ReCALL (2023), 35: 3, 339–355
doi:10.1017/S0958344022000222

RESEARCH ARTICLE

Charting the landscape of data-driven learning using a bibliometric analysis

Jihua Dong
Shandong University, China (dongjihua@sdu.edu.cn)

Yanan Zhao
Shandong University, China (202220253@mail.sdu.edu.cn)

Louisa Buckingham
The University of Auckland, New Zealand (l.buckingham@auckland.ac.nz)

Abstract

This study employs a bibliometric approach to analyse common research themes, high-impact publications and research venues, identify the most recent transformative research, and map the developmental stages of data-driven learning (DDL) since its genesis. A dataset of 126 articles and 3,297 cited references (1994–2021) retrieved from the Web of Science was analysed using CiteSpace 6.1.R2. The analysis uncovered the principal research themes and high-impact publications, and the most recent transformative research in the DDL field. The following evolutionary stages of DDL were determined based on Shneider's (2009) scientific model and the timeline generated by CiteSpace, namely, the conceptualising stage (1980s–1998), the maturing stage (1998–2011), and the expansion stage (2011–now), with Stage 4 just emerging. Finally, the analysis discerned potential future research directions, including the implementation of DDL in larger-scale classroom practice and the role of variables in DDL.

Keywords: data-driven learning; co-citation analysis; structural variation analysis; bibliometric analysis

1. Introduction

Data-driven learning (DDL) involves researcher-like inductive explorations of language use, and was described by Johns (1991) as "the attempt to cut out the middleman as far as possible and to give the learner direct access to the data" (p. 30). It has been a long-standing focus of language learning research and has been attested to be useful in guiding language learners to "explore language corpora and come to their own conclusions" (Boulton, 2011: 575). By providing second language (L2) learners with a large amount of "naturally-occurring language" (Boulton, 2009a: 37), the DDL approach entails a range of activities where learners are not taught by traditional, often teacher-centred, deductive approaches, but are encouraged to explore corpus data independently and identify patterns of language use. This can enable learners to discover language patterns through authentic language data (Boulton & Cobb, 2017), which can effectively enhance students' learning motivation, engagement and autonomy (Gilquin & Granger, 2010). A number of review studies on DDL research have been conducted, which have contributed valuable insights into the development of DDL studies. Nevertheless, the high-impact publications, main research venues and developmental stages in the DDL field still remain to be explored.

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

With a view to addressing these omissions, this study employs bibliometric analysis to map the studies on DDL over the period 1994–2021 in terms of the common research themes, high-impact publications, main research venues, the developmental stages, and the latest transformative publications in this field. Bibliometric analysis is a statistical analysis of datasets comprising literature published within a specific time period (Pritchard & Wittig, 1981). The following research questions (RQs) guide our analysis:

1. What are the common research themes, high-impact publications and main research venues in the DDL field?
2. Using Shneider's (2009) model of evolutionary stages, what developmental stages can be identified in the DDL field over time?
3. What publications can be identified by structural variation analysis (SVA) as having potentially high impact?

2. Literature review

2.1 Reviews in DDL

Numerous studies have endeavoured to synthesize DDL research and have drawn attention to specific topics in the field. For instance, Chambers (2007) investigated 12 empirical studies from the 1990s onwards to explore learners' corpus consultation and stressed the importance of evidence for assessing the effectiveness of DDL. Boulton (2008) analysed 39 DDL papers and identified a primary focus on learners' interactions with and attitudes toward DDL. Boulton (2010b) further surveyed 27 empirical research studies on students' learning outcomes and found a scarcity of research investigating variables such as learners' motivation and attitudes. Also, Boulton and Tyne (2013), in their critical review, pointed out the need for classroom practice and collaboration between researchers and practical instructors. From a chronological view, Boulton (2017) offered a research timeline throughout the existence of DDL.

Another notable focus has been on a particular aspect of DDL. Yoon (2011) examined the use of DDL in writing classes and concluded that concordancing exercises are useful for L2 writers. Boulton (2012) reviewed 20 empirical studies of corpus use in English for specific purposes (ESP) and found that corpora can be used as effective learning tools and reference resources. In a review of 18 empirical studies of DDL in L2 writing, Luo and Zhou (2017) identified the great potential of DDL activities in L2 writing classes, but they also found that the use of corpora was not superior to traditional tools when used as a reference tool. Chen and Flowerdew (2018) synthesized 37 empirical studies in academic writing in terms of their main application and called for more studies to expand this field. The third strand of systematic reviews involves the construction of a corpus comprising DDL studies and the identification of common themes by using software or corpus-based analysis. Pérez-Paredes (2022) examined the utilisation of DDL by compiling journal articles from 2011 to 2015 into a corpus, and found that the topics of syllabus integration and teacher training are rarely discussed in DDL. In the latest study, Boulton and Vyatkina (2021) conducted a large-scale and systematic corpus-based analysis of DDL studies and identified the publication scope, research themes, and future directions of DDL research.

Meta-analyses of DDL research have also been undertaken. For instance, Mizumoto and Chujo's (2015) examination of the effectiveness of DDL for learning lexico-grammatical items provided support for the use of DDL for vocabulary acquisition. The meta-analyses by Boulton and Cobb (2017) and Lee, Warschauer and Lee (2019) were identified as prominent publications in our bibliometric analysis and are discussed in detail in Section 4.4.

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

2.2 Bibliometrics

Bibliometrics is conceptualised as "the application of mathematical and statistical methods to books and other media of communication" (Pritchard, 1969: 348). This quantitative approach employs bibliometric data from scientific databases such as Web of Science (WoS) and Scopus and has been used to identify research networks, research themes and research trends (Lei & Liu, 2019). The use of scientific databases enables a comprehensive, structured and balanced coverage of literature (Birkle, Pendlebury, Schnell & Adams, 2020) by employing inclusive bibliometric data (e.g. number of publications and citations, occurrences of keywords, and references). This makes it possible to evaluate the impact of journals, authors, and publications and the productivity of institutions (Lei & Liu, 2019). More recent bibliometric analyses have used customised software such as CiteSpace and VOSviewer to construct and visualise bibliometric maps. CiteSpace is an information visualisation software that can analyse and visualise trends and patterns in a field, and can facilitate various analyses, such as co-citation analysis, SVA, and collaboration networks (Chen, 2012).

Previous enquiries in this line have applied bibliometric analysis to map out the development of research fields, such as applied linguistics or computer-assisted language learning (Chen, Zou, Xie & Su, 2021; Jung, 2005; Liu & Zhang, 2021), L2 vocabulary acquisition (Meara, 2012), corpus linguistics (Park & Nam, 2017), multilingualism (Lin & Lei, 2020), English for academic purposes (EAP) (Hyland & Jiang, 2021a), and ESP (Hyland & Jiang, 2021b; Liu & Hu, 2021). A recent application of this approach in DDL is He and Wei (2019), who investigated the role of corpora in EAP research from 2009 to 2018.

2.3 Evolutionary model

Shneider's (2009) four-stage evolutionary model was employed in this study to trace the development of DDL research. According to Shneider (2009), the evolution of a scientific discipline can be mapped into four stages. The first stage is primarily concerned with introducing language (i.e. terms and concepts) to a field. The second stage tends to display a primary focus on the principal techniques and tools. The third stage focuses on broadening the existing focus of interest to new areas. Research at stage four typically involves codifying knowledge through reflective reviews, meta-analyses or textbook publications. Shneider's (2009) four-stage model has been employed in bibliometric analyses of various disciplines, including information science, engineering, and ESP (e.g. Chen, 2017; Liu & Hu, 2021). The review by Liu and Hu (2021) revealed three evolutionary stages of ESP, namely the "initial conceptualising stage" (1970s–1990s), "the maturing stage" (1990s–2000s), and "the flourishing stage" (2000s–).

3. Methodology

3.1 Data collection

The dataset was retrieved from the WoS core collection database. The search provided 412 articles (1994–2021), which were narrowed down to 126 by excluding papers irrelevant to DDL. The earliest publication in the dataset appeared in 1994, thus considered the starting point. A flowchart of detailed procedures and relevant descriptions is provided in Appendix A (available in supplementary material). The inclusion of the most recent studies (November 2021) in the dataset enabled an up-to-date analysis of DDL research to capture the latest citations.

Although our initial search in WoS focused primarily on research articles (citing papers), CiteSpace captures the cited papers in the references, which enables the inclusion of a wider range of publication types (e.g. dissertations, theses, book chapters, and meta-analyses). This thereby broadens the scope of the study and enables us to identify prominent or frequently co-cited publications from a wide range of document types in this field.

3.2 Co-citation analysis and SVA

Co-citation analysis is a common bibliometric approach used to measure the topic similarity between two or more documents. Co-citation is measured by “the frequency with which two or more publications are referenced in another publication” (Aryadoust, Zakaria, Lim & Chen, 2020: 2). If two documents are cited in one article, they are regarded as co-cited documents; the more co-citations two documents have, the greater their semantic relatedness. Highly co-cited pairs of publications grouped into the same cluster can display commonalities in research themes (Chen, Ibekwe-SanJuan & Hou, 2010). Co-citation counts can be used to generate a scientific map of knowledge in a field, which consists of clusters of co-cited publications. The identification of key research themes using co-citation contributes to understanding the evolution of common research themes in a field (Chen, 2017).

Given that co-citation analysis relies heavily on citation counts, it may be intuitively presumed that citation counts would be affected by factors such as early online publication and open access. We checked the dataset in this study and identified eight preprints out of 126 articles. According to Craig, Plume, McVeigh, Pringle and Amin (2007), the effect caused by the differing duration “diminishes with larger counting intervals” (p. 9), thus the influence of preprints on citation is marginal. Also, the co-citation analysis used in this study is measured by calculating the co-citation of two references. On the one hand, early access publications in citing articles do not influence the results, as WoS combines early access and published papers into one record. On the other hand, if one preprint or open-access article is highly cited, it does not influence the prominence of the theme identified in this study unless the text is repeatedly co-cited with another text. Even if a particular preprint or open-access article is cited highly in conjunction with other articles on the same topic, the validity of this co-citation analysis still remains unaffected as long as the cited text is related to the theme of the cluster scrutinised.

SVA is a predictive model operationalised as a function in CiteSpace (Chen, 2012), which aims to determine the transformative potential of a new publication in a field (Sebastian & Chen, 2021). The variation can be quantified based on information in the publication, mainly cited references. The higher the degree of variation, the more transformative a publication will be. Unlike co-citation analysis, which requires the information of accumulated co-citation counts, SVA is advantageous in assessing the transformative potential of ideas conveyed even in a very recent publication. This contributes to mitigating the chronological bias inherent in co-citation. By identifying studies with transformative potential, SVA can serve as a good indicator of potentially high-impact publications (regardless of their publication dates), and can thereby signal the direction of future research in a field. This methodological approach has been attested to be effective in identifying studies of high transformative potential, such as Nobel Prize-winning publications (Sebastian & Chen, 2021).

3.3 Network generation and analysis

The following parameters were used to address the research questions raised in this study. First, the modularity (Q) index and average silhouette score were adopted to measure the quality of the network, following Chen *et al.* (2010). The modularity score determines the clearness of boundaries between each pair of clusters, and high modularity scores signal the decomposition of recognisable clusters. The average silhouette is used to determine the quality of a clustering structure, and high average silhouette scores indicate the high reliability of the clusters in this study. When addressing RQ1, which concerns the common research themes and high-impact publications, the following three metrics, sigma (Σ), betweenness centrality, and burst, were used. Sigma, a measure of a publication’s novelty, was mainly used to identify prominent publications. The distribution of co-citations from the dataset and the number of prominent publications in different journals were used to identify the main research venues. To address RQ2, we then mapped these clusters onto

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

the evolutionary stages of a discipline identified in Shneider's (2009) model based on the defined qualities for each stage, including time frames, interrelations, and the embodiment of the characteristics specified in this framework. In order to find transformative publications in co-citation networks (RQ3), this study used centrality divergence (C_{KL}) and the harmonic mean (H) scores. The detailed descriptions of the metrics in addressing the RQs are available in Appendix A, and the properties set in the analysis as well as screenshots of CiteSpace are elaborated in Appendix B (refer to supplementary material). Apart from the automatic analysis using CiteSpace, a manual analysis of the labels based on a close reading of the data source and the automatic labels in CiteSpace was conducted by two of the authors. To ensure the consistency of the coding, Cohen's kappa was employed and the coefficient was found to be 0.93, indicating a high agreement in the labelling. Inconsistencies were resolved in a follow-up discussion.

4. Results and discussion

This section first presents the findings of RQ1 and RQ2. As the common research themes and high-impact publications are embedded in the evolutionary stages, this study integrates the first two research questions, namely common research themes, high-impact publications and research venues (RQ1) and the developmental stages (RQ2) in Sections 4.1–4.4. This is then followed by a detailed account of the latest transformative research in DDL (RQ3).

4.1 Baseline network interpretation

A network of 469 co-cited references and 1,793 co-citation links was created by CiteSpace from our bibliometric dataset. The modularity score of the network was 0.79, indicating clear boundaries between each pair of clusters; the high quality of clustering configuration was attested by the high silhouette score of 0.92 (see Figure A3 in Appendix A). Forty-five clusters were identified automatically, 11 of which contained more than 10 studies, and were thus worthy of further investigation.

Figure 1 presents the timeline view of clusters from a diachronic perspective of DDL. The time span of each main cluster is presented by separate horizontal lines. Each cluster is arranged horizontally with the direction of time from left to right. Clusters are sequenced in vertical order by size. When starting from the top of Figure 1 and moving down line by line, we can see the co-cited references in the main clusters. The larger the tree ring is, the more highly co-cited the publication is. Coloured lines represent the co-citation links between each pair of publications. A detailed illustration of publications in Clusters #0–#2 is presented in Figure C1 in Appendix C (presented in supplementary material).

Table 1 displays these main clusters, sequenced by the number of co-cited publications in each cluster, from the largest, Cluster #0, to the smallest, Cluster #10, as well as the oldest, Cluster #17. As can be seen, the mean year of publication varied from 2003 (Cluster #3) to 2020 (Clusters #5 and #8). The time period of each cluster's activeness was determined by considering the publishing years of all studies in each cluster. The largest cluster, Cluster #0, labelled "Effectiveness of DDL", was active for 10 years (2011–2021), and involved 81 co-cited publications. One prominent publication identified in this cluster is Boulton and Cobb (2017), a meta-analysis examining the effectiveness of DDL (elaborated in Section 4.4.1). The two smallest clusters, #9 and #10, "Corpus-based materials in pedagogy" and "Discipline-specific corpora", each comprised 12 studies. Examples of these studies include Frankenberg-Garcia (2012) in Cluster #9, which re-tested the benefits of corpus-based examples for learners' comprehension and the capacity of error correction; and Borja (2007) in Cluster #10, which provided an overview of translation-specific corpora in Spain for translators and translating researchers. The oldest major cluster is Cluster #3, "Teacher education in DDL", which comprised 38 publications. The most recent clusters, #5 ("Pedagogical implications of DDL") and #8 ("Language teachers' lesson

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Table 1. Major clusters of co-cited references

Cluster ID	Size	Silhouette	Mean (Year)	Time period	Label
0	81	0.894	2017	2011–2021	Effectiveness of DDL
1	46	0.862	2012	2005–2014	Learners' interaction with DDL
2	45	0.966	2009	2004–2008	Critical evaluation of DDL in classrooms
3	38	1	2003	1998–2007	Teacher education in DDL
4	38	0.905	2011	2005–2013	Classroom practice
5	33	0.945	2020	2014–2021	Pedagogical implications of DDL
6	30	0.994	2005	2000–2006	Early attempts in DDL
7	25	0.958	2015	2010–2015	Variables affecting DDL
8	18	0.966	2020	2016–2021	Language teachers' lesson planning
9	12	0.989	2013	2008–2012	Corpus-based materials in pedagogy
10	12	1	2009	2003–2011	Discipline-specific corpora

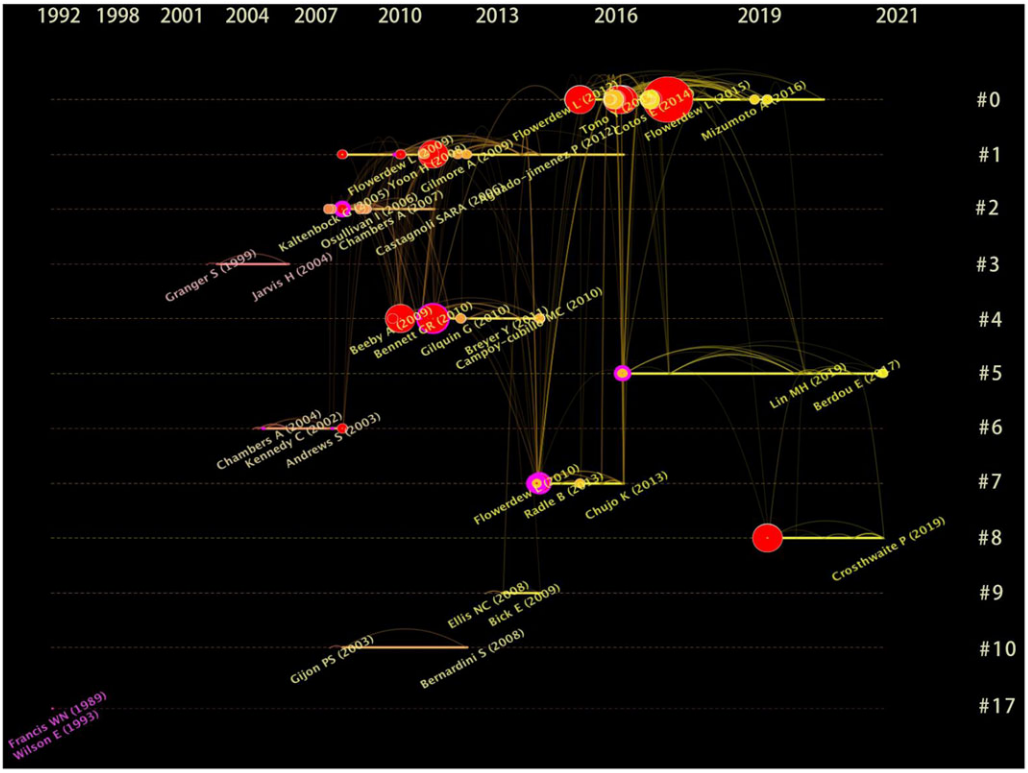

Figure 1. The timeline view of the network

planning"), have a mean publication year of 2020 and are still evolving in 2021 (as indicated by the co-cited publications in 2021 of this cluster). Cluster #5 included a meta-analysis by Cobb and Boulton (2015) focusing on the application of DDL in classrooms, and the review by Boulton (2017) on the explicit use of corpora in L2 learning and teaching. A representative publication

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Table 2. Citations bursts until 2021

Author	Year	Strength	Begin
Daskalovska N	2015	3.98	2016
Boulton A & Cobb T	2017	8.93	2017
Vyatkina N	2016	3.62	2017
Lee H, Warschauer M & Lee J H	2019	5.12	2019

in Cluster #8 is Zareva (2017), which surveyed L2 teachers' attitudes towards the application of DDL in teaching grammar. It is interesting to note that the oldest cluster, Cluster #17, contained the earliest co-cited publication, the Brown Corpus of American English (Francis & Kučera, 1989), albeit this cluster containing only four studies. Also of note, more than one cluster may be active at a time. For instance, Clusters #0, #4 and #10 were all active in 2011, which indicates that a variety of themes are valued during the same time period. This overlap in the time frames of clusters can be explained by the non-linear development of scientific fields, in which more than one prominent research topic emerges simultaneously. Also, many studies combine several research themes; for example, Boulton (2010a) in Cluster #4 examined both the effectiveness of DDL and low-level learners' attitudes.

Table 2 displays four publications with the most recent bursts (until 2021). Typically, there is a gap between the publishing year and the burst year, which is called the post-publication lag. Taking Daskalovska (2015) as an example, the starting year of the citation burst is 2016, one year after the publishing date. In some cases, however, the two years may coincide, such as Boulton and Cobb (2017) and Lee *et al.* (2019).

Based on the research focus of each cluster and Shneider's (2009) four-stage model, we identified the following three major stages in the development of the DDL field (RQ2): the conceptualising stage (1980s–1998), marked by the establishment of a new research object; the maturing stage (1998–2011), characterised by the development of research techniques and methods; and the expansion stage (2011–now), which features the application of instruments in new research domains and addresses new research questions, and the recent emergence of some features of Stage 4. But the dividing point between evolutionary stages is not clear-cut, and temporal overlap between adjacent stages may occur. For instance, the transition year of 2011 featured a theme focusing on the techniques and applications of those techniques. This can be explained by the inherently non-linear development of disciplines.

4.2 Stage 1: The conceptualising stage (1980s–1998)

Like other scientific disciplines, the feature of the first stage in DDL research is represented as the introduction of "new objects and phenomena" to signal the emergence of a certain discipline (Shneider, 2009: 217). Typical for the first stage is the coinage of new terminology to describe the subjects in a field. The term "data-driven learning" dates from Johns (1991), while prior to this various other terms had been used, including classroom concordancing, the microcomputer-based approach to foreign language learning (Johns, 1988), concordancing (Bloch, 2009) and corpus-based learning (Cobb & Boulton, 2015). Johns (1991) proposed that language learners use concordancers to explore authentic language data. Other publications from this first stage include Murphy (1996), which reported the use of DDL to assist in vocabulary learning, and Kita and Ogata (1997), which reported the use of DDL for the acquisition of collocation knowledge. However, the literature in Stage 1 was not identified in the co-citation network, which can be explained by the low level of co-citations of articles from this stage. According to Shneider (2009), first-stage research, while usually creative and inventive,

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

often possesses methodological weaknesses or inaccuracies, and is thus usually less cited than studies in the subsequent stages.

4.3 Stage 2: The maturing stage (1998–2011)

The most significant feature of Stage 2 concerns the creation of "a toolbox of methods and techniques" (Shneider, 2009: 217). The most representative clusters in this stage are #1, #2, #6, #9 and #10 (see Table 1), and are characterised by tool-centred publications. An additional significant feature entails a deeper analysis of the field. Part of Clusters #1, #2, #4, #6 and #7 are dominated by learner-centred publications, which reflect primary investigations on the effectiveness of DDL and its implementation. The primary focus on the implementation of DDL in this stage complies with the characteristics of Stage 2 defined by Shneider (2009).

4.3.1 Tool-centred research

The most salient theme in Stage 2 is concerned with tool-centred publications, which focus on developing specific types of tools and techniques, including software and corpora, for different groups of learners. Clusters #1, #2, #6, #9 and #10 (see Table 1) are representative of such themes. Of these, Cluster #1 is the largest, with 46 publications. One prominent publication in Cluster #1 is Bloch (2009) (Σ: 1.07), which designed the interface for a web-based concordancing program for academic writing. Although no high-impact publications were identified in the remaining clusters, the manual analysis revealed the dominance of tool-centred approaches to language teaching and learning. Anthony (2004) in Cluster #2 focused on the update of the corpus toolkit, AntConc, to assist in corpus building and analysis. Davies (2008), with the Corpus of Contemporary American English (COCA) in Cluster #9, and Burnard (2004), with the BNC Baby Corpus in Cluster #6, were identified as two important corpora. This is in line with previous findings of Park and Nam (2017), who found that COCA is the most cited in DDL. However, this work was not identified as a prominent publication here, possibly due to the relatively low co-citation with other studies on a similar theme, as it is possible that researchers may draw on a single tool or source in their empirical analyses. Other publications involve the indexing system by Köhler, Philippi, Specht and Rüegg (2006) and the accuracy of part-of-speech tagging in corpora by Coden, Pakhomov, Ando, Duffy and Chute (2005) in Cluster #10, and the use of the English Interview Corpus in language teaching by Braun (2005) in Cluster #2.

4.3.2 Learner-centred research

Another notable theme in Stage 2 is learner-centred research, as shown in Cluster #1. The principal focus of this cluster is the introduction of DDL in language learning (e.g. Kennedy & Miceli, 2010) and the role of corpus consultation in language learning (e.g. O'Sullivan, 2007). Kennedy and Miceli (2010), a prominent publication (Σ: 1.16), evaluated corpora as an aid to creative writing among intermediate-level language learners. O'Sullivan (2007), the second prominent publication (Σ: 1.15), investigated the role of corpus consultation in process-oriented learning. The third prominent publication conducted by Vannestål and Lindquist (2007) (Σ: 1.08) integrated the use of corpora into university English grammar courses.

Three prominent publications were identified in Cluster #4, all authored by Boulton (2010a, 2009a, 2009b). Boulton (2010a), with the highest sigma value of 2.59 in this stage, tested the assumption that DDL is unsuitable for low-level learners. The study demonstrated the effectiveness of paper-based concordance materials (teacher prepared) for low-level learners, thereby eliminating the cognitive burden presented by the use of software and computers. Boulton (2009a) (Σ: 1.27) provided evidence for the suitability of DDL among low-level learners, and Boulton (2009b) (Σ: 1.14) attempted to popularise DDL in language learning classrooms.

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Table 3. High-impact effectiveness-centred publications

Burst	Centrality	Sigma (Σ)	Author	Year	Cluster ID
8.93	0.04	1.42	Boulton A & Cobb T	2017	0
5.12	0.07	1.4	Lee H, Warschauer M & Lee J H	2019	8
4.52	0.06	1.31	Smart J	2014	0
4.17	0.07	1.31	Huang ZP	2014	0
3.98	0.03	1.14	Daskalovska N	2015	0
3.62	0.02	1.08	Vyatkina N	2016	0
0.00	1.00	1.00	Frankenberg-Garcia A	2014	0

The theme of learner-centred studies is also evident in Clusters #2 and #6. Chambers (2007), a prominent publication (Σ: 2.08), was an early attempt to synthesize DDL studies, while Cresswell (2007) (Σ: 1.15) found that learning styles affect learning outcomes, both of which are book chapters. Similarly, Chambers (2005), a prominent publication in Cluster #6 (Σ: 1.22), reported that individual differences (like learning styles and motivation) influence the success of DDL activities.

Cluster #7 also displays a primary focus on the learner-centred theme. Yoon (2011) has a strong citation burst (Σ: 1.45), and reviews 12 empirical studies that focus on the effectiveness and evaluation of DDL for L2 writing. Yoon noted that learners' acquisition of linguistic knowledge in writing and their autonomy can be facilitated by DDL, and pointed out the importance of studies focusing on teacher training and classroom implementations, as well as variables that affect learners' behaviours and learning outcomes.

4.4 Stage 3: The expansion stage (2011–now)

Conforming to Shneider's (2009) four-stage model, studies in the third stage tend to apply the methods and techniques developed in the second stage to address new problems in different domains, such as speaking competence in Cluster #0 (Geluso & Yamaguchi, 2014). Thus, Stage 3 represents the theme of "expansion" in this field. Studies in this stage primarily focused on the application of DDL to a broader range of domains. The expansion stage included part of Clusters #1, #4, #7 and #9, plus the intact Clusters #0, #5 and #8. The emergence of themes such as "variables affecting DDL", in Cluster #7, and "language teachers' lesson planning", in Cluster #8, is illustrative of the focus on new subjects and phenomena. Two main focal points of third-stage publications on DDL were identified, namely, effectiveness-centred and pedagogy-centred research.

4.4.1 Effectiveness-centred publications

Effectiveness-centred publications concentrated on the impact of DDL on learning outcomes, which are primarily determined by quantitative methods such as tests. Table 3 presents the studies with a strong focus on the effectiveness of DDL for learning collocations. As can be seen, two main clusters contained publications related to the effectiveness of DDL. Publications addressing the effectiveness of DDL were the most prominent in Cluster #0, among which Boulton and Cobb (2017) is the publication with the highest sigma across this stage. In this study, the authors undertook a meta-analysis to measure the effectiveness of DDL for language acquisition, and concluded with a call for longitudinal research and the incorporation of delayed post-testing. Another prominent publication, Smart (2014) examined the effectiveness of paper-based DDL for English as a Second Language (ESL) grammar and found more effective learning outcomes

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

of inductive learning with printed corpus-based materials than deductive corpus-based and traditional approaches.

Huang (2014) focused on patterns of abstract nouns in L2 writing; Daskalovska (2015) concentrated on verb–adverb collocations, and Vyatkina (2016) analysed verb–preposition collocations. These studies reached a consensus on the positive role that DDL plays in facilitating learners' acquisition of collocations. Although Frankenberg-Garcia (2014) was not identified as a prominent publication, a close examination revealed that it enjoys high impact, as indicated by its high co-citation frequency (8). This paper examined the impact of corpus-based examples on language comprehension and production, and found improvements in learners' awareness of grammatical properties. In Cluster #8, Lee *et al.* (2019), a meta-analysis of the effectiveness of DDL for L2 vocabulary acquisition and the variables affecting learning outcomes, possessed the second highest sigma score.

4.4.2 Pedagogy-centred publications

Pedagogy-centred publications feature primarily in Clusters #0, #4 and #7. Unlike effectiveness-centred research, which typically measures the effectiveness of DDL through tests, pedagogy-centred research has a strong focus on the implementation of DDL in classroom environments (e.g. Charles, 2015; Flowerdew, 2012), the learners' perceptions of DDL (e.g. Charles, 2014; Geluso & Yamaguchi, 2014), and factors influencing its implementation (e.g. Cotos, 2014).

Geluso and Yamaguchi (2014) in Cluster #0 presented a curriculum design focusing on spoken fluency and surveyed students' attitudes towards DDL (co-citation frequency: 8). Cotos (2014), a frequently co-cited publication in Cluster #0 (with a co-citation frequency of 7), focused on the role of corpora in students' language learning by comparing their interactions with a local learner corpus and a native-speaker corpus. Charles (2014), also co-cited seven times in Cluster #0, conducted qualitative research on the use of self-built corpora from a longitudinal perspective. Pérez-Paredes, Sánchez-Tornel and Calero (2012), a frequently co-cited publication in Cluster #4 (co-cited five times), examined learners' search strategies in DDL activities. Flowerdew (2012), the most representative and most co-cited publication in Cluster #7, focused on applications of DDL in classrooms, and discussed the impediments to DDL in pedagogy and the pedagogical application of corpora.

Additionally, the analysis of the labels shows that DDL studies have displayed some features of the fourth stage. For instance, Cluster #5 ("Pedagogical implications of DDL") and #8 ("Language teachers' lesson planning") reflect the emerging Stage 4 in the DDL field. Vyatkina (2020) and Chambers (2019) in Cluster #5, as well as O'Keeffe (2021) in Cluster #8, agreed on the positive impact of DDL on learning outcomes and the advantages of DDL practices in various contexts. The need for theoretical underpinnings from the area of second language acquisition was also emphasised by O'Keeffe (2021) and Lee *et al.* (2019). These calls conform to the features of Stage 4 that involve broader applications of knowledge generated in the first three stages for various practical purposes (Shneider, 2009). Another notable feature of the fourth stage is the publication of meta-analyses and reviews (Shneider, 2009), and several examples have evidenced the emergence of the fourth stage (e.g. Boulton & Vyatkina, 2021; Lee *et al.*, 2019). However, current research has still not fully addressed the role played by variables in DDL such as the relative explicitness of instruction and cognitive learning processes (Chambers, 2019), which indicates the ongoing Stage 3. Thus, the current research status characterises the end of Stage 3 and the beginning of Stage 4.

Regarding the main research venues in RQ1, this study carried out a journal co-citation analysis to identify the most frequently co-cited journals in DDL. Table 4 lists the top 10 journals, sequenced by co-citation frequency. Among them, *Computer Assisted Language Learning* (96), *ReCALL* (87) and *Language Learning & Technology* (76) are identified as the most frequently co-cited journals in the field, and are thus the main venues for DDL research. This corresponds

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Table 4. Top 10 co-cited journals

Rank	Journal	Co-citation frequency	Number of burst publications
1	*Computer Assisted Language Learning*	96	1
2	*ReCALL*	87	7
3	*Language Learning & Technology*	76	3
4	*System*	74	0
5	*Applied Linguistics*	64	1
6	*Language Learning*	59	2
7	*TESOL Quarterly*	57	0
7	*English for Specific Purposes*	57	1
9	*Journal of Second Language Writing*	52	0
10	*ELT Journal*	51	0

to the common aim of all three journals, which is to encourage technology-mediated language learning and teaching, especially those involving innovative practices. Other prominent journals include *System* and *Applied Linguistics*, with a co-citation count of 74 and 64, respectively. Figure C2 in Appendix C displays a list of highly cited journals.

Finally, we calculated the impact of the journals by considering the number of prominent publications appearing there. As co-citation counts of one journal are closely associated with the number of publications, the distribution of burst publications offers a more objective view of influential journals. The analysis shows that the 19 burst publications across three stages are distributed unevenly across nine journals. These publications were published predominantly by *ReCALL* (seven papers), which indicates that *ReCALL* is the main source of prominent DDL studies, and a prominent repository for research on DDL. Other important publication venues are *Language Learning & Technology* (three papers) and *Language Learning* (two papers). There are also other journals that possess a single burst publication: *Applied Linguistics*, *Computer Assisted Language Learning*, *English for Specific Purposes*, *Indian Journal of Applied Linguistics*, and *Journal of English for Academic Purposes*. This indicates wide interest in DDL from other journals in applied linguistics.

4.5 Latest transformative research

To answer RQ3, an SVA was conducted to identify transformative research in the last three years (2019–2021) and predict future directions of DDL research (the specific results are shown in Table 5). The analysis identified seven transformative publications (sequenced by the C_{KL} score, as introduced in Appendix A). Among these studies, five belong to empirical studies and the remaining two are reviews.

More specifically, two transformative studies focused on teacher education in DDL. The first one, Chen, Flowerdew and Anthony (2019), reported the success of a teacher training workshop that introduced corpus-based academic writing pedagogy to English teachers in Hong Kong. In the second study, Crosthwaite, Luciana and Wijaya (2021) examined the effectiveness of a DDL training program for teachers. This shows that teacher training is an urgent need and a prerequisite for large-scale classroom implementations of DDL, which is in line with the results in Chen and Flowerdew (2018).

Three transformative studies examined different factors in DDL activities. Sun and Hu (2020) investigated the difference between paper- and computer-based corpus-informed exercises to

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Table 5. Transformative research in 2019–2021

C_{KL}	H	Publishing year	Author
0.25	0.76	2019	Crosthwaite P, Wong L L C & Cheung J
0.25	0.98	2020	Sun X & Hu G
0.24	0.75	2020	Vyatkina N
0.18	0.56	2019	Chen M, Flowerdew J & Anthony L
0.03	0.08	2021	Boulton A & Vyatkina N
0.03	0.09	2020	Crosthwaite P, Storch N & Schweinberger M
0.02	0.07	2021	Crosthwaite P, Luciana & Wijaya D

support Chinese undergraduates' acquisition of hedging in writing. Crosthwaite, Storch and Schweinberger (2020) examined the effectiveness of DDL for learners' resolution of errors, with consideration of different degrees of directness in the written corrective feedback provided by teachers. Crosthwaite, Wong and Cheung (2019) identified corpus query and usage patterns based on actual data collected from an online corpus platform. This shows that current DDL studies display a predominant interest in implementing DDL in classroom practices. The investigation of variables affecting the effectiveness of DDL, such as the type of activities, the role of written corrective feedback, and learners' query strategy in using corpora, are at the centre of current work.

Of interest is that two transformative review studies, Boulton and Vyatkina (2021) and Vyatkina (2020), contribute to reporting similar DDL future development directions. Both publications point out that future studies may need to focus on the integration of DDL for teaching LOTEs (languages other than English), DDL practices among learners of different proficiency and age levels, and open-access resources of DDL integrated with user guides and exercise collections for specific corpora. Boulton and Vyatkina (2021) also emphasised the necessity of advancing theories in DDL and considering different forms of learner interaction with corpora (e.g. multimedia corpora with video and sound).

5. Conclusion

This study provided a diachronic and systematic review of the development of the DDL field by implementing a co-citation analysis, SVA and close manual analysis of a corpus of 126 publications collected from the WoS core collection. In addressing RQ1 (common themes, publications and venues) and RQ2 (developmental stages), this study identified 11 main clusters and 19 prominent publications, as well as three major evolutionary stages of DDL research (namely, the conceptualising stage, the maturing stage and the expansion stage). These stages represent a shift in academic interest from the establishment of techniques and testing the effectiveness of DDL for language acquisition to the implementation of DDL in classroom practice with consideration of a range of variables, with new features of Stage 4 emerging. Current interest involves more nuanced research results that incorporate different variables, the review of knowledge generated in the first three stages and the practical implementation of knowledge in this field. The results for research venues in RQ1 indicated that *Computer Assisted Language Learning*, *ReCALL* and *Language Learning & Technology* are the main venues for DDL research, while *ReCALL* is the most influential venue for DDL research in terms of prominent publications in this field. The findings from RQ3 (publications with potentially high impact) show that the main areas of future research are the implementation of DDL in classroom teaching and teacher training.

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

The analysis shows that researchers have reached a consensus that DDL plays a positive role in promoting learning outcomes; however, little is yet known about different variables inherent in various pedagogical approaches to DDL, and individual learner differences have only begun to be addressed. Therefore, future studies may consider expanding inquiries in this line by including a more specific analysis, such as introducing DDL in classrooms, organising teacher training workshops, and examining the effect of variables (both activity and learner related) in DDL. More nuanced study designs are needed to assess DDL in different pedagogical contexts with different levels of learners.

Despite the advantages inherent in a large-scale bibliometric analysis, limitations in this approach need to be recognised. First, co-citation analysis presents a bias that favours older publications, as recent but potentially high-impact publications have had less opportunity to be cited. In this study, although SVA was employed to mitigate this bias, the analysis was used to identify the studies of transformative potential. It thus could not fully address the inherent problem of co-citation analysis in failing to compensate for recent publications. Future approaches may need to consider solving this issue by assigning greater weighting (e.g. through a weighting algorithm or normalisation) to recent papers to mitigate the chronological bias. Similarly, regarding the influence of open access on citations, 3 out of 17 impactful research articles in our dataset were found to be open access. Although it is beyond the scope of this study to examine the relationship between citation counts and relevant factors such as open access, it is certainly of value and interest to explore this issue in the future. Second, we acknowledge the Matthew effect, according to which well-known authors are more likely to be cited than less well-known authors. While the approach uses co-citations to measure the importance of individual studies, future studies may take into account other factors related to citation practice. Third, it is necessary to point out that the analysis in this study is based on articles in the core collections of WoS and their reference lists. This might undermine the impact of some highly cited or influential publications that are not indexed in the WoS core collection or not co-cited in their reference lists. For example, influential publications by Johns (1991) and Davies (2008) were not identified as highly co-cited studies in the co-citation analysis, possibly due to their relatively low co-occurrence with studies on similar themes despite their high impact as a single study. Another potential explanation for why tools like COCA and AntConc were not identified as prominent publications is inappropriate citation. Some authors might use AntConc as a tool without citation or variously cite one of the several papers or different versions of one software. Thus, caution may be needed when using the result of co-citation solely to gauge the influence of a study. Future bibliometric analyses should thus include literature from various academic platforms and a wider range of document types (e.g. dissertations, theses, book chapters, and meta-analyses). Additionally, although the bibliometric analysis based on the labels automatically generated from the citing articles and their references using CiteSpace can produce stable results for DDL studies between 1994 and 2021, further bibliometric analyses may be needed to explore new research themes, future developmental stages, and prominent publications to keep abreast of the evolving landscape of DDL.

Employing a bibliometric approach, this study provided a comprehensive picture of the development of DDL with respect to its developmental stages, the state-of-the-art, common research themes, high-impact publications, research venues and potential research directions. There is a clear need for bibliometric studies to analyse further or more detailed aspects of DDL, such as author co-citation and collaborations across regions. Future bibliometric studies (particularly those that employ CiteSpace) may compare their results with this study to identify potential changes in research direction. Researchers can also use CiteSpace to familiarise themselves with existing knowledge or identify the latest trends in a new field.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344022000222

Acknowledgements. We would like to acknowledge our appreciation for the support received from the National Social Science Foundation of China (No. 21FYYB052) and Social Science Foundation of Shaanxi Province (No. 2020K025).

Ethical statement and competing interests. The research was conducted in accordance with Shandong University's code of conduct for responsible research. No ethical issues exist in this study. The authors declare no competing interests.

References

Anthony, L. (2004) *AntConc* (Version 3.0.1). Tokyo: Waseda University. http://www.antlab.sci.waseda.ac.jp/

Aryadoust, V., Zakaria, A., Lim, M. H. & Chen, C. (2020) An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology*, 11: 1–29. https://doi.org/10.3389/fpsyg.2020.01941

Birkle, C., Pendlebury, D. A., Schnell, J. & Adams, J. (2020) Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1): 363–376. https://doi.org/10.1162/qss_a_00018

Bloch, J. (2009) The design of an online concordancing program for teaching about reporting verbs. *Language Learning & Technology*, 13(1): 59–78. https://doi.org/10125/44168

Borja, A. (2007) Corpora for translators in Spain. The CDJ-GITRAD Corpus and the GENTT Project. In Anderman, G. & Rogers, M. (eds.), *Incorporating corpora: The linguist and the translator*. Clevedon: Multilingual Matters, 243–265. https://doi.org/10.21832/9781853599873-016

Boulton, A. (2008) But where's the proof? The need for empirical evidence for data-driven learning. In Edwardes, M. (ed.), *Proceedings of the BAAL annual conference 2007*. London: Scitsiugnil Press, 13–16.

Boulton, A. (2009a) Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, 21(1): 37–54. https://doi.org/10.1017/S0958344009000068

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

Boulton, A. (2010a) Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3): 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

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

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

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

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

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

Boulton, A. & Tyne, H. (2013) Corpus linguistics and data-driven learning: A critical overview. *Bulletin Suisse de Linguistique Appliquée*, 97: 97–118.

Boulton, A. & Vyatkina, N. (2021) Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, 25(3): 66–89. https://doi.org/10125/73450

Braun, S. (2005) From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1): 47–64. https://doi.org/10.1017/S0958344005000510

Burnard, L. (ed.) (2004) *BNC Baby* [CD-ROM]. Oxford: Oxford University Research and Technology Service. http://www.natcorp.ox.ac.uk/corpus/babyinfo.html

Chambers, A. (2005) Integrating corpus consultation in language studies. *Language Learning & Technology*, 9(2): 111–125. https://doi.org/10125/44022

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

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

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

Charles, M. (2015) Same task, different corpus: The role of personal corpora in EAP classes. In Leńko-Szymańska, A. & Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 131–154. https://doi.org/10.1075/scl.69.07cha

Chen, C. (2012) Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3): 431–449. https://doi.org/10.1002/asi.21694

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Chen, C. (2017) Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2): 1–40. https://doi.org/10.1515/jdis-2017-0006

Chen, C., Ibekwe-SanJuan, F. & Hou, J. (2010) The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7): 1386–1409. https://doi.org/10.1002/asi.21309

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

Chen, M., Flowerdew, J. & Anthony, L. (2019) Introducing in-service English language teachers to data-driven learning for academic writing. *System*, 87: 102148. https://doi.org/10.1016/j.system.2019.102148

Chen, X. L., Zou, D., Xie, H. R. & Su, F. (2021) Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology*, 25(3): 151–185.

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

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H. & Chute, C. G. (2005) Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6): 422–430. https://doi.org/10.1016/j.jbi.2005.02.009

Cotos, E. (2014) Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2): 202–224. https://doi.org/10.1017/S0958344014000019

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J. & Amin, M. (2007) Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics*, 1(3): 239–248. https://doi.org/10.1016/j.joi.2007.04.001

Cresswell, A. (2007) Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In Hidalgo, E., Quereda, L. & Santana, J. (eds.), *Corpora in the foreign language classroom*. Amsterdam: Rodopi, 267–287. https://doi.org/10.1163/9789401203906_018

Crosthwaite, P., Luciana, & Wijaya, D. (2021) Exploring language teachers' lesson planning for corpus-based language teaching: A focus on developing TPACK for corpora and DDL. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2021.1995001

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

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

Daskalovska, N. (2015) Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2): 130–144. https://doi.org/10.1080/09588221.2013.803982

Davies, M. (2008) *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Flowerdew, L. (2012) *Corpora and language education*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9780230355569

Francis, W. N. & Kučera, H. (1989) *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University, Department of Linguistics.

Frankenberg-Garcia, A. (2012) Learners' use of corpus examples. *International Journal of Lexicography*, 25(3): 273–296. https://doi.org/10.1093/ijl/ecs011

Frankenberg-Garcia, A. (2014) The use of corpus examples for language comprehension and production. *ReCALL*, 26(2): 128–146. https://doi.org/10.1017/S0958344014000093

Geluso, J. & Yamaguchi, A. (2014) Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 26(2): 225–242. https://doi.org/10.1017/S0958344014000044

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

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

Huang, Z. (2014) The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL*, 26(2): 163–183. https://doi.org/10.1017/S0958344014000020

Hyland, K. & Jiang, F. K. (2021a) A bibliometric study of EAP research: Who is doing what, where and when? *Journal of English for Academic Purposes*, 49: 100929. https://doi.org/10.1016/j.jeap.2020.100929

Hyland, K. & Jiang, F. K. (2021b) Delivering relevance: The emergence of ESP as a discipline. *English for Specific Purposes*, 64: 13–25. https://doi.org/10.1016/j.esp.2021.06.002

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

Johns, T. (1991) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *English Language Research Journal*, 4: 27–45.

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

Jung, U. O. H. (2005) CALL: Past, present and future – A bibliometric approach. *ReCALL*, 17(1): 4–17. https://doi.org/10.1017/S0958344005000212

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

Kita, K. & Ogata, H. (1997) Collocations in language learning: Corpus-based automatic compilation of collocation and bilingual collocation concordancer. *Computer Assisted Language Learning*, 10(3): 229–238. https://doi.org/10.1080/0958822970100303

Köhler, J., Philippi, S., Specht, M. & Rüegg, A. (2006) Ontology based text indexing and querying for the semantic web. *Knowledge-Based Systems*, 19(8): 744–754. https://doi.org/10.1016/j.knosys.2006.04.015

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

Lei, L. & Liu, D. (2019) Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, 40(3): 540–561. https://doi.org/10.1093/applin/amy003

Lin, Z. & Lei, L. (2020) The research trends of multilingualism in applied linguistics and education (2000–2019): A bibliometric analysis. *Sustainability*, 12(15): 6058. https://doi.org/10.3390/su12156058

Liu, S. & Zhang, S. (2021) A bibliometric analysis of computer-assisted English learning from 2001 to 2020. *International Journal of Emerging Technologies in Learning (iJET)*, 16(14): 53–67. https://doi.org/10.3991/ijet.v16i14.24151

Liu, Y. & Hu, G. (2021) Mapping the field of English for specific purposes (1980–2018): A co-citation analysis. *English for Specific Purposes*, 61: 97–116. https://doi.org/10.1016/j.esp.2020.10.003

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/10.3991/ijet.v12i03.6523

Meara, P. (2012) The bibliometrics of vocabulary acquisition: An exploratory study. *RELC Journal*, 43(1): 7–22. https://doi.org/10.1177/0033688212439339

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

Murphy, B. (1996) Computer corpora and vocabulary study. *The Language Learning Journal*, 14(1): 53–57. https://doi.org/10.1080/09571739685200391

O'Keeffe, A. (2021) Data-driven learning – A call for a broader research gaze. *Language Teaching*, 54(2): 259–272. https://doi.org/10.1017/S0261444820000245

O'Sullivan, Í. (2007) Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3): 269–286. https://doi.org/10.1017/S095834400700033X

Park, H. & Nam, D. (2017) Corpus linguistics research trends from 1997 to 2016: A co-citation analysis. *Linguistic Research*, 34(3): 427–457. https://doi.org/10.17250/KHISLI.34.3.201712.008

Pérez-Paredes, P. (2022) A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1–2): 36–61. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., Sánchez-Tornel, M. & Calero, J. M. A. (2012) Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics*, 17(4): 482–515. https://doi.org/10.1075/ijcl.17.4.02par

Pritchard, A. (1969) Statistical bibliography or bibliometrics? *Journal of Documentation*, 25: 348–349. https://doi.org/10.1108/eb026482

Pritchard, A. & Wittig, G. R. (1981) *Bibliometrics: A bibliography and index*. Watford: ALLM Books.

Sebastian, Y. & Chen, C. (2021) The boundary-spanning mechanisms of Nobel Prize winning papers. *PLOS ONE*, 16(8): e0254744. https://doi.org/10.1371/journal.pone.0254744

Shneider, A. M. (2009) Four stages of a scientific discipline; four types of scientist. *Trends in Biochemical Sciences*, 34(5): 217–223. https://doi.org/10.1016/j.tibs.2009.02.002

Smart, J. (2014) The role of guided induction in paper-based data-driven learning. *ReCALL*, 26(2): 184–201. https://doi.org/10.1017/S0958344014000081

Sun, X. & Hu, G. (2020) Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168820954459

Vannestål, M. E. & Lindquist, H. (2007) Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course. *ReCALL*, 19(3): 329–350. https://doi.org/10.1017/S0958344007000638

Vyatkina, N. (2016) Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL*, 28(2): 207–226. https://doi.org/10.1017/S0958344015000269

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

Yoon, C. (2011) Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10(3): 130–139. https://doi.org/10.1016/j.jeap.2011.03.003

Zareva, A. (2017) Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal*, 71(1): 69–79. https://doi.org/10.1093/elt/ccw045

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press

About the authors

Jihua Dong is Professor, Qilu Young Scholar, and Taishan Young Scholar at Shandong University, China. Her research interests include corpus linguistics, data-driven teaching, and academic writing. She has published in *English for Specific Purposes*, *International Journal of Corpus Linguistics*, *Journal of English for Academic Purposes*, and *System*, among others.

Yanan Zhao is a PhD student in the School of Foreign Languages and Literature at Shandong University. She has obtained a master's degree in languages and linguistics from the University of Melbourne, Australia. Her research interests include data-driven learning, second language learning and teaching, and corpus linguistics.

Louisa Buckingham lectures in applied linguistics at the University of Auckland. She has published on corpus-informed discourse analysis, language learning and sociolinguistics. She has published in various journals, including *TESOL Quarterly*, *System*, *Journal of English for Academic Purposes*, *English for Specific Purposes*, and *Journal of Multilingual and Multicultural Development*.

Author ORCiD. Jihua Dong, https://orcid.org/0000-0001-7864-2319
Author ORCiD. Yanan Zhao, https://orcid.org/0000-0001-6840-9580
Author ORCiD. Louisa Buckingham, https://orcid.org/0000-0001-9423-0664

https://doi.org/10.1017/S0958344022000222 Published online by Cambridge University Press