Thirty years of data-driven learning: Taking stock and charting new directions

Alex Boulton

This is a pre-publication version. Where possible, please refer to the definitive text (OA):
Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning & Technology*, 25(3), 66–89. https://doi.org/10125/73450

Thirty years of data-driven learning: Taking stock and charting new directions over time

ABSTRACT

The tools and techniques of corpus linguistics have many uses in language pedagogy, most directly with language teachers and learners searching and using corpora themselves. This is often associated with work by Tim Johns who used the term Data-Driven Learning (DDL) back in 1990. This paper examines the growing body of empirical research in DDL over three decades (1989-2019), with rigorous trawls uncovering 489 separate publications, including 117 in internationally ranked journals, all divided into five time periods. Following a brief overview of previous syntheses, the study introduces our collection, outlining the coding procedures and conversion into a corpus of over 2.5 million words. The main part of the analysis focuses on the concluding sections of the papers to see what recommendations and future avenues of research are proposed in each time period. We use manual coding and semi-automated corpus keyword analysis to explore whether those points are in fact addressed in later publications as an indication of the evolution of the field.

Keywords: Data-Driven Learning, DDL, Corpus Linguistics, Concordancing, Corpus-Based Language Learning

INTRODUCTION

Since corpora emerged in their modern form in the 1960s as large collections of electronic texts designed to represent an area of language use, numerous applications have been found for them in language teaching. These applications have been both indirect, with specialists using corpus-derived information for improved language descriptions leading to new dictionaries and other resources, and direct, with language teachers and learners searching and using corpora themselves. The latter type of application was given prominence in work by Tim Johns at Birmingham University who used the term Data-Driven Learning (DDL) as early as 1990. Many researchers associate DDL with a central hard-core involving "the hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning of advanced usage" (Boulton, 2011, p. 572). However, it can extend in many directions away from this prototypical core so that some activities may be more or less 'DDL-like', a broader definition with fuzzy boundaries encompassing many different ways of "using the tools and techniques of corpus linguistics for pedagogical purposes" (Gilquin & Granger, 2010, p. 359). This is the definition adopted here, with the addition that the pedagogical purposes are for foreign or second language (L2) learning, teaching or use.

In line with the usual evolution of new techniques in applied linguistics (Shintani et al., 2013), early DDL publications were mostly descriptive or speculative, containing suggestions for

corpus-based teaching activities; less than a dozen empirical studies had appeared by the end of the century. Since then, the body of empirical DDL research has been growing rapidly, including two special *LLT* issues in 2001 and 2017. The general topic of 'corpora' was fifth of 43 themes (Gillespie, 2020) detected in a selection of journals in CALL (Computer-Assisted Language Learning). The aim of this paper is to survey the evolution of DDL over 30 years, with a focus on empirical studies. Specifically, the integrated double analysis – manually coding the scope of the papers and semi-automatically analyzing them as a corpus for recurrent keywords – allowed us to examine not only what has been done, but also what future directions were suggested by the researchers, how (if at all) these suggestions changed over time, and how (if at all) they were realized in subsequently conducted and published studies.

A number of DDL researchers have provided position papers on DDL, recently including two plenary speeches by key DDL names in *Language Teaching*. Chambers (2019) seeks to "bridge the research-practice gap", noting that DDL is mainly the practice of aficionados, and underlining the need for research among 'regular' teachers (i.e., non-researchers) – a request that could apply to countless other areas of CALL research. O'Keeffe (2020) echoes Flowerdew (2015) in noting the lack of theoretical underpinnings in DDL and calling for a "broader research gaze" beyond constructivism and sociocultural theory to position it more firmly within the field of Second Language Acquisition (SLA). Tribble (2015) is yet another example of such position papers that take stock of the field, though they often tend towards a personal or theoretical stance rather than a meticulous review as such.

A number of scholars have already provided different types of syntheses for DDL, beginning with Chambers (2007). Her in-depth analysis of 12 studies reached the primary conclusion that the approach was worthwhile, but that there was a serious lack of stringent quantitative evaluation, a challenge taken up in numerous studies that followed. Currently, however, such complaints have reversed, with Pérez-Paredes (2019) now lamenting that "the fact that the pool of papers examined is fundamentally empirical may have contributed to a lack of theoretical positioning" (p. 17). As the first synthesis, it is not surprising that Chambers' sample is fairly small, though it is by no means exhaustive. A later attempt by Boulton (2010) to focus on learning outcomes from 27 papers also found grounds for optimism. Some syntheses restrict the field: Boulton (2012a) to ESP (English for Specific Purposes) through 20 DDL papers; Luo and Zhou (2017) to writing in 18 papers published between 2010-2016; Chen and Flowerdew (2018) to DDL in EAP (English for Academic Purposes) from 37 papers, mainly from 2010-2017. Godwin-Jones (2017) organized his survey by the type of DDL technology used. These studies tend to highlight the tremendous diversity of objectives and instruments, tools and uses, while at the same time noting areas that are in need of further research. As with the position papers discussed earlier, such narrative syntheses provide valuable insights, though they are often limited in scope, with fairly serendipitous collection procedures and manual analysis. This may be deliberate as in Boulton (2017), which provided a timeline of personally selected papers for individual comment, but may at times be due to a lack of methodological rigor. An original perspective is provided in He and Wei (2019), a mainly bibliometric analysis of co-citation clusters (i.e., groups of articles which feature the same citations) in 328 papers collected using the keywords *EAP* and *corpus* from journals in the Social Science Citation Index between 2009 and 2018. Most recently, Pérez-Paredes' (2019) systematic review converted the papers he examined into a corpus, allowing the identification of frequent key clusters alongside a coding

scheme that focuses particularly on normalization (Chambers & Bax, 2006) of DDL within CALL. The limitations here are the focus on just five major CALL journals from 2011-2015, for a total of 32 papers. The most inclusive study to date is Boulton (in press), finding 351 empirical studies up until 2018. The methodology is used as a foundation for the present paper, with the trawls and coding now extended and analyzed historically, combined with a corpus keyword analysis of the conclusion sections.

Finally, there have been a number of meta-analyses of DDL. One advantage of this type of synthesis is the attempted rigor in collecting and analyzing the data; however, only quantitative data can be included, so valuable qualitative research is neglected. The first to meta-analyze DDL were Mizumoto and Chujo (2015), who surveyed 14 studies, all in Japan and featuring the second author, mainly for lexicogrammar among lower-level learners. Their pre/post-test comparison found medium effect sizes overall. This was followed by Cobb and Boulton (2015) with 21 studies, updated in Boulton and Cobb (2017) for 64 publications and over 3000 participants. The rigorous trawls and broad sweep to include potentially any application of DDL result in the most comprehensive meta-analysis to date, finding large effect sizes in both pre/post-test and control/experimental designs. Their overall finding was that "DDL works pretty well in almost any context where it has been extensively tried" (p. 386), with medium or large effect sizes recorded for almost all moderator variables based on at least 10 samples. The most recent meta-analysis can be found in Lee et al. (2019), who focused exclusively on vocabulary in 29 studies. Their more sophisticated multi-level analysis provides valuable insights in this area, with medium effect sizes the norm. These last two papers both provide the complete data for further analysis, but the conclusion is clear: it is not now so much a question of whether DDL works, but how it may best be used with different learners in different contexts for different purposes.

Against this backdrop of reviews and syntheses of DDL, the present paper describes rigorous trawls conducted to uncover a near-exhaustive collection of 489 empirical evaluations of some aspect of DDL over three decades (1989-2019), published in English in academic journals, book chapters and conference proceedings; this makes it the largest such collection to date. The texts were converted into a corpus of over 2.5 million words to aid the analysis using corpus tools. The papers were read and coded for context, participants, research design and methodology, learning tasks and other variables, all leading up to the final theme of the technology used, i.e. the corpora and tools. Rather than simply treating the collection as a uniform whole, the papers are divided into five time periods for a historical perspective: as new research appears, it is to be expected that new areas will be addressed, whether pedagogical, technical or methodological. The main part of the analysis thus focuses on the conclusion sections for recommendations for future work, and assesses to what extent this has been converted into new research strands. Specifically, two kinds of data will be compared in the analysis: (1) the scope of the papers, as captured by the coding scheme of publication types and research designs; (2) the themes for future directions identified in the Conclusion sections, particularly via the grouping of key words and phrases. The aim here is to map the field, identify gaps in the existing research, and determine areas in need of further exploration.

Research questions

The following research questions guided our study:

1. How did the publication scope of empirical DDL research change from 1989 through 2019?
2. What themes for future research directions were identified in these studies and how did the frequency of these themes change over time?
3. How were these suggested future directions realized in the study designs in subsequent periods?

Questions (1) and (2) were explored first, and the results were used to inform the analysis conducted in response to question (3), which is the main research question of our study.

METHODOLOGY

Collection

The objective was to analyze empirical DDL studies, with DDL defined as the use of corpus tools and techniques for pedagogical purposes in a foreign/second language; studies with native speakers only were excluded, with the exception of initial or in-service teacher training. Note that this definition is deliberately neutral on both the 'text' to be explored (which is not necessarily a corpus *per se*) and the procedures (whether the learners work 'hands-on' with the electronic corpora and tools, or 'hands-off' with print-outs or other pre-processed data derived from corpora, typically concordances, frequency lists, collocations, etc.). Empirical studies are also broadly defined, as any publication that provides some type of evaluation of DDL, from experimental designs exploring learning outcomes to observational studies of learning processes to feedback questionnaire studies; this may be the entire focus of the paper, or just a minor section. Papers were excluded if they only described corpora or tools, analyzed corpora (notably learner corpora), talked about potential or actual examples of DDL but with no evaluation other than the teacher-researcher's impressions, and so on.

Since one objective was to create a searchable corpus of published research, only full texts in English were included, notably journal articles, book chapters and conference proceedings. Conference posters and oral presentations, slides, notes and other text types were excluded, as were PhD and master's dissertations since their length would potentially bias the corpus; they are also unsystematically available, though extracts are often reworked in other formats. On the other hand, short write-ups of conference papers or even in some cases long conference abstracts are included where they meet the other inclusion criteria. It is also clear that dozens or hundreds of papers in other languages do exist, at least in the European languages the authors are familiar with, and in several Asian languages.

The collection was started over 10 years ago, with recent versions being described in Boulton and Cobb (2017) and Boulton (in press). New trawls were conducted for the present study by a graduate research assistant under the authors' guidance. These used combinations of key words including *DDL*, *data-driven*, *corpus/corpora*, *concordanc**, and *Johns*; contextualizers were often needed to reduce noise, notably *language* and *learning*. The main databases searched were LLBA, MLA, ERIC, JSTOR, DOAJ, Web of Science, Academia, ResearchGate and Google

Scholar (filetype:pdf to reduce noise); a number of new resources such as 1findr, SemanticScholar and ORCiD produced few or no additional items and were quickly abandoned. When 50 consecutive results produced no new hits, that source was discontinued. When a potential hit was identified, the abstract was read and the paper located and downloaded; borderline cases were discussed between the two authors. All the reference lists were scoured for further leads, and any source (journal, publisher or conference series) that provided two or more hits was individually followed up. Sometimes the process of locating the full text produced yet more results, as searching the title online could lead to the author's homepage with other studies, or publications that cited the original target.

Our objective was to include all and any papers that fit the criteria, regardless of other considerations. For example, some studies are reported in more than one paper, often with substantial overlap of the text itself (e.g., short conference papers later developed in journals); we also identified four cases of apparent plagiarism, but kept the papers here as our aim is to look at what is available, regardless of its potentially severe demerits. Quality is a tricky topic and largely subjective (Lipsey & Wilson, 2001): on reading a paper, one might be impressed by the freshness of the research questions or the elegance of the design; or, conversely, bored by the lack of originality or shocked by apparent flaws in the methodology. To sidestep this issue, many systematic reviews include articles from certain journals only (e.g., Pham et al., 2014; Pérez-Paredes, 2019), the downside being that large quantities of research go unacknowledged. Our approach was to ignore quality itself, but to conduct parallel analyses of just those journals that feature in the top 100 of linguistics in the JCR rankings (Journal Citation Reports; Clarivate Web of Science, 2019). Such bibliometric systems have been roundly criticized, and a journal's impact factor can in no way be taken as a guarantee of the quality (or lack of it) in individual papers; nonetheless, it does provide a rough-and-ready indicator of the likely impact and readership a paper may have (Schöpfel & Prost, 2009). To avoid doubling the wordcount by systematically reporting for JCR and other publications, the comparisons will be brief and only reported where substantial differences are identified. Full data can be found in the online supplement so that others can sort and process the data in other ways. The next two sections explain how we came up with the taxonomies for the two types of data used in our analysis – publication scope (captured in the article coding sheet) and future direction themes (captured via corpus analysis of keywords in article conclusions).

Publication scope: coding sheet and procedure

The collection was first coded as a form of scoping review to "map the existing literature in a field of interest in terms of the volume, nature, and characteristics of the primary research", especially useful when the topic is "of a complex or heterogeneous nature" (Pham et al., 2014, p. 371). The original manual was drawn up by Boulton and Cobb (2017), informed by other meta-analyses and syntheses in applied linguistics. The spreadsheet essentially divides into four sections:

- Publication: date and source of publication, length.
- Population: L1 and L2, country and second/foreign language context, proficiency, university or other situation, specialization and comparing language for general, specific or academic purposes (LGP, LSP, LAP).

- Procedures: duration, corpora and tools (if hands-on), interaction type, language focus, and whether it has DDL as a learning aid or a reference resource, or looked at learner attitudes or behavior during use.
- Design: number of participants in control and experimental groups, the instruments used and whether the data were analyzed statistically or qualitatively.

The first author had read and coded the first batch (401 papers). The second author independently read and coded a random subset of 10, and disagreements were resolved; this process was repeated until no new problems arose, subsequent to which she read and coded the 88 new papers found. The way the collection has grown over time does however mean that results should be taken as indicative rather than absolute. This is inevitable in a study of this scale as a complete re-read of all 489 papers would be prohibitively time-consuming. Some categories are entirely factual and very rarely problematic (e.g., date and source of publication), while others involve "best guesses due to insufficient information given in the primary studies" (Lin, 2014, p. 135). This particularly concerns the level of proficiency; most often, we used the labels given by the authors of each paper despite noted problems (Burston & Arispe, 2018). Duration is another sticking point, since it is variously given in minutes, hours, sessions, weeks, months, semesters, etc.: a semester may be fairly intensive with several long classes over 15 weeks entirely devoted to DDL, or it may be half a dozen classes where only a few minutes are dedicated to corpus work.

Identification of future direction themes: corpus keyword analysis

With a large collection of publications as we have here, corpus tools can be useful in helping to explore specific questions rather than relying on regular reading alone; however, we wanted to retain the human analysis and not rely entirely on automatic black-box document clustering, summarization or other procedures. For this, we used the freeware AntConc (Anthony, 2019). The first author converted the complete set of texts to txt format, cleaned and checked to ensure that only the original text remained. In particular, meta-data (affiliations, contact, acknowledgements, etc.) were removed along with headers and footers, figures and tables, primary and secondary data extracts (with the exception of academic quotations), references and appendices. The texts were checked (e.g. for ligatures, hyphens, mathematical symbols, diacritics, etc.) but not edited in any other way; any errors in the original were left in place. This gave a full corpus of 2,563,589 tokens which could be explored in its entirety to help with the analyses.

To identify themes suggested by article authors for future research directions, we used a combination of automated and manual analysis. First, we created a Conclusions corpus of final sections comprising 253,569 tokens. Next, we used AntConc tools to identify words and *n*-grams that were significantly more frequent in Conclusions compared to the full corpus. The Keyword tool analysis yielded 140 types and 66,699 tokens with positive keyness. Many of the Conclusions keywords were metalinguistic text structuring devices, including subheadings, typical of the sub-genre 'research article conclusions' (e.g., *conclusion(s)*, *limitations*, *implications*), which were ignored for the current analysis. Words that could potentially be relevant to our main theme of interest (future research directions), were explored individually

with the AntConc Concordance tool. A final list of 11 words was identified that primarily appeared in relevant contexts and in immediate clusters with other relevant words (

APPENDIX **A**). For the key *n*-gram comparison, the Conclusions corpus and full corpus *n*-gram lists were created first. The search parameters were set to 3-4 words, with the minimum occurrence of 3 times in at least 3 articles. Next, the Keyword tool was used to search the Conclusions corpus list for key *n*-grams in comparison to the full corpus list. This comparison yielded 436 key *n*-gram types for 11,890 *n*-gram tokens. Many of the Conclusions key *n*-grams (similar to keywords, see above) were metalinguistic text structuring devices (e.g., *the findings of*, *the results of*, *the present study*, *discussion and conclusion*), which were again ignored for the current analysis. The remaining *n*-grams (41 types, 646 tokens) were combined in clusters by common theme (

APPENDIX **B**).

As the next step, we looked for keywords that characterized each time period in comparison with all other time periods. After using the Keyword tool and weeding out irrelevant hits from the total of 133, we identified 18 keywords in the 1989-2003 Conclusions subcorpus, 23 in 2004-2007, 11 in 2008-2011, 19 in 2012-2015, and 30 in 2016-2019 (APPENDIX C). The key *n*-gram comparison per time period yielded no obviously meaningful results.

In the final theme identification step, we grouped the resulting 15 themes (inferred from the 11 keywords and 41 key *n*-grams from the whole Conclusions corpus and 101 period-specific keywords) into 4 overarching categories (Table 1).

Thematic categories	**Themes**	**Keywords and *n*-grams**
Theory and methodology	• sample size • data collection instruments • duration • theoretical considerations	11 overall keywords (retained from 140), APPENDIX **A** 41 overall key *n*-grams (retained from 436), APPENDIX **B** 101 period-specific keywords (retained from 133), APPENDIX C
Learning contexts	• language proficiency • institution • disciplinary specialization • geographic distribution	
Implementation	• interaction with DDL • consultation • guidance • autonomy • language skills	
Technology	• software or interface • specific corpora • corpus size	

Table 1. Taxonomy of future directions themes from the Conclusions corpus.

RESULTS

Our analysis was divided into two steps as reported in the two sections below. The first describes the scope of the articles in our collection vis-à-vis the coding sheet rubrics; the second discusses changes in all identified themes in relation to the scope of the studies across time periods. The first section is shorter as it reports on the analyses conducted to enable the main analysis reported in the second section.

Scope of the articles in the collection

The initial pool in Boulton and Cobb (2017), collected using similar procedures as here, consisted of 181 DDL papers (not counting 17 PhD dissertations and 7 papers in other

languages); Boulton (in press) unearthed 351. The present study uncovered 489 in total, an increase of 170%. This rise is partly due to the large numbers of papers being published, partly to better indexing and the use of additional databases and improved procedures. This underlines, first, the general health of DDL as a research field; and second, that varied and rigorous methods are essential to tap the large numbers of studies that often fly under the radar, but that even now it would be naïve to claim we have an exhaustive collection.

The first study to meet our criteria is Baten et al. from 1989; the end cut-off point was set at 2019, the last full year prior to writing. While early publications are essential for a historical overview, it takes time to build momentum and for a substantial body to appear, and we grouped these from 1989-2003 for a total of 44 papers. This enabled us to settle upon four equal time periods of four years each for the rest as numbers increase: 2004-2007, 2008-2011, 2012-2015, 2016-2019. As each has different numbers of publications, main results will be given as rounded percentages for comparison purposes.

Of the 489 papers, 361 (74%) are journal articles from 185 different sources, 135 of them featuring only one paper. This suggests that a focus only on DDL-friendly sources will lead to major omissions in coverage. The most popular journals for DDL publications are *ReCALL* (28 papers), *CALL* (25), *LLT* (23) and *System* (11); no other journals reach double figures. It is worth noting that all 4 of these journals are in the JCR100 list, which in total accounts for 117 papers in the corpus (24%) from 15 different journals. This subset of nearly a quarter of the corpus is used as a comparison for the rest of the collection – not as a guide to quality in itself, but as an indicator of whether surveys limited to just top journals reflect the entire body of research we have identified. Figure 1 shows the evolution of publication, with the dotted line representing a three-year smoothing (i.e., averaged over a three-year period). Figure 2 shows the papers grouped into periods as outlined above, for the JCR100 list and others.

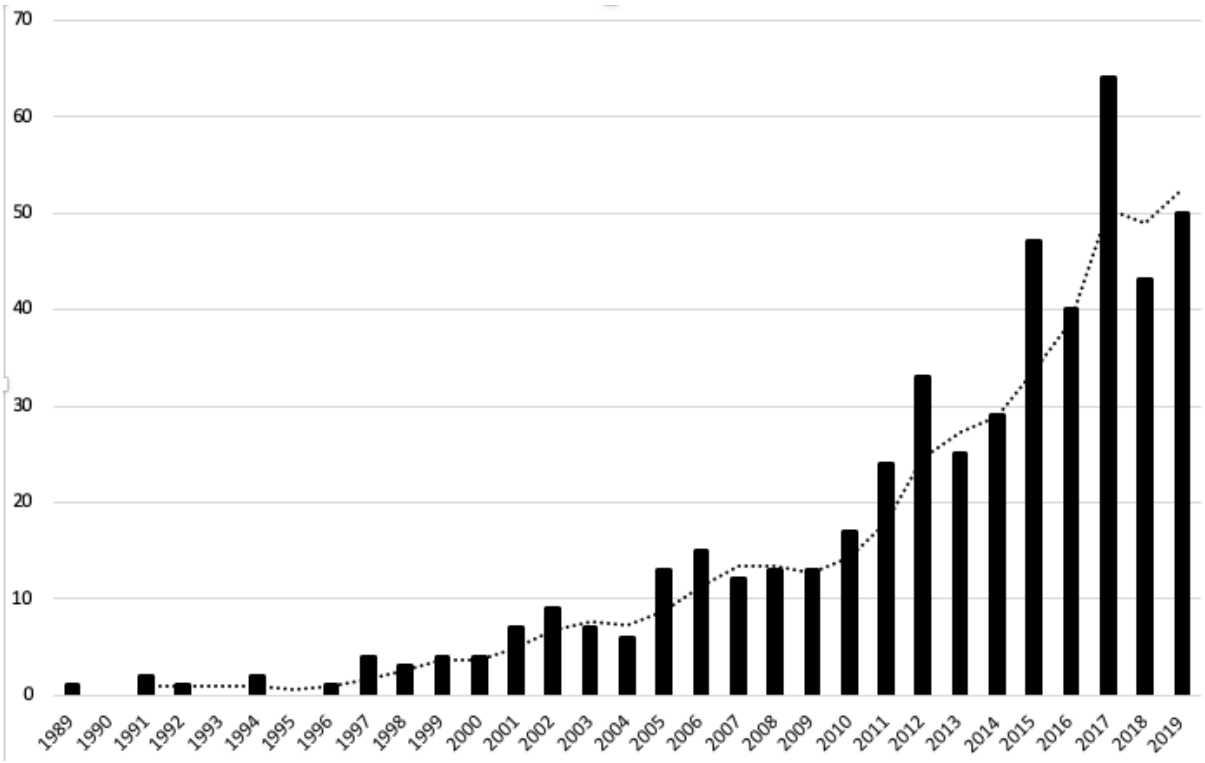

Figure 1. Article counts by publication date

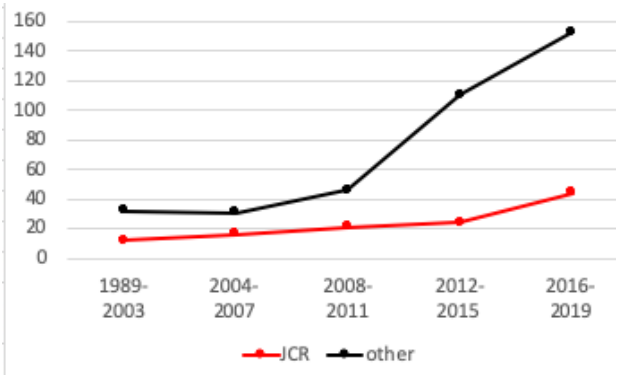

Figure 2. Counts of JCR100 and other papers by period

Of the rest, 62 are overtly conference proceedings (13%), although many more conference presentations were written up as journal articles or chapters in books deriving from conferences – the biennial TaLC (Teaching and Language Corpora) conferences in particular have been publishing selected papers since 1994. Overall there are 61 book chapters (12% of the total), including 13 from Peter Lang, 10 from Routledge, 7 from Rodopi and 6 from John Benjamins. Finally, 5 are miscellaneous – individual working papers or unspecified sources (1%).

Changes in the identified themes in relation to the scope of the studies across time periods

A first group of themes concerns the nature of the research and how it is conducted. DDL research has primarily been experimental throughout its history, to the extent that the lack of

theorization has been pinpointed as a potential problem. The keywords associated with the theme **theory** in article Conclusions (

APPENDIX **B**) show how references to this theme have changed over time, clearly in line with specific SLA theories adopted at each time in DDL. The concept of *transfer* associated with the generativist theory is used throughout the timeline but is key only in the earliest period. A number of terms associated with constructivist SLA theory appear as key during three of the later periods: *hypertext* (2004-2007), *intelligences*, *encoding* (2012-2015), and *enhancement* (2016-2019), while sociocultural theory comes to the fore in 2008-2011 with the concepts *social*, *materializations*, *scaffolding*. It must be noted that some of these terms are associated with a single author and appear only in Conclusions of the few theoretically oriented articles. The fact that the authors draw attention to these theories shows a desire to inspire more theory-grounded research in the future. However, few papers in our empirical collection actually do this (cf. Pérez-Paredes, 2019, discussed above).

Methodology themes

Methodology as a theme in its own right only appeared in 2016-2019 Conclusions through the keywords *experimental*, *questionnaire*, *scores* (41 occurrences,

APPENDIX **B**). In general, prior to 2004, 43% of the studies were purely qualitative, giving no indication of numerical data whatsoever – a possibly ironic finding given the nature of corpus linguistics and DDL, though it is difficult to generalize given the range of objectives covered in different papers. On the other hand, 25% give raw numbers, percentages or descriptive statistics, and 32% performed some kind of statistical analysis. However, the figures stabilize after that for overall rates of 15%, 35% and 50% respectively. The JCR100 have a slightly greater portion of statistical analysis, but the difference is small (54% vs 49%).

Sample size is often acknowledged as a limitation in Conclusions, with calls for a larger number of participants in future studies (Figure 3). While not a big concern in the earliest period, it has regularly been mentioned in 18-22% of Conclusions since 2004. Despite these persistent requests, the average group size has remained remarkably stable between 38 and 42, presumably as this reflects convenience sampling and typical (university) class sizes. The JCR100 papers involve substantially larger groups than others (45 vs 36); however, the standard deviation is very high at 54. The median is 26, with 178 studies involving fewer than 20 DDL students, including 61 with fewer than 10, and 12 just one or two. At the other end of the scale, 30 had 100 or more experimental participants, the largest study involving 526. Clearly larger sample sizes increase power, but in DDL as in other areas of applied linguistics, researchers are limited by institutional considerations.

Figure 3. Sample concordances for the theme *sample size* from the Conclusions corpus

The choice of **instrument** clearly depends on what the researchers want to focus on, though a general finding is that most studies use more than one instrument (mean = 1.9). The single most popular instrument is the questionnaire, used in 53% of papers, though only in 49 cases is it the sole tool. These papers are mainly interested in learners' perceptions of DDL and, in some cases, what they did. Other perception instruments include individual interviews (21%) or group discussions / focus interviews (8%), while behavior is also monitored by a wide variety of self-reports (journals, diaries, logs, think-aloud, stimulated recall: 17%), teacher observation or field notes (8%), tracking or screen recording (7%). A few (2%) analyzed project work or reports of some kind, 23% analyzed productions (usually writing or self-correction, occasionally translation or, in 3 cases, spoken production), with more controlled tests featuring in 45%. The JCR100 are more likely to use tracking (15% vs 5%) and questionnaires (60% vs 51%), but actually slightly less likely use 'tests' (43% vs 46%). This is perhaps because their aim is not (or no longer) to see if DDL 'works' but more subtle phenomena. Less formal instruments (reports, discussions, observations) have generally decreased over time (together, they were used in 43% of publications prior to 2003, declining in each period to 22% in 2016-2019), but so has the use of tracking, from a high of 13% in 2004-2007 to 5% in 2016-2019. Other changes over time are minimal or not obviously directional, an interesting finding in itself.

There have been numerous and increasing calls for more **longitudinal research or longer DDL treatments**, explicit or implied by mentioning that their own study was too short. In the three latest periods, 19-22% of the Conclusions mentioned this theme. In reality though, there is no obvious tendency towards longer or more ecological research, although this is difficult to assess given differing reporting practices (hours, weeks, sessions, semesters, etc.). Sometimes it is inherent in laboratory designs, with 19 studies lasting less than an hour, 34 a single session. From 2008 on, at least, the studies seem to have similar duration regardless of how they are measured: 13 hours, 9 sessions, 8 weeks, or slightly over 1 semester for the longer ones (Figure 4). Nor is there any major difference between the JCR100 and other papers, although the former do show somewhat higher means for hours (15 vs 13) and sessions (12 vs 8). Only 9% included a longitudinal analysis via delayed post-tests, with another 5% including delayed questionnaires distributed after the end of the experiment. Therefore, despite the recognized need, little is being done to investigate extended DDL or follow up after use, likely due to planning considerations and the number of variables involved, as well as pressure to publish.

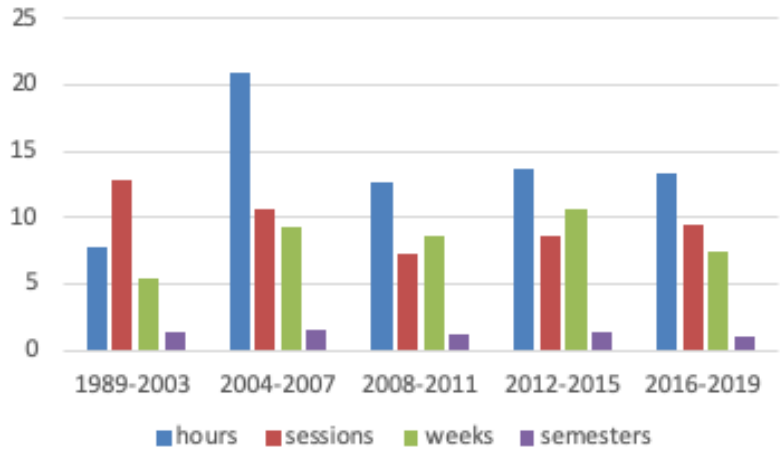

Figure 4. Mean duration with the measures available

Learning / contextual themes

A second group of themes concerns the learners themselves and their particular situations. **Language proficiency** has been cited frequently in the Conclusions, between 22% and 31% of articles mentioning it in each time period. This generally calls for a wider variety of proficiency levels to be explored, especially lower levels (Figure 5). Actual research seems to have responded to these calls, at least to some extent: in the first two periods, only 22% were with lower levels compared to 37% today. It may be that having established the credentials of DDL at higher levels, researchers are now experimenting to see how far they can go. More prosaically, it may be that the tools are more amenable, that there are more relevant corpora for lower levels, or that the techniques have been refined and adapted to new contexts. Nonetheless, more work would seem to be needed with lower levels of proficiency, which is not necessarily synonymous with age.

beyond a number of participants' current level	of language proficiency. 3. Some concordances were culture-bound,
implementation in schools in Korea. Irrespective	of language proficiency, students benefited from CCL for
on groups of students with different levels	of language proficiency. The result corroborated with the
ting. Quasi-experimental research that uses tests	of language proficiency to more accurately measure the
corpus consultation by students of various levels	of language proficiency. To achieve this goal, teachers
ectively scaffold learners' collocation learning.	Low-level students benefited more from collocation instruction (
tivation by working collaboratively. However, CCL	low-level students did not perform better than
of verbs. An interesting finding was that	low-level students from the ICL group had
be a decrease in the attitudes of	low-level students. In this sense, the underlying
high-level students, medium-level students and	low-level students respectively by using the method
alized expressions than medium-level students and	low-level students. That is to say, learners
in the high-level groups than the	low-level students. This result corresponded to the
were higher than the posttest scores of	low-level students, we can infer that all
small and large corpus and studying with	low-level students, which gave us understanding about
students did not perform better than ICL	low-level students with lower mean score. As
activity was measured with a group of	low-level students with the results indicating that
be due to the affective factor that	low-level students would be afraid to speak
to be better thought with deductive teaching	for low-level EFL students. A third reason
is a trade-off in using corpus	for low-level EFL students. When low-level
acquire a positive attitude toward corpus use	for low-level EFL students. In fact, getting
such as sentence-writing are more difficult	for low-level learners than other exercises and
computers in the classroom, and being suitable	for low-level learners. The effectiveness of the
appropriate only for advanced learners and not	for low-level learners. Hadley (2002), for example, claimed
tive discovery learning, the threshold hypothesis	for low level of learners, and capacity to
is seen that deductive teaching is indispensible	for low-level students. In this sense, the
. Firstly, this approach might not be suitable	for low-level students and even higher level
that will be conducted with participants from	different proficiency levels.
ttings, that would include more participants with	different proficiency levels, and that would use a
herefore, partial replications with larger-scale,	different proficiency levels and target items are needed
, and their learning outcomes across learners of	different proficiency levels. Another question is whether there
weaknesses of using corpus-based materials at	different proficiency levels can also be explored in
ly an individual activity; therefore, learners of	different proficiency levels could be learning at their
play important roles when group members have	different proficiency levels (e.g., Kim & McDonough 2008). Others (
milar. Next, conducting corpus-based studies with	different proficiency levels (elementary, pre-intermediate and high)
and implementing, this framework with students of	different proficiency levels. Furthermore, with the shift from

Figure 5. Sample concordances to the theme proficiency from the Conclusions corpus

The theme of **institutions** reflects the recently growing interest in expanding DDL applications to new learner populations and learning environments. The evidence is especially solid in regard to such keywords as *elementary*, *young*, *mobile*, each appearing in a range of 6 to 7 article Conclusions published in 2016-2019 (

APPENDIX **B**). The coding sheet however shows that calls for DDL extension outside tertiary education still remains a minority interest. Among the 477 studies that give some indication of the institutional context, just 9% are with younger learners, in schools (elementary, middle, junior high or high) or other pre-university courses, but the university constitutes the mass of participants, accounting for 85% overall. This may indicate that DDL is most appropriate at these levels, or that researchers simply follow suit, or, prosaically, that these are the participants that they have access to in their own teaching. Of these, 39% are undergraduates compared to just 16% of graduates, including doctoral students; a further 30% combine undergraduates and graduates or merely note that they were conducted at university. Only 7% are in other contexts, mainly *language institutes* (which may cover a multitude of situations), or on rare occasion with various types of professionals. There seems to be little change over time and no major differences between JCR100 and other papers. There is clearly much to be done to investigate how DDL may be used outside class, after the end of a course, or in non-university contexts.

Specialization comes up in other Conclusions keywords: *translation*, *translators*, *engineering*, *legal*, *law* (

APPENDIX **B**). Disciplinary or other specialization is indicated in 363 papers: 47% feature students majoring in languages, linguistics, translation, language teaching, or some other language-related degree: 19% are in other humanities and social sciences, 16% in STEM subjects (science, technology, engineering, medicine), and 18% mixed. These figures are fairly stable, although STEM subjects increased from 8% in the first two periods to 17% in the last two; the difference between JCR100 and other papers is negligible. However, DDL is often used for 'general' English purposes even in highly specialized disciplines; indeed, general language accounts for twice as many studies (66%) as LSP and LAP combined (15% and 16% respectively), with mixed studies relatively rare (about 2%, usually an LSP/LAP crossover); no notable differences were found over time or between JCR100 and other papers. This may seem unintuitive: since general language can be found in generic resources (dictionaries, textbooks, etc.), it might be thought that DDL is most appropriate for specific needs in narrow fields.

Geographic distribution is important for individual researchers, as highlighted by national keywords (e.g., *Thai*) calling for more work in their own country. Figure 6 shows the evolution by region in the 5 time periods. The suspicion that DDL is a largely European phenomenon (e.g., Hadley, 2002) may have some basis in the earliest periods; also that it has been slow to take off in North America. The last two periods since 2012 have seen large increases in the main regions, with sudden huge increases in Asia and the Middle East in particular, even though these regions may be associated more with a teacher-fronted knowledge-dispensing educational model rather than with the kind of autonomous discovery-based processes involved in DDL. This may imply that DDL as a largely inductive process comes as a welcome change from usual practice.

Figure 6. Evolution by region

The **target language** was overwhelmingly English in 432 papers (89% of those stated), of which 95 (22%) were JCR100. Eleven other languages are listed, notably including 6 in French, 5 German and 3 Italian in the JCR100 studies. The country of origin generally indicates the **first language**, with the exception of studies in inner-circle English-speaking countries where they tend to be mixed with English as the L1: only 45% of papers are foreign- as opposed to second-language learning there, as opposed to 85% overall.

Implementation themes

Interaction with DDL has remained stable, with the main procedures being introduced right from the start: keywords such as *concordancing*, *concordance*, *concordancer*, *collocations*, *collocate* change only slightly over time (

APPENDIX **B**). Nonetheless, researchers continuously call for more studies using these terms, i.e. more of the same. It is also clear that *DDL* has established its footing as the top key term in the latest batch of article Conclusions. Computer-based (hands-on) DDL has been by far the most popular theme here, attracting attention in at least 38% of Conclusions in each period, peaking in 2004-2007 at 54% (107 and 177 occurrences in the last two periods). The authors call for more hands-on DDL applications in which learners explore corpora *autonomously* or *on their own* (Figure 7). In contrast, *paper-based* was barely mentioned as a theme at the beginning and occurs in only about 9% of Conclusions in the three latest periods. This distribution is somewhat more balanced in what has actually been done as nearly two thirds of papers (63%) have participants working directly with a concordance and a quarter (25%) with prepared materials, usually paper but occasionally slide projections; 11% have used corpora integrated into a wider CALL program. However, 21% use two approaches, usually paper leading up to hands-on concordancing, but rarely comparing them directly. While participants are often encouraged to collaborate in pairs or small groups, with *collabo** as a search term appearing in 88 of the 489 papers, it is rarely a main objective (only 11 use it 5 times or more). No convincing patterns emerge over time or between JCR100 and other publications. While specific tasks were not coded, it seems likely that new tools provide new affordances, though the basics were in place decades ago and may not have changed that much.

corpus should not be used on their	own to help learners to come to terms
can be used freely and on their	own initiative by all students from beginner to
that can provide useful models for their	own writing.
of having anonymised examples of their students'	own writing to work on in class, in
learning and are able to choose their	own words to investigate and make their own
own words to investigate and make their	own discoveries. The major advantage of the DDL
student was able to work at her	own pace. The instructor found more time to
may work better with Thai students than	autonomous concordancing. Also, as the students themselves s
'language awareness' and empowers them to be	autonomous and responsible for their own learning. A
to be autonomous and responsible for their	own learning. A number of conclusions can be
is that teachers and learners build their	own text corpora that they need, as suggested
group discussion. For the first task, our	own corpus of Business texts was consulted as
tentative but hopefully useful light on students'	own evaluations of its value in L2 writing.
that they will embrace corpora on their	own, or with minimal guidance from the instructor.
cited in this study, cannot on its	own lead to any large-scale integration of
individual centre should have to create its	own resources. Learners' needs are sufficiently simil
earning environment which aims to promote learner	autonomy, could then become the focus for future
nguage-learning environment which favours learner	autonomy and discovery learning. It is, however, the

Figure 7. Sample concordances to the theme autonomy from the Conclusions corpus

Consultation, i.e. using corpora as reference tools, was suggested as a future direction in 25% of the earliest Conclusions (1989-2003) and in almost 50% in 2004-2007. Actual research has responded to these growing calls with the number of articles exploring this theme changing from 15% in 1989-2003 to about 30% in each subsequent period. In fact, just over a quarter (26%) overall use a corpus as a reference resource, compared to 42% looking at learning outcomes). It is noteworthy that the JCR100 papers are less focused on outcomes than the others (34% vs 44%), perhaps as pre/post-test designs are fairly obvious and easy to administer, and are relatively more interested in learners' behavior (32% vs 20%), which is more difficult to track. Other objects of study include the processes involved (23%) and, especially, the participants' attitudes and perceptions in both JCR100 and other papers (56%). Typically collected via

questionnaires or interviews as we saw earlier, this emic perspective often occurs as a secondary feature alongside the main focus and is almost *de rigueur*. With a large body of supportive research, the question now is less 'does it work' than 'how well does it work' with different learners with different needs in different contexts, using different resources, etc.

Guidance or training has consistently been mentioned as a theme in 22-26% of the Conclusions for each period with a spike of attention in 2004-2007 (more than 50%). Overwhelmingly, the authors have concluded that DDL requires *extensive*, *planned*, *ongoing*, *meticulous*, etc. teacher *guidance* (Figure 8). Although this did not figure as a category in our top-down coding sheet, a search for this theme in article abstracts yielded only 47 hits in total (in contrast to 182 hits in Conclusions), with only a handful of articles stating the amount of guidance needed for successful DDL as their research goal. This shows a disconnect between researchers' claims about the importance of and the need for more guidance and the small number of attempts to empirically support these claims.

alone to acquire the patterns without any	guidance, especially those who are used to the
design of a tutorial program with appropriate	guidance could be helpful. Furthermore poor reading profic
in this study suggest that, with careful	guidance (i.e., a teacher-tested worksheet) and
nvolved (N = 37), the role of corpus consultation	guidance needs to be further addressed taking into
et al. (2012), who found that despite corpus	guidance and training only a group of learners, 46
al introduction to concordance work and extensive	guidance in using concordancing strategies is recommended.
for the first time, effective and extensive	guidance from the instructor is of paramount importance.
nductive thinking strategies. Therefore extensive	guidance in using inductive learning strategies is highly
er to inductive learning methods. Thus, extensive	guidance in using inductive learning strategies is recomme
the instructor's full support and firm	guidance. Although these techniques could be included in
their own learning. They strongly long for	guidance to develop their analytical skills. Therefore, th
05; Hadley, 2002; Yeh et al., 2007). With further	guidance through the use of prompts in students'
tantly, during the training, teachers should give	guidance and observe learners closely in order to
hers might consider supporting learning by giving	guidance through 'leading questions', which might help lea
assist the instructor to offer students intensive	guidance on tasks and to examine the long-
assist the instructor to offer students intensive	guidance on tasks and to examine the long-
.'s (2011) study relied on explicit teacher-led	guidance on corpus consultation. The DDL students also
broadened toward direct corpus use with less	guidance adapted to the local context. Given the
the present study confirms. Teachers' meticulous	guidance and vigilant individualized feedback are likely t
corpora on their own, or with minimal	guidance from the instructor. In conclusion, it should
from authentic language data. Consequently, more	guidance should be offered by teachers if concordancing
. In this regard, teachers may need more	guidance and practice opportunities to use corpus-based
teaching, some admitted that they needed more	guidance on how to select appropriate language problems

Figure 8. Sample concordances to the guidance theme from the Conclusions corpus

Language skills as mentioned in article Conclusions have been changing from period to period. Syntax and oral language skills were more prominent in the earliest period (1989-2003), then attention shifted to pragmatics and discourse in 2004-2007, collocations, colligations, and lexical and prosodic targets in 2008-2011, and verbs in 2016-2019 (

APPENDIX **B**). Further, interest in chunks has become prominent in the two latest periods (keywords *chunks*, *formulaic*, *MWE*). While some of this keyness (e.g., pragmatics in 2004-2007) can be attributed to prominence of a few specific authors, other targets have attracted a wider interest (e.g., verbs being mentioned in Conclusions of 15 studies from 2016-2019). Despite these temporal differences in Conclusions, there is little obvious change over time in the collection as a whole. Language skills are mentioned as a focus in 125; of these, writing is far the most frequent (98), followed by translation and reading (17 each), with speaking (8) and listening (4) far behind. The figures are relatively stable over time, and there is little difference between JCR100 and other papers. Vocabulary, variously defined, is clearly a recurrent objective, listed in 123 papers, though it fades into lexicogrammar (including collocates, chunks, grammar-function words, error-correction, etc.), which occurs in 256 papers, with considerable overlap. Wider areas such as discourse, pragmatics and rhetorical functions, discourse markers and cohesion, etc. are highlighted in only 41 cases (nearly all non-JCR100), presumably as they are less amenable to the basic surface queries of corpus tools. This leaves 23 that are interested in corpus linguistics *per se* (e.g., how learners go about building or using a corpus, corpus literacy, etc.), and a diverse 18 which range from literary analysis and cultural awareness to systemic-functional linguistics, language awareness, interference, cognitive skills and critical thinking – though again, none of these are in the JCR100. While vocabulary and lexicogrammar are the most obvious focus, then, especially for writing, DDL has moved into wider areas of language.

Technological themes

Technology is an obvious candidate for change over time, and indeed some hits for this theme in the earliest period Conclusions clearly referred to now superannuated media, tools or corpora (e.g., *cards*, *PET*), while others became key in the latest period – possibilities like *mobile* or *telecollaborative*, and newer and currently developing corpora like COCA and FLAX (

APPENDIX **B**). Most data about technology, however, come from our top-down analysis of actual research.

In terms of **software or interface**, the BYU suite at English-Corpora.org is the most commonly used in 31% of all papers, though with a slight reduction in the final period. Other integrated interfaces include SketchEngine and the BNCweb in 11 cases each, VLC [1] in 9 and MICASE/MICUSP in 8; no others feature more than 5 times. Some older stand-alone software is tending to disappear: MicroConcord, MonoConc and ParaConc (in 5 or 6 papers overall) do not feature at all in 2016-2019; similarly, WordSmith Tools, despite appearing in 33 papers overall, is down from a high of 30% of papers in 2004-2007 to just under 2% in 2016-2019, perhaps due to the availability of free rivals that are arguably more user-friendly for pedagogical purposes. LexTutor has its ups and downs, being used a total of 23 times, while the most popular downloadable tool overall is AntConc with 43 recorded uses. This first appeared in 2002 but has no recorded use in our DDL collection until 2009. Other tools such as FLAX are starting to make an appearance (5 papers so far), though use of the researchers' own tools seems also to be in decline (20 cases, accounting for just 4% or 5% of papers in the last 3 periods). Other tools, though excellent for research, are virtually absent here (e.g. CQPweb). There are essentially no differences between JCR100 and other papers. This then is one area of real change over time as new technology is developed and made available, often linking corpora and other resources. For example, COCA has recently integrated links to outside resources in its interface, from dictionaries and automatic translation to videos, images, Google searches, and full texts.

Corpus is an unconvincing keyword in this context, with the specific corpora often but not always named – 354 times among the 391 papers that feature hands-on work. Overall, 27% used more than one corpus, a figure increasing over the 5 time periods from 8% to 28%. Self-compiled corpora were used as at least one source in 42% of studies, though the figure is decreasing (56% to 29%), presumably as reliable outside corpora become more readily available. In line with this, just a third of papers named the corpus in 1989-2003 compared to over 70% in the last two periods. A significant number (19%) use more or less specialized corpora, whether disciplinary or genre, particularly academic texts, and more among the JCR100 papers (24% vs 17% for the rest). Some of these are ready-made corpora or subcorpora (e.g. MICASE or COCA_academic), though a large number are compiled from research articles as an obvious and relatively 'clean' source of texts, often but not always for disciplinary academic writing. Learner corpora remain difficult to come by or create, with fewer than 10% of studies using them in any time period; even fewer require learners to create their own corpora, rarely going above 5%, and virtually none using graded corpora specially written or compiled for the target proficiency level (under 3%, and in the final period only). Corpora of literary texts make a small contribution (3-6%) but disappear entirely in the latest period. Manuals were a popular source in 1989-2007 (17%) but have never since risen above 3%. The single most popular corpus is the BNC in one form or another in 24% of papers, ahead of COCA at 20%. COCA was first made available in 2008 (see Davies, 2009) and by 2016-2019 is being used in nearly a third of all studies (32%), while the BNC has remained relatively steady at around 25-30% since 2008. Some of the older corpora are fading in popularity (the COBUILD / Bank of English and Brown families are down to under 3% and 2% respectively in the latest period), while others have yet to break on to the scene in a large way: MICASE/MICUSP and BASE / BAWE together account for just 4% of

[1] Here and in some other cases, the original links may have changed.

studies in 2016-2019. Despite their obvious appeal, the potential of corpora and tools such as WordAndPhrase or SkELL remain largely untapped here, reported in just 2 and 3 papers respectively. Surprisingly, perhaps, only 19 used the web directly as a 'corpus', hovering at around 5% for the entire span, despite anecdotal evidence that many learners are simply Googling words and phrases for answers to their language questions. Parallel corpora (often English paired with Chinese or Japanese) have been popular at times, featuring in 9% of papers, while multi-modal corpora are almost invisible – just 9 studies, with MICASE and Backbone the only two to occur more than once. New, larger, rigorous and specialized corpora are thus making an appearance, though there seems to be a dearth or underuse of substantially original types of corpus, especially spoken corpora, learner corpora, graded corpora, parallel corpora, and so on, all of which may be highly relevant for pedagogical purposes.

Of the 250 hands-on studies that note the **size** of at least one corpus in terms of tokens (as opposed to texts, sentences, *n*-grams, or characters), and disregarding the web, the smallest is 1.2k words of student texts (clearly not a 'corpus' in the usual sense, but amenable nonetheless to exploration with corpus tools and techniques), the largest 2.8bn; exactly half are under 100m words, very often self-compiled corpora. Figure 9 gives the percentage of corpora in different bands (≥1k, ≥10k, etc.) for the five time periods, where such information is explicitly given. There is a striking decline in popularity for smaller corpora over time, while corpora of 100m words or more take the lion's share in the last three periods; corpora of over 1bn words only make an appearance in 2016-2019. The mean corpus size increases apace, from 15m tokens in early days to over a quarter of a billion in the last period, skewed by the mega-corpora (the median and mode are both 100m). Though none of the JCR100 papers feature corpora under 10k tokens, the other papers have a higher mean corpus size (199.6m vs 142.6m tokens), partly as they flirt more with very large corpora (5 of the 6 uses), partly also due to the increasing numbers of such publications in later periods when large corpora are more readily available. If "there's no data like more data", this may not be true for pedagogical purposes, where both small and large have their own advantages.

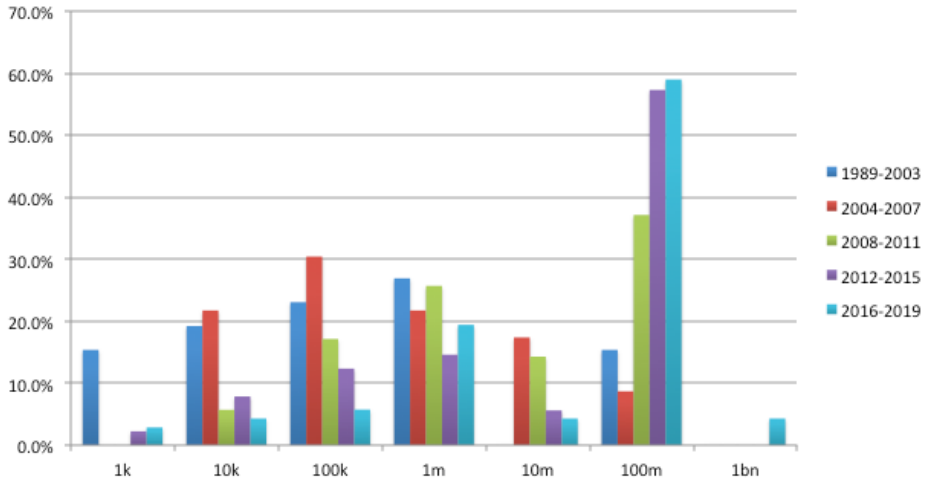

Figure 9. Corpus sizes reported over time

DISCUSSION AND OUTLOOK

Our overview has shown that the body of empirical DDL research has been growing rapidly and has by now developed into a fully-fledged CALL subfield. With nearly 200 empirical studies published in 2016-2019 alone and quantitative research being on the rise, Yao's (2019) statement that "empirical quantitative studies on the [DDL] topic have been limited" (p. 18) seems hardly justified. In what follows, we bring together our main findings and outline our suggestions for future research.

The lack of a theoretical-ideological base may have helped DDL to get as far as it has, but after 30 years it would seem necessary to push further and expect theories to drive continued development. Some researchers have tried to pin DDL to underlying theories such as constructivism (Cobb, 1999) or noticing and sociocultural theory (Flowerdew, 2015), but these attempt to justify the approach after the event. Pérez-Paredes (2019) and O'Keeffe (2020) are among those calling for research to be more theory-led in future. Of course, an approach survives on its merits and works because it works (or doesn't), regardless of theory; but this does leave us rather hungry for more.

Methodologically, researchers have been employing a variety of data collection and analysis instruments and methods. Nevertheless, although research quality was not our goal, we could not help but notice that statistical analyses have frequently been insufficiently robust and reporting practices non-standard. Next, we found persistent and growing calls for greater sample sizes and duration but no growth trend in actual studies. Further, the field has been growing more diverse in terms of geographical regions and first languages spoken by DDL participants. However, English still heavily dominates as the target language, although the number of studies exploring DDL for other languages has increased recently (Vyatkina, 2020a). Finally, research has overwhelmingly been conducted with university students in language-for-general-purposes classes. Our findings thus confirm Gillespie's (2020) wider observations about empirical CALL studies: "they are conducted over a few weeks or a semester at most, or infrequently over one academic year; are preliminary, and usually not followed up; and involve a small number of students (sometimes fewer than 10), more often at beginners or intermediate level, in a single institution" (p. 138). There are clear pragmatic reasons behind using convenience samples and short interventions at institutions where researchers teach. Nevertheless, there is clearly much to be done with longer, ecological settings to investigate how DDL may be used outside class, after the end of a course, or in non-university (professional, primary and secondary schools) contexts. Greater collaboration, both within and between institutions, might help to not only overcome these limitations but also explore contextual differences. More outreach to language teachers in form of open access corpora, accessibly written DDL guides (e.g., Poole, 2019; Vyatkina, 2020b), and training opportunities should also broaden the DDL research base.

Another area where we noticed little change in DDL research was types of learner interaction with technology. The development of corpora and corpus tools has been driven by research but also by pedagogical and other needs. These tools, in turn, have also been the driving force behind pedagogical innovation inasmuch as they allow previously time-consuming, difficult or impossible activities. The simple early tools were mainly used for the concordances they produced, quickly adding frequency lists (lemmatized or not), collocations, clusters (bundles or

n-grams), keywords, distributions, and so on. Each of these allows a new way to interact with the language. However, our overview shows that the main procedures – examining concordances and frequency information – have remained largely unchanged since the early research days. There are virtually no studies that compare different types of learner-corpus interaction or different types of corpora (specialized corpora, open-access corpora, web as corpus). Other underexplored possibilities include parallel corpora of translated texts in two or more languages, and multimodal corpora – written text aligned with sound and/or video recordings. Though these are among the most difficult to produce and access, they are potentially of tremendous interest for teaching and research. Some readily-available tools have managed to harvest online data to provide a DDL-like interface for multiple languages such as Linguee or DeepL, or videos such as PlayPhrase with TV and film extracts, or extracts from YouTube in YouGlish – despite the name, again available for different languages. Similarly, a promising area is combining corpus tools with other CALL practices such as multimedia glossing (e.g., Frankenberg-Garcia et al., 2019). We now have far greater choice of tools and corpora, which are faster, more aesthetic, and user-friendly (Godwin-Jones, 2017), and our hope is that they will attract the attention of future DDL researchers.

Overall, we found that much of research covers the same ground, with many studies confirming that DDL 'works'. Going forward, more nuanced, comparative research is needed. What works in different contexts and for different learner profiles? Are particular procedures, corpora or tools more appropriate in some settings than others? The field can now move away from experimental/control (or DDL/non-DDL) designs toward designs with two or more experimental groups.

Another area worth noting is the development of long-term, higher-level, non-language skills. Such skills as critical thinking, independent learning, and learner autonomy feature prominently in article Conclusions as alleged benefits from DDL, but we found virtually no direct exploration of these concepts as research objectives. Therefore, Boulton (2012b) observation still holds: "it is notable that much of the research to date focuses on targets that are easy to measure in a highly controlled experimental environment – short-term learning outcomes in vocabulary and lexico-grammar, as well as error-correction and Likert-scale questionnaires of learner attitudes, etc., [yet] there is a notable dearth of studies looking at the major advantages that are generally attributed to DDL" (p. 86). While operationally defining these abstract constructs certainly is not straightforward, we hope that future innovative research will pursue this challenging yet most promising direction that would go hand in hand with much needed theorization of DDL and its cross-fertilization with broader SLA and sister fields like educational psychology.

A final comment: although not specifically targeted in our analysis and largely impressionistic, it is our conviction from reading the papers that quantitative researchers should take advantage of contemporary multifactorial methods that allow for reliably exploring multiple variables simultaneously. More mixed-methods studies are needed that provide both rigorous quantitative accounts and ecologically valid qualitative analyses of DDL processes and outcomes. Studies that set out to test previous research findings should be clearly situated as replication studies, which have been lacking to date. Partial replication studies are especially welcome that would explore specific variables such as proficiency level or target language while keeping other design elements constant. Finally, a promising direction would be more integration of DDL and SLA.

Employing such SLA constructs as implicit and explicit knowledge, receptive and productive knowledge, knowledge breadth and depth, controlled and free production, and testing DDL effectiveness for the development of these knowledge types and language skills would undoubtedly bring both fields forward.

SUPPLEMENTARY MATERIALS

The entire coding sheet is available online as a simple Excel spreadsheet for others to sort and search at will. This contains the full references and abstracts which are usually freely available, but the rest of the corpus consists largely of copyright material which cannot be openly shared.

ACKNOWLEDGEMENTS

The authors would like to thank Tom Cobb and James Thomas, as the present study draws on previous collaboration and inspiration. Particular thanks go to A. Jakob Johnson who conducted most of the latest round of trawls, listing the papers which the authors then assessed for inclusion; he also conducted bottom-up manual coding for future research directions themes. The project was supported in part by the College of Liberal Arts and Sciences, University of Kansas.

REFERENCES

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

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

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

Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Goźdź-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 563-580). Peter Lang. https://doi.org/10.3726/978-3-653-04563-5

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

Boulton, A. (2012b). Computer corpora in language learning: DST approaches to research. *Mélanges Crapel, 33*, 79-91.

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

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

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

Burston, J., & Arispe, K. (2018). Looking for a needle in a haystack: CALL and advanced language proficiency. *CALICO Journal*, *35*(1), 77-102. https://doi.org/10.1558/cj.31594

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

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

Chambers, A., & Bax, S. (2006). Making CALL work: Towards normalisation. *System*, *34*(4), 465-479. https://doi.org/10.1016/j.system.2006.08.001

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

Cobb, T. (1999). Applying constructivism: A test for the learner as scientist. *Educational Technology Research & Development*, *47*(3), 15-31. https://doi.org/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 University Press. DOI 10.1017/CBO9781139764377.027

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-188. https://doi.org/10.1075/ijcl.14.2.02dav

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15-36). John Benjamins. https://doi.org/10.1075/scl.69.02flo

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

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

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

Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology, 21*(3), 9–27. http://llt.msu.edu/issues/october2017/emerging.pdf

Hadley, G. (2002). Sensing the winds of change: An introduction to data-driven learning. *RELC Journal*, *33*(2), 99-124. https://doi.org/10.1177/003368820203300205

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

Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, *10*, 14-34.

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

Lin, H. (2014). Establishing an empirical link between computer-mediated communication (CMC) and SLA: A meta-analysis of the research. *Language Learning & Technology*, *18*(3): 120-147. https://doi.org/10125/44387

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

Luo, Q., & Zhou, J. (2017). Data-driven learning in second language writing class: A survey of empirical studies. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(3), 182-196. https://doi.org/10.1093/applin/amy012

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

O'Keeffe, A. (2020). Data-driven learning: A call for a broader research gaze. *Language Teaching*. Advance online publication. https://doi.org/10.1017/S0261444820000245

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

Pham, M.T., Rajić, A., Greig, J.D., Sargeant, J.M., Papadopoulos, A., & McEwen, S.A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*, 371-385. https://doi.org/10.1002/jrsm.1123

Poole, R. (2019). *A guide to using corpora for English language learners*. Edinburgh University Press.

Schöpfel, J., & Prost, H. (2009). Le JCR facteur d'impact (IF) et le SCImago Journal Rank Indicator (SJR) des revues françaises: Une étude comparative [The JCR impact factor (IF) and the SCImago Journal Rank Indicator (SJR) for French journals: A comparative study]. *Psychologie Française*, *54*(4), 287-305. https://doi.org/10.1016/j.psfr.2009.07.002

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

Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37-62). John Benjamins.

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

Vyatkina, N. (Ed.). (2020b). *Incorporating corpora: Using corpora to teach German to English-speaking learners* [Online instructional materials]. Lawrence, KS: KU Open Language Resource Center. https://corpora.ku.edu

Yao, G. (2019). Vocabulary learning through data-driven learning in the context of Spanish as a foreign language. *Research in Corpus Linguistics*, *7,* 18-46. https://doi.org/10.32714/ricl.07.02

APPENDIX A

Selected keywords in the Conclusions corpus (vis-à-vis the full corpus) ranked by keyness. Keyword statistic: log-likelihood 4-term. Keyword statistic threshold: $p<.05$ (+Bonferroni). Keyword effect size measure: Dice coefficient. Keyword effect size threshold: all values.

Keyword	Rank	Frequency	Keyness	Effect size
DDL	15	1208	164.54	0.0092
proficiency	32	354	82.6	0.0028
longitudinal	42	47	53.4	0.0004
awareness	56	217	36.33	0.0017
guidance	62	93	35.48	0.0007
consultation	67	222	34.01	0.0017
EFL	77	328	31.7	0.0026
own	79	326	31.05	0.0025
collaboration	82	42	30.43	0.0003
integration	86	66	29.43	0.0005
autonomous	129	70	21.99	0.0006

APPENDIX B

Key n-grams in the Conclusions corpus (vis-à-vis the full corpus) clustered in 5 themes and ranked by keyness in each cluster. Keyword statistic: log-likelihood 4-term. Keyword statistic threshold: $p<.05$ (+Bonferroni). Keyword effect size measure: Dice coefficient. Keyword effect size threshold: all values.

Theme	Key n-gram	Rank	Freq.	Keyness	Effect size
Sample Size	number of participants	29	39	90.54	0.0012
	small sample size	76	19	48.86	0.0006
	small number of	112	26	38.69	0.0008
	larger number of	141	17	35.01	0.0005
	the sample size	160	14	33.54	0.0004
	a larger number	167	14	32.88	0.0004
	a larger number of	168	14	32.88	0.0004
	small number of participants	180	12	31.42	0.0004
	a larger sample	232	10	28.1	0.0003
	larger sample size	272	8	26.33	0.0002
	limited number of	281	19	25.88	0.0006
	a larger scale	296	8	25.2	0.0002
	the small sample size	304	10	24.94	0.0003
	the small sample	308	11	24.8	0.0003
	a larger sample size	327	7	24.11	0.0002
	sample size and	365	8	23.2	0.0002
	number of subjects	421	8	21.47	0.0002
Duration	the long term	70	24	49.93	0.0007
	a longer period	156	13	33.91	0.0004
	a long term	206	12	29.23	0.0004
	over a longer	213	10	29	0.0003
	more time to	275	12	26.08	0.0004
	a short term	320	8	24.16	0.0002
	the long term effects	325	8	24.16	0.0002
	a longitudinal study	335	9	23.96	0.0003
	long term effects of	420	8	21.47	0.0002
Autonomy	their own learning	96	28	43.11	0.0009
	on their own	131	45	35.65	0.0014
	for their own	197	23	30.02	0.0007
Consultation	of corpus consultation	350	31	23.54	0.0009
	that corpus consultation	385	14	22.48	0.0004
Paper-based DDL	paper based DDL	177	32	31.66	0.001
	paper based materials	398	16	22.03	0.0005
	paper based DDL materials	399	12	22	0.0004
Populations	pre service teachers	166	18	32.99	0.0006
	the EFL learners	228	14	28.24	0.0004
	to the EFL	326	8	24.16	0.0002
Proficiency	different proficiency levels	196	16	30.1	0.0005
	of language proficiency	333	18	24.08	0.0006
	low level students	383	14	22.86	0.0004
	for low level	435	9	21.09	0.0003

APPENDIX C

Keywords in the temporal subcorpora of the Conclusions corpus (vis-à-vis all other temporal subcorpora) ranked by keyness. Fr=Frequency; Key=Keyness; ES=Effect Size. Keyword statistic: log-likelihood 4-term. Keyword statistic threshold: $p<0.05$ (+Bonferroni). Keyword effect size measure: Dice coefficient. Keyword effect size threshold: all values.

1989-2003				2004-2007				2008-2011				2012-2015				2016-2019			
Keyword	**Fr**	**Key**	**ES**	**Keyword**	**Fr**	**Key**	**ES**	**Keyword**	**Fr**	**Key**	**ES**	**Keyword**	**Fr**	**Key**	**ES**	**Keyword**	**Fr**	**Key**	**ES**
subject	37	96.71	0.0035	project	42	62.53	0.0028	concordancer	62	48.22	0.0035	chunks	44	54.94	0.0013	DDL	732	211.73	0.0143
CLUES/clues	22	91.34	0.0021	concordancing	79	52.75	0.0052	SkE	11	43.27	0.0006	collocations	167	44.47	0.0051	participants	257	77.05	0.0051
strategies	39	43.66	0.0037	text	58	51.91	0.0039	prompts	15	38.02	0.0008	intelligences	15	40.51	0.0005	workshop	43	64.61	0.0008
concordancing	60	43.42	0.0056	template	11	47.05	0.0007	materializations	9	35.4	0.0005	lexical	126	31.88	0.0038	activities	259	59.58	0.0051
texts	47	39.76	0.0044	word	73	45.02	0.0048	WBC	8	31.47	0.0005	translation	54	30.97	0.0016	formulaic	40	50.7	0.0008
student	54	37.73	0.005	telecollaborative	11	40.42	0.0007	social	13	22.66	0.0007	usages	19	29.75	0.0006	COCA	60	50.38	0.0012
PET	6	29.77	0.0006	intervention	27	34.41	0.0018	entries	8	22.07	0.0005	translators	22	29.42	0.0007	lines	116	46.06	0.0023
methodology	23	26.64	0.0022	learner	80	29.3	0.0053	MICASE	7	21.81	0.0004	students	773	23.4	0.0227	elementary	28	43.72	0.0006
cards	5	24.81	0.0005	templates	8	28.19	0.0005	scaffolding	19	21.29	0.0011	engineering	15	23.35	0.0005	reporting	22	40.38	0.0004
compartment	5	24.81	0.0005	occurrences	13	26.62	0.0009	triadic	5	19.67	0.0003	corpora	290	22.86	0.0087	collocate	21	38.55	0.0004
oral	17	24.74	0.0016	ConcApp	6	25.66	0.0004	prepared	18	19.11	0.001	errors	75	22.02	0.0023	mobile	20	36.71	0.0004
syntax	11	23.75	0.001	hypertext	6	25.66	0.0004					colligations	8	21.61	0.0002	young	29	36.38	0.0006
transfer	13	23.75	0.0012	pragmatic	16	24.94	0.0011					encoding	8	21.61	0.0002	FLAX	17	31.21	0.0003
work	53	22.56	0.0049	derived	13	24	0.0009					prosodic	8	21.61	0.0002	verbs	63	25.75	0.0012
specific	40	21.77	0.0037	words	82	22.47	0.0054					corpus	756	21.55	0.0222	students’	14	25.7	0.0003
way	44	21.27	0.0041	scholarly	8	22.08	0.0005					legal	32	21.26	0.001	platform	20	25.35	0.0004
treasure	4	19.85	0.0004	news	11	21.78	0.0007					law	14	21.14	0.0004	findings	211	23.97	0.0042
concordance	68	18.97	0.0063	Telekorp	5	21.39	0.0003					examples	86	21.13	0.0026	thesis	19	23.71	0.0004
				web	30	20.26	0.002					tags	11	19.75	0.0003	experimental	103	23.22	0.002
				properties	6	20.17	0.0004									collaboration	32	22.83	0.0006
				week	13	19.72	0.0009									studies	217	22.71	0.0043
				discourse	31	19.61	0.0021									scores	47	22.32	0.0009
				electronic	12	18.95	0.0008									study	791	22.15	0.0154
																Thai	22	22.14	0.0004
																satisfaction	12	22.03	0.0002
																lectures	25	21.55	0.0005
																informed	49	20.78	0.001
																enhancement	26	20.76	0.0005
																questionnaire	54	20.22	0.0011
																MWE	11	20.19	0.0002