediate levels as much as advanced, for general as much as specific/academic purposes, for local as much as large corpora, for hands-on concordancing as much as for paper-based exploration, for learning as much as reference, and particularly for vocabulary and lexicogrammar. Many of these findings go against common perceptions, and the elements missing from the list (e.g., skills or other language areas) are for the most part missing because there is as yet insufficient research rather than because research evidence is against them. There is nothing to suggest that they are inherently unamenable to a DDL approach but are rather just difficult to operationalize. 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.

A meta-analysis cannot identify which theoretical underpinnings lead to these results. However, once we know something of the effects, this may feed back into theoretical and pedagogical discussions and inform future research. We are now in a position to respond to many of the practitioner misgivings we have heard over the years with what seems to be pretty solid evidence. In no particular order, these are that learners seem able to perceive language patterns despite the lines chopped off the concordance output and that DDL activities are not confined to advanced learners, nor exclusively to simplified corpora or mediated data, nor to hands-off or paper-based activities, nor for learning

goals limited to vocabulary and collocation. The evidence on all these seems clear.

Although meta-analysis has until recently remained a fairly marginal methodology in applied linguistics, it has already generated some traditions, and one of these leads us to a complaint about the quality of the work that we have reviewed. Our complaints about DDL work are not extensive, and the increasing effect sizes speak for themselves. Still, an initial count of 205 DDL studies was reduced to 64 partly due to our inclusion criteria but also because of missing data and incomplete reporting. Even for studies that were included, the coding sheet in the Supporting Information online shows blank cells for seemingly basic information such as corpora and software used, language objectives and test instruments, materials and procedures, and participant information. Full meta-compliant recommendations for good reporting can be found in Larson-Hall and Plonsky (2015).

Some of the weaker results in our collection reflect a plethora of research questions, especially in Ph.D. dissertations. As in other areas of applied linguistics, unambiguous research questions, motivated and doable experimental tasks, focused objectives, transparent study designs, and outcome measures clearly tied to learning tasks all led more frequently to strong results than the alternatives did. Other weak effects are simply a factor of the limited number of studies focusing on a given area, especially for languages other than English, in contexts outside higher education, for personalized use outside class, and with focus on specific skills or language areas in addition to vocabulary or lexicogrammar. We cannot know if researchers are right to avoid DDL in these cases until it has actually been studied there.

Another pressing need is for more longitudinal and delayed posttesting. If a prime rationale for DDL is that it should be good for autonomy, learning to learn, consciousness raising, and other forms of long-term change in thought or action, then this should be evidenced in strong results on delayed posttests. We have avoided this complex issue here for reasons of space, but we should note that of the very few delayed tests reported, many are within Ph.D. dissertations, and that Ph.D. dissertations overall exhibit substantial differences in effects according to design (large within-groups, small between-groups). Nonetheless, the delayed effects are anomalously low, suggesting that it is not just the difficulty inherent in testing for one factor in the complexity of long-term studies and delayed tests, but that something else is likely going on. Given the theoretical importance of delayed testing in DDL, we hope that our exposure of this problem will encourage researchers to give it greater attention in future, or to reexamine the studies covered here with a view to clarifying this particular issue.

Conclusion

As far as we know, our meta-analysis represents the first time that DDL work has been brought together for all to see and consider as a whole—including for DDL researchers. For all of us, it should have become more apparent why this work is worth doing, what questions are most worth asking, which designs worth pursuing, and what data must be included to assure one's work will be included in the next meta-analysis—particularly information from the period between immediate and delayed posttest. With the to-do list properly laid out, we conclude that the future of DDL looks rather bright. There is a corpus revolution underway in both applied linguistics and language instruction (e.g., McCarthy, 2004) and what we have found here suggests that even learners can participate. This would ideally be confirmed in a subsequent remake of the present study in 5 years using some or all of the categories and variables that we have identified.

Final revised version accepted 10 October 2016

Notes

1 All instances except two (lines 57 and 414) in this random sample of 20 are adverbial, metaphorical, nonanatomical uses of *back.* This proportion of 90% nonanatomical uses will be found to be similar in virtually any other corpus (except a medical corpus where some samples will fall to about 85% unless the topic is specifically the *human back*). Although dictionaries may have pedagogical reasons for presenting prototypical, concrete meanings first, the majority (e.g., http://www.wordreference.com/definition/back) start their entries for this word with the nominal/literal/anatomical sense, implying that this is the most important sense to learn.

2 References to all meta-analyzed studies are included in Appendix S1 in the Supporting Information online.

3 Actually they suggested 0.7 and 1.0 for medium and large C/E effects, but this is a hybrid of meta-analyses and primary studies, unlike their P/P recommendation, which includes meta-analyses only. We have adopted the latter procedure here.

4 http://www.scimagojr.com/journalrank.php?category=1203. Impact factor is of debatable relevance to quality, but it does give a general idea of the relative prestige of various journals.

References

Aston, G. (1998). *Learning English with the British National Corpus: 6th Jornada de Corpus*. Barcelona, Spain: Universitat Pompeu Fabra. Retrieved from http://sslmit.unibo.it/~guy/barc.htm

Barrett, L., Dunbar, R., & Lycett, J. (Eds.). (2002). *Human evolutionary psychology*. Basingstoke, UK: Palgrave.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

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

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

Boulton, A. (2015). Applying data-driven learning to the web. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 267–295). Amsterdam: John Benjamins.

Boulton, A. (in press). Corpora in language teaching and learning: Research timeline. *Language Teaching*, *50*.

Burston, J. (2013). Mobile-assisted language learning: A selected annotated bibliography of implementation studies 1994-2012. *Language Learning & Technology*, *17*, 157–225.

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*, 4–20. doi:10.1017/S0958344014000159

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

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

Chinnery, D. (2008). You've got some GALL: Google-assisted language learning. *Language Learning & Technology*, *12*, 3–11.

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development*, *47*, 15–33. doi:10.1007/BF02299631

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge, UK: Cambridge University Press.

Ellis, N. C. (2006). Meta-analysis, human cognition, and language learning. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 301–322). Amsterdam: John Benjamins.

Ellis, R. (2015). Introduction to the special issue on synthesizing research on form-focused instruction—the complementary contributions of narrative review and meta-analysis: Complementarity in research syntheses. *Applied Linguistics*, *36*, 285–289. doi:10.1093/applin/amv015

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517.

Felix, U. (2008). The unreasonable effectiveness of CALL: What have we learned in two decades of research? *ReCALL*, *20*, 141–161. doi:10.1017/S0958344008000323

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

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50. doi:10.1111/1467-9922.00147

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology supported language learning. *ReCALL*, *25*, 165–198. doi:10.1017/S0958344013000013

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

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.2307/1164588

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

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558. doi:10.1111/0023-8333.00164

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal*, *51*, 208–216. doi:10.1093/elt/51.3.208

Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165–211). Amsterdam: John Benjamins.

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. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 107–117). Amsterdam: Rodopi.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on web as corpus. *Computational Linguistics*, *29*, 333–347. doi:10.1162/089120103322711569

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*, 590–619. doi:10.1093/applin/aml029

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159. doi:10.1111/lang.12115

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366. doi:10.1093/applin/amu040

Li, S., Shintani, N., & Ellis, R. (2012). Doing meta-analysis in SLA: Practice, choices, and standards. *Contemporary Foreign Language Studies*, *384*, 1–17.

Lin, H. (2014). Establishing an empirical link between computer-mediated communication (CMC) and SLA: A meta-analysis of the research. *Language Learning & Technology*, *18*, 120–147.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.

McCarthy, M. (2004). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga: The Irish Yearbook of Applied Linguistics*, *20*, 30–52.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*, 129–148. doi:10.1093/applin/amq035

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

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art—and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 5–24). Frankfurt, Germany: Peter Lang.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528. doi:10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2001). Does type of instruction make a difference? Substantive findings from a meta-analytic review. *Language Learning*, *55*, 157–213. doi:10.1111/j.1467-1770.2001.tb00017.x

Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.

Norris, J. M., & Ortega, L. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly*, *41*, 805–815. doi:10.1002/j.1545-7249.2007.tb00105.x

Oghigian, K., & Chujo, K. (2010). An effective way to use corpus exercises to learn grammar basics in English. *Language Education in Asia*, *1*, 200–214.

Ortega, L. (2010). Research synthesis. In B. Paltridge & A. Phakiti (Eds.), *Companion to research methods in applied linguistics* (pp. 111–126). London: Continuum.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. doi:10.1017/S0267190510000115

Pawson, R., & Tilley, N. (2004). Realist evaluation. In H.-U. Otto, A. Polutta, & H. Ziegler (Eds.), *Evidence-based practice: Modernising the knowledge base of social work?* (pp. 151–182). Opladen, Germany: Barbara Budrich.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*, 993–1038. doi:10.1111/j.1467-9922.2011.00663.x

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470. doi:10.1111/j.1540-4781.2014.12058.x

Plonsky, L., & Brown, D. (2014). Domain definition and search techniques in meta-analyses of L2 research (or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–268. doi:10.1177/0267658314536436

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591–610. doi:10.1093/applin/amu001

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

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

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

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

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

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

Smith, S. (2011). Learner construction of corpora for general English in Taiwan. *Computer Assisted Language Learning*, *24*, 291–316. doi:10.1080/09588221.2011.557024

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*, 263–308. doi:10.1111/j.1467-9922.2010.00562.x

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Taylor, J. (2012). *The mental corpus: How language is represented in the mind*. Oxford, UK: Oxford University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Yeh, Y., Liou, H.-C., & Li, Y.-H. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, *20*, 131–152. doi:10.1080/09588220701331451
Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*, 39–58. doi:10.1080/09588221.2010.523285

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. References to All 205 Studies Included in the Meta-Analysis.
Appendix S2. Coding Scheme.
Appendix S3. Procedure for Calculating CIs.
Appendix S4. Effect Sizes Pooled for Each Unique Sample.
Appendix S5. Effect Sizes for Each Subquestion.
Appendix S6. Effect Sizes for Analyses of Moderator Variables.
Appendix S7. Summary of Robust Moderator Variables.