International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 19, No. 2, 2025

Data-driven Learning in Second Language Writing

-- A systematic review of efficacy, challenges, and future directions

Huilin Luo*

School of Foreign Languages, Southwest Petroleum University, Chengdu, Sichuan 610500, China
* **Corresponding author:** Huilin Luo (Email: 1444657726@qq.com)

Abstract: This systematic review investigates the integration of Data-Driven Learning (DDL) in second language writing pedagogy through empirical and theoretical perspectives, analyzing pedagogical impacts, methodological constraints, and future directions. Findings demonstrate DDL's effectiveness in improving writing accuracy, fluency, and learner autonomy through exposure to authentic linguistic corpora. Current limitations include restricted corpus accessibility, variable teacher competencies, learner adaptation barriers, and a predominance of short-term studies. Comparative analysis reveals distinct research trajectories: domestic studies prioritize error correction algorithms and automated feedback systems, while international scholarship emphasizes lexical-grammatical development and feedback mechanism validation. The synthesis identifies critical research gaps requiring large-scale longitudinal investigations, extended pedagogical interventions, and specialized teacher training programs. The paper concludes by proposing an integrative framework for optimizing DDL implementation, advocating for cross-disciplinary collaborations between computational linguistics and language pedagogy, coupled with technological innovations in corpus interface design. These findings contribute to advancing evidence-based practices in data-enhanced language education.

Keywords: Computer-assisted Language Learning, Corpus Linguistics, Data-Driven Learning (DDL), Second Language Writing, Writing Pedagogy.

1. Introduction

In the field of corpus linguistics and computer-assisted instruction, data-driven learning (DDL) has attracted much attention as an important research component. Corpus linguistics provides insights into language usage patterns through the study of corpora, providing empirical support for language learning and teaching. DDL, on the other hand, is a method that utilizes real language data from corpora to help learners extract language patterns from them and facilitate language learning. In recent years there have been numerous studies related to writing for DDL. From the perspective of learners and educators, DDL can stimulate students' interest in writing and improve writing accuracy and fluency, but it also faces challenges such as teacher familiarity and individual student differences. Studies at home and abroad have shown that although DDL has many advantages in writing teaching, it also needs more in-depth research and improvement to better fulfill its role in language learning. Future research should focus on individual learner differences, increase the sample size, extend the intervention time, and enhance teacher training to promote the further application and development of DDL in writing instruction.DDL has the potential to become an important tool in the field of writing instruction, providing students with more personalized and efficient learning support. This review describes the current research hotspot of second language writing in combination with DDL, starts from the core concepts related to DDL and writing, develops and elaborates the current status of domestic and international research on DDL in combination with second language writing and its research focus, research gaps, strengths and weaknesses in DDL second language writing teaching and its future direction, and finally concludes by summarizing the main points and findings of the domestic and international research reviews on DDL and the strengths and weaknesses of DDL for L2 writing, and puts forward future prospects and research directions. And put forward the future outlook and research direction.

2. Core Concepts of DDL and its Application in Writing

DDL belongs to the focus of research in the field of corpus linguistics and computer-assisted instruction. Corpus linguistics, as the name suggests, is a discipline that studies corpora. Specifically, corpus linguistics refers to the study of language using corpora. Corpus linguistics includes the use of electronic text collections for language analysis. Corpus linguistics involves the collection and analysis of authentic texts to provide evidence for describing the nature, structure, and use of language. Corpus linguistics has a long tradition of using texts as the empirical basis for linguistic description, and of studying all aspects of language, including phonology, vocabulary, grammar, and discourse. Corpus linguistics offers new insights into how language changes systematically across different historical, regional and sociolinguistic contexts, genres and registers (Lusta A et al., 2023).

Corpus as the center of gravity of corpus linguistics research has been defined in detail and specifically by major scholars. A corpus refers to a large, principled, computer-readable collection of texts that can be analyzed for patterns of language use in different contexts and is thought to help improve learners' performance in second language writing (Ş AHIN KıZıL A, 2023). Similarly, some scholars consider a corpus as a large amount of real-world text used for a specific purpose, a typical set of texts with annotations (e.g., additional information may include words marked as lexical, "POS"). Some scholars consider a corpus to be a finite-sized body of machine-readable text, sampled in order to maximize the representation of the linguistic diversity under consideration. This definition suggests that a corpus is

characterized by samplability, representativeness, finite size, machine-readability, and referentiality; a corpus is a collection of real language, both written and spoken, compiled for a specific purpose. From the above definitions it is easy to conclude that a corpus is a large computer-readable database of a limited size and representative of the universal language that has been created for a specific purpose and brings together authentic languages. Nowadays, the term corpus is more commonly used for computerized databases created for linguistic research (2023). With the popularity of corpora came the concept of corpus literacy, which refers to an individual's ability to use a corpus, which includes navigating and searching the corpus, interpreting search results, and recognizing the limitations of the corpus and the advantages of language learning (2023).

By summarizing the definition of corpus, we can find that corpus can provide a large amount of authentic linguistic data, which can be used in teaching to provide teachers and students with linguistic bases and authentic linguistic scenarios, and help to promote language learning. With the popularization of corpus, computer-assisted language teaching related to corpus has received a lot of attention in recent years, and all kinds of information technology are becoming more and more popular and are closely intertwined with human activities, and today's foreign language teaching, if it lacks the technological elements, it will be difficult to create an authentic and diversified language learning environment, and the fusion of related information technology and foreign language teaching has become an inevitable trend (Xu & Liu , 2019). Computer-assisted language learning (CALL), abbreviated as CALL, refers to all kinds of integration of information technology and foreign language teaching in a broad sense (2019). Driven by big data and deep learning, the wide application of artificial intelligence has triggered important changes in the field of language education. In this context, computer-assisted language learning research has been accelerated, and CALL has entered the stage of in-depth application after 2010 (Tian & Peng, 2022). Among the current research and classroom examples of computer-assisted foreign language teaching, DDL for language learning based on authentic discourse is one of the notable research trends and hotspots. DDL is a term that appeared only in 1991. Simply put, DDL is the process by which learners extract inferred language patterns from an authentic corpus (SATAKE Y, 2020). From the learner's point of view, DDL (Data-Driven Learning) is a method of learning language based on corpus, especially on concordance-based material, i.e., the learner, with a certain question, uses retrieval software to discover patterns and draw conclusions based on observing and analyzing a large amount of authentic corpus, and masters a certain grammatical structure through real-time practice. That is to say, learners with a certain problem, using search software to discover patterns and draw conclusions based on observing and analyzing a large amount of real corpus, and through real-time practice, mastering a certain grammatical structure or the usage of a certain word, so this method of learning is also known as the "research-then-theory" method (Gan & Zou, 2010). From an educator's point of view, DDL is a more radical approach to language teaching than traditional teaching in which learners engage with the corpus either directly or indirectly through the materials (Sun W & Park E, 2023). Crosthwaite (2019) defines DDL as a pedagogical approach in which learners either engage with language data indirectly (i.e., indirectly) through the use of paper-based materials drawn from a corpus or indirectly (i.e., indirectly) through the use of paper-based materials drawn from a corpus or indirectly access to linguistic data (i.e., indirect DDL) or directly by accessing online corpus data to generate index lines (i.e., direct DDL). The direct approach to DDL allows learners almost unlimited access to authentic linguistic data, fosters learner autonomy, and promotes discovery-based learning. However, its adoption places high demands on teachers' familiarity with corpus tools and runs the risk of overwhelming and distracting learners, especially those who are new to corpus consultation. Indirect use of corpora provides only limited access to corpus data and restricts learners' own data-based discovery. However, through the use of corpus information, printed materials, and well-designed side activities, indirect methods can be more stress-free and make corpus data immediately available for immediate benefit (Sun X & Hu G, 2020). From a classroom application perspective, DDL applies computer-generated indexed line entries in the classroom to motivate students to discover symmetrical patterns in the target language and to enhance activities and tasks based on the target this word output (Uftah M, 2023). Summarizing the previous findings, Liu (2016) mentioned that the features of DDL compared with traditional foreign language teaching include student-centeredness, authenticity of language materials, learning process that emphasizes students' self-exploration and discovery and advocacy of bottom-up inductive learning.

DDL, as a hot area in foreign language teaching, also has a focus on research. According to Pérez-Paredes P (2019) in a systematic review of data-driven learning in five journals published over a five-year period of relevant studies related to corpus and DDL instruction, it was found that the main focus of DDL research is the use of indexing lines and collocations in the development of college students' writing skills, and that DDL has a positive effect on assisted learning in the area of second language writing. Writing is an essential skill for language learners, but is often considered challenging for learners, especially in the context of English as a Foreign Language (EFL). In writing, as in many other different realities, the amount and type of naturally occurring linguistic data is relatively limited (2023). Second language writing is the process of encoding in a fixed, regular, unimodal, static context using the English writing system (Satake Y, 2020). Therefore, this review unfolds the current status and focus of domestic and international research on DDL combined with second language writing, research gaps, strengths and weaknesses in teaching second language writing in DDL and future directions in the context of DDL's current research hotspot, second language writing.

3. Related Studies on DDL Writing

Research on Data-Driven Learning (DDL) in second language (L2) writing has gained momentum, focusing on its theoretical foundations, pedagogical applications, and empirical evidence. This section synthesizes domestic and international literature to highlight key findings, gaps, and future directions, emphasizing DDL's role in enhancing writing skills through authentic corpus data and learner-centered discovery.

3.1. Literature Review at Home

The combination of DDL and second language writing has

been launched relevant research in China. New teaching methods in the context of digital learning have been emphasized, and language learning based on authentic discourse, such as DDL, is a hot topic of teaching research in the context of information technology. The hotspots in the last decade cover the field of DDL writing (2022). In a review of related studies, Tian Zhen (2022) and others also found that data-driven learning, based on natural discourse (discourse-based approach) and language retrieval (hands-on concordancing), positively affects language learning and can effectively improve learners' articulation in academic writing. Computer-generated corrective feedback can improve the quality of learners' second language writing however the effectiveness of error correction is influenced by factors including feedback timing, feedback type (real-time versus diachronic correction), feedback training, and learner interaction. In addition, benefiting from the development and application of automated assessment systems, machine scoring of writing outcomes has also been shown to improve learners' writing ability. In the research on corpora mentioned by Liu (2016) and others, there are a lot of research results on the application of corpora to second language writing teaching, such as the analysis of second language writing outcomes, especially the analysis of writing errors, and the diagnosis of language skills. Liu (2016) study introduces the pedagogical application and findings of CQPweb, which was applied to the teaching of the course "Writing in English for Science and Technology" for one semester, and it was found that the appropriateness of the choice of words and phrases, the completeness of the structural elements, and the reasonableness of the distribution of structural points of the compositions of the students who had been assisted by CQPweb had been improved to different degrees. Gan (2010) found that the aptness of word choice, the completeness of structural elements, and the rationality of the distribution of structural points of students' compositions were all improved to different degrees by CQPweb. According to Gan (2010), the corpus-based data-driven learning approach makes it possible to develop students' independent learning ability in English writing. The network has enough information reserve and output capacity to provide teachers with second language acquisition information that is larger than students' current language ability with vivid, graphic, multidimensional and three-dimensional information sources, to stimulate students' learning interest, to improve teaching efficiency per unit of time, and to cultivate students' self-directed learning ability. Xu (2019) mentioned that the automatic correction and feedback system for second language writing helps learners to correct errors in writing. Therefore, the current focus of DDL combined with writing research in China focuses on writing articulation, writing assessment and feedback for error correction.

However, there are still some limitations and unexplored areas for teaching writing in the context of DDL. Liu (2016) mentioned that most of the existing results reflect the application of corpus by experts, and rarely involve the direct application of corpus by students. Although there are general corpora available on the Internet (e.g., BYU Corpus series corpus, Sketch Engine, etc.), the online use of English for Specialized Purposes (ESP) corpora is still difficult, and ESP teaching is faced with a variety of problems such as the shortage of teaching resources, outdated teaching materials, and outdated teaching methods. The corpus has not been widely used in English writing classrooms, and the instruction of corpus-assisted learning should be more detailed and simplified, and the training in this area should be appropriately increased in length and intensity, so as to avoid affecting the experimental results and efficiency due to the low level of proficiency. Second, research on DDL-assisted writing instruction should strengthen the comprehensive examination of learners' individual differences as well as cognitive, affective, and socio-cultural factors (2022). Xu (2019) mentioned that the practical application effect of corpus also depends on teachers' corpus expertise. For example, limited by the level of relevant expertise of foreign language teachers, the research results and technologies related to spoken corpus have not yet been widely and effectively applied to foreign language teaching, and the related DDL theory research is not yet sufficient. Gan (2010) study shows that corpora are not very effective for chapter structure modification and logical relations. Therefore, the current research on teaching writing in DDL needs to consider the compaction of corpus resources, the popularization of the corpus, the session-specific writing tasks that the corpus contributes to, the effect of individual learner factors on DDL, the proficiency of teachers in the corpus, and the articulation of teaching and learning.

The strengths and weaknesses of teaching writing with DDL have been initially summarized among previous empirical and theoretical studies. Gan (2010) mentioned that DDL combined with writing teaching makes the teacher's analysis of writing skills and theoretical presentation more intuitive, develops students' independent learning ability and ability to use modern technology to process manuscripts, facilitates the teacher's corrections and explanations, improves efficiency, and is conducive to the construction of a harmonious teacher-student relationship; Liu (2016) study based on the CQPweb to assist students' writing also confirmed that the DDL group produced a higher quality of writing than the non DDL group produced higher quality compositions; the advantages and potentials of corpus application in foreign language teaching were affirmed in the study of Xu (2019). Of course, there are some shortcomings in the use of DDL in writing classrooms. For example, the limitations of corpus tools, such as CQPweb, from students' perspectives are centered on: 1) the search method is too complex, and it is difficult to write search expressions; 2) the corpus resources are not rich enough, the classification of disciplinary specialties is not refined enough, and there is a lack of representative texts in some disciplines. Xu (2019) mentioned that there are still some inconveniences in the use of corpus, such as insufficient types of corpus, teachers' lack of corpus expertise and training, students' lack of basic understanding of corpus and lack of autonomy, which makes it difficult to complete the identity transition. Therefore, the problems of using DDL in writing teaching mainly focus on three aspects: teacher's perspective, student's perspective and corpus perspective.

In view of the gaps in the current research on DDL writing and the limitations of DDL use in teaching and learning, related literature also suggests future research directions. Based on her own research, Gan (2010) argued that future research could combine teacher instruction and DDL to explore other issues in the writing instruction classroom, emphasizing the importance of teachers in the DDL writing classroom. Tian (2022) believes that personalized tracking of learners' language development paths based on big data will be more common in future research. Liu (2016) argued that

research related to DDL writing instruction should be conducted in a broader temporal and spatial context, and that empirical research on CQPweb-assisted ESP writing instruction should be carried out to not only compare the effects of writing instruction in experimental and control classes horizontally, but also to follow up on the learners who persist in using CQPweb after the course is over, and to observe the changes in their writing proficiency over time. Xu (2019) believes that teachers should clarify the advantages, disadvantages, similarities and differences of all kinds of teaching tools, especially the characteristics, advantages and limitations of all kinds of information technology, digital tools, decision-making in the context of the real teaching environment and systematic assessment of the effect of informationized foreign language teaching. Therefore, future empirical research on DDL writing instruction should focus on: delayed post-testing and tracking observation of DDL on learners' writing, DDL empirical research should be appropriately scaled up and teachers' decision-making and evaluation of DDL in writing classrooms.

3.2. Literature Review Abroad

Foreign research on DDL writing instruction precedes domestic research, and the relevant empirical research is more abundant. First of all, based on the systematic review articles, we can have a general understanding of the current status of research related to DDL writing instruction in recent years. These studies follow the PRISMA framework, which is called Preferred Reporting Items for Systematic Reviews and Meta-Analyses. The framework was first proposed by Moher et al. in 2009, revised in 2020, and re-emphasized by Page et al. in 2021. this approach. The guidelines for this framework include 27 items designed to improve the transparency and consistency of systematic reviews. Adherence to this guide is essential to ensure transparency, clarity, and thoroughness in the systematic evaluation of articles (2023). Pérez-Paredes P (2019) summarizes prior relevant literature and finds that DDL research ranges from endorsing the benefits of DDL and claiming its superiority over other learning methods to articulating the top issues learners have when confronted with DDL. Whereas the notable meta-analysis conducted on DDL writing showed that DDL studies produced moderate to high effect sizes in both within-group and between-group designs, and secondly, extensive empirical research is needed for the teaching of writing in DDL.The systematic review study by Lusta A (2023) mentioned that in the field of corpora and DDL, quantitative synthesis methodology has made a significant contribution to the development of corpora knowledge, as mentioned in the four noteworthy meta-analyses and four bibliometric analyses demonstrate. However, the lack of comprehensive qualitative research is evident, with only two critical reviews and one systematic review published to date. The meta-analytic reviews show that while quantitative research in the field of DDL research is growing, DDL research still needs a stronger foundation to drive continued development. Regarding the research methods used in DDL-related prior attainment research methods, mixed methods were used the most, followed by quantitative and finally qualitative. Specific methodological research methods include: surveys, assessments, reports, interviews, observations, questionnaires, data tracking, and output data; theories that are often used to support empirical research on DDL include the attention hypothesis, socio-cultural theories, constructivism, and usage-based theories (2023). It was also mentioned in the study that research related to the teaching of writing with DDL has been largely limited to university classrooms, that DDL continues to attract interest from language pedagogy, especially in writing and vocabulary, and that recent research has emphasized the integration of different language skills through DDL approaches. Pérez-paredes P (2019) also found that the focus of DDL research is shifting from corpus linguistics to language pedagogy and is increasingly emphasizing second language users rather than the technology itself; the main focus of current DDL research is on the use of indexical lines and collocations in the development of college students' writing skills.Sun's (2023) study mentions that DDL is effective in improving learners' collocations, which leads to better long-term memory and performance, fewer collocational errors in writing, and writing more fluent and authentic sentences; the predominance of Hands-on teaching methods in the relevant studies suggests that teachers have an active role in utilizing corpus technology to facilitate language learning and teaching. Through the systematic review of the articles, we can find that the current research on DDL writing has been effective, is a major trend in current research, and focuses on the learners themselves and the guiding role of teachers in the higher education environment.

Specific empirical research on DDL writing instruction is thriving abroad.Crosthwaite P's (2017) study focuses on second language writing feedback, suggesting that more and more studies are focusing on the value of corpus-based Data-Driven Learning (DDL) for written error correction in a second language (L2), with generally positive results. However, a potential challenge for language teachers involved in this process is how to provide feedback on students' written DDL compositions. Using 32 learners from mainland China with English language learning backgrounds as subjects, he investigated how to provide feedback on DDL students' written compositions, identifying the types of errors that teachers could address in a timely manner by focusing their feedback so that they could conduct a corpus query and how the nature of this feedback could affect the success rate of the query made. The results indicated that students used the corpus to correct word choice, word form, collocation, and phrase errors, but less frequently used the corpus to correct deletion or morphosyntactic errors, which is consistent with Sun's (2023) findings. It was also found in the study that students' overall attitudes towards corpus use to aid second language teaching were positive, but the reason why some students were unable or unwilling to address these types of errors through corpus consultation was most likely due to the influence of teacher feedback on revision errors in written submissions and that the time and effort spent analyzing the index line data and understanding the teacher's feedback on their writing was perceived to be difficult. Satake Y(2020) also conducted a study on error correction feedback in writing and found that the corpus had a high accuracy in error correction for coronal and prepositional omissions, but that a dictionary would be more effective for lexical error correction. Crosthwaite P(2020) conducted a study on whether and how the form of Written Corrective Feedback (WCF) provided for L2 writing could contribute to an effective corpus for L2 error resolution. queries was conducted and found that the less WCF the better if learners were to successfully revise errors through corpus queries,while more direct types of WCF tended to result in students revising errors without querying the corpus. Larsen-Walker M (2017) mentioned that previous

research on second language writing has emphasized the importance of articulation for fluent academic writing, however second language writers tend to overuse and misuse linking adverbs (LAs), including subordinating conjunctions (because) and transition words (however), which reduces the articulation and readability of their texts, and therefore initiated a study on linking adverbs in DDL-assisted writing. The results of the study showed a slight improvement in students' ability to use LAs correctly after using the corpus and that errors in the use of prepositions could be corrected after referring to the corpus, whereas errors in subject-verb agreement could not. In light of the fact that many previous studies have established that vague constraints are a hallmark of successful writing for students at the high school and/or undergraduate level, that appropriate vague constraints contribute positively to the overall quality of the writing, especially to the effectiveness of the argument, and that more vague means are a characteristic of stronger writing and higher scores on essays, Sun X (2020) conducted a study to determine whether or not direct and indirect DDLs were effective was studied. It was found that direct DDL had a practically meaningful (though statistically insignificant) effect on the frequency of ambiguous language use, whereas indirect DDL had a similar effect on the kind of ambiguous language use as the frequency significance of direct DDL, and a statistically significant and large effect on the frequency of ambiguous language use.In a study by Crosthwaite P (2019), it was suggested that for the use of multidisciplinary corpus use of Relatively few large-scale DDL studies have been conducted, and little is currently known about graduate students' disciplinary corpus use or querying habits, so a study was conducted among graduate students in multiple disciplinary contexts, and it was found that there were cross-disciplinary differences in students' preferences for corpus-assisted writing features and keyword choices. He also mentioned that since DDL is still largely seen as an extracurricular activity that takes place outside of regular class hours or in non-credit courses during vacations, the scope of most DDL studies has been relatively small. ŞAHIN KıZıL A (2023) conducted a study about whether DDL has an effect on writing CAF. The results of the study showed that writing revision using DDL had a significant positive effect on fluency and lexical complexity, and the DDL group outperformed the non-DDL group in terms of lexical diversity and fluent writing. However, there was no statistically significant evidence of the effect of DDL on accuracy and grammatical complexity.In a study by Muftan M (2023), it was found that BNCweb-assisted writing can help students to improve fluency, consistency, and complexity, and most of the students had a positive attitude towards this, but the lack of time for students was the biggest problem of corpus use. From the above studies, it can be found that research on teaching writing with DDL focuses on the effectiveness of written feedback on writing, the use of specific words in writing (e.g., connecting adverbs, ambiguities, etc.), and the varying effectiveness of direct DDL and indirect DDL. In the above mentioned researches, at the same time, they also reflect their own limitations and gaps, many of which mention that the sample content is too small, and other problems include the short duration of DDL interventions, failure to study the medium- and long-term effects of DDL, and failure to consider the effects of learners themselves.

From the foreign empirical studies on DDL writing, it is not difficult to summarize the advantages and disadvantages of DDL as an instructional method for assisting second language writing. DDL is based on naturally occurring linguistic data and allows the learner to interact with authentic inputs; in contrast to the rule-based learning that has been prioritized in traditional language learning materials, DDL assumes a discovery-based learning in which the learner learns by reviewing the review of indexed line output to infer rules and explore rule patterns. Finally, in contrast to traditional textbooks that treat grammar and vocabulary as separate components, DDL takes a lexical-grammatical approach to language teaching. Keyword-context, learners can easily and effectively notice lexico-grammatical patterns that can eventually be recycled in their writing.DDL enables learners to generate more accurate and complex syntactic structures. Since DDL activities provide learners with the opportunity to notice, discover, and analyze correct linguistic patterns through authentic input, it has the potential to improve the accuracy of learners' writing. Noticing and learning lexico-grammatical patterns in corpus data can also help learners to write more fluently. (Ş AHIN K ı Z ı L A,2023). DDL encourages active participation in the learning process, which usually requires students to independently discover or study linguistic rules based on observing and examining indexing results. Unlike rule-based language learning, which tends to isolate grammar and vocabulary, DDL encourages a more lexico-grammatical approach by allowing students to use an indexer to access common lexical or grammatical patterns of the search item. Given these characteristics, DDL is recommended as an excellent strategy for facilitating second (L2) or foreign (FL) language learning. Using a corpus while writing can help them become more accurate writers because the extended time spent on DDL tasks helps them realize their weaknesses. It also helps them recognize correct language patterns. In DDL, it is often the case that students conduct research and identify applicable collocations or idiomatic expressions that are relevant to their writing. Most of the students reported that frequent use of collocations in Keywords in Context (KWIC) style helped to increase their level of collocation awareness, which ultimately helped them to use these collocations more effectively in their writing activities. They usually identify inappropriate arbitrary pairings of verbs with nouns or adjectives with nouns without assessing their applicability in the target language (2023).

However, there is a lack of usefulness of DDL for low-level learners, differences in the success of individual learners (students' learning styles, disciplinary backgrounds, motivation, linguistic competence) in using DDL, inability of learners to adopt an inductive approach to language learning, conflict with background culture, difficulty in using authentic language rather than textbook language, lack of teachers' knowledge about DDL, lack of DDL materials in English textbooks, teachers' negativity and skepticism, and problems with technical or linguistic knowledge and investment of time or training (Crosthwaite P, 2019). Muftan M (2023) also mentions that corpus is too much of a time investment and too demanding for the learners, complex operations, and that the use of corpus alone is not sufficiently accurate in paraphrasing (it has to be combined with others). Comparing with the national literature review, we can get similar conclusions: the advantages of DDL used in writing include increasing students' autonomy, providing students with the opportunity to access authentic corpus, and thus improving learners' writing authenticity; the disadvantages of DDL used in writing include the inability of individual students to realize

corpus knowledge, teachers' unfamiliarity with and negative attitude towards DDL, the scarcity of corpus materials in classrooms and lack of technology, etc.

Therefore, based on the above findings and conclusions, the researcher provides suggestions for future research directions on writing in DDL. First, in terms of sample size Ş AHIN KıZıL A (2023), Crosthwaite P (2020), Larsen-Walker M (2017) and Sun X (2020) have made demands on sample size, arguing that the sample size reduces the ability of statistical tests to reliably detect certain effects of the DDL treatment, and that therefore, future research should be conducted over a longer period of time and with a larger study sample, and try to overcome the effects caused by some confounding variables. Secondly, in terms of DDL intervention duration, Muftan M (2023) and Sun X (2020) both mentioned that duration is not enough and further research could test whether prolonged DDL interventions produce better and more sustainable learning outcomes.Crosthwaite P (2019) also mentioned that there is still a need to study the effects of corpus and DDL on medium- and long-term writing or language development of longitudinal effects. Finally, from the learner's perspective, Sun X (2020), Satake Y (2020) and Crosthwaite P (2019) also agree that there is a need to understand individual and disciplinary differences in corpus use, and that there is a need to investigate how learner characteristics, such as learning styles, linguistic proficiency, language learning ability, and metacognitive awareness, interact with the characteristics of different DDL methods, and that further research needs to consider how learner characteristics, such as learning style, language proficiency, language learning ability, and metacognitive awareness interact with each other, and further research needs to consider learner level.

4. Findings and Challenges in DDL Writing Research

A comprehensive review of domestic and international literature shows that research on DDL writing instruction has achieved certain results. Domestic studies focus on writing articulation, writing assessment and error correction feedback, emphasizing the positive impact of data-driven learning on language learning. However, domestic studies have also pointed out some limitations, such as experts applying the corpus more than students and teachers' limited level of expertise on the corpus. In contrast, foreign studies are richer in empirical research in the field of DDL writing instruction, focusing on writing feedback, word use and other aspects. Although DDL is believed to improve students' writing accuracy and fluency and help them discover correct language patterns, it also faces many challenges, such as the lack of usefulness for low-level learners and teachers' insufficient understanding of DDL.

DDL, as an effective tool, helps to enhance students' writing accuracy and fluency, stimulate learning interest, and improve writing authenticity. However, there are some limitations in its application, such as low popularity with the corpus, limitations for low-level learners, teachers' lack of proficiency in DDL and the corpus, and technical and linguistic knowledge input issues. Future research should strengthen the study of individual learner differences, increase the sample size, extend the intervention time, and optimize teacher training to promote more effective use of DDL in writing instruction. Through continuous and in-depth research and improvement, DDL is expected to become an important tool in the field of writing instruction, providing students with more personalized and effective learning support.

5. Conclusion

This systematic review synthesizes the evolving landscape of Data-Driven Learning (DDL) in second language (L2) writing, underscoring its transformative potential while addressing persistent challenges. By integrating empirical evidence from domestic and international studies, the review highlights DDL's efficacy in enhancing writing accuracy, fluency, and learner autonomy through authentic corpus engagement and inductive discovery. Key findings reveal that DDL fosters deeper linguistic awareness, particularly in collocation use and error correction, aligning with the needs of L2 writers in academic and specialized contexts.

Notable disparities exist between research trajectories: Chinese scholars emphasize automated feedback systems and error correction mechanisms, while international studies focus on lexical-grammatical development and feedback design. Despite these advancements, critical gaps remain, including limited corpus accessibility for ESP writing, variable teacher expertise, and understudied learner factors (e.g., motivation, proficiency levels). Methodological constraints, such as small sample sizes and short intervention durations, also hinder the generalizability of findings.

We proposes an integrative framework for DDL implementation, advocating for collaborative efforts between computational linguists and language educators to refine corpus tools and design user-friendly interfaces. Future research must prioritize large-scale longitudinal studies to assess sustained effects, explore personalized DDL models leveraging AI and big data, and develop targeted teacher training programs to bridge the technology-pedagogy divide. By addressing these gaps, DDL can evolve from a niche methodology to a mainstream tool in data-enhanced language education, empowering learners to navigate authentic linguistic environments and fostering more effective, learner-centered writing instruction.

Acknowledgments

I gratefully acknowledge the authors of all cited works whose contributions laid the foundation for this research.

References

[1] Crosthwaite P. Retesting the limits of data-driven learning: feedback and error correction [J]. Computer Assisted Language Learning, 2017, 30(6): 447-73.

[2] Crosthwaite P, Wong L L C, Cheung J. Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning [J]. ReCALL, 2019, 31(3): 255-75.

[3] Crosthwaite P, Storch N, Schweinberger M. Less is more? The impact of written corrective feedback on corpus-assisted L2 error resolution [J]. Journal of Second Language Writing, 2020, 49.

[4] Gan Min, Zou Ling. English Writing Teaching under the DDL Model [J]. Educational Academic Monthly, 2010, (5): 106-107.

[5] Larsen-Walker M. Can Data Driven Learning address L2 writers' habitual errors with English linking adverbials? [J]. System, 2017, 69: 26-37.

[6] Liu Ping, Wu Liangping, Liu Liya. A Study on the Application of CQPweb in ESP Writing Teaching [J]. Foreign Language World, 2016, (5): 11-19.

[7] Lusta A, Demirel Ö, Mohammadzadeh B. Language corpus and data driven learning (DDL) in language classrooms: A systematic review [J]. Heliyon, 2023, 9(12).

[8] Muftah M. Data-driven learning (DDL) activities: do they truly promote EFL students' writing skills development? [J]. Education and Information Technologies, 2023, 28(10): 13179-205.

[9] Pérez-Paredes P. A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015 [J]. Computer Assisted Language Learning, 2019, 35(1-2):36-61.

[10] Şahin Kızıl A. Data-driven learning: English as a foreign language writing and complexity, accuracy and fluency measures [J]. Journal of Computer Assisted Learning, 2023, 39(4): 1382-95.

[11] Satake Y. How error types affect the accuracy of L2 error correction with corpus use [J]. Journal of Second Language Writing, 2020, 50.

[12] Sun W, Park E. EFL Learners' Collocation Acquisition and Learning in Corpus-Based Instruction: A Systematic Review [J]. Sustainability, 2023, 15(17).

[13] Sun X, Hu G. Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class [J]. Language Teaching Research, 2020, 27(3): 660-88.

[14] Tian Zhen, Peng Yajing. Research Progress in Computer-Assisted Language Learning under the Background of Artificial Intelligence (2011—2021) [J]. Foreign Language World, 2022, (3): 53-60.

[15] Xu Jinfen, Liu Wenbo. Innovative Foreign Language Teaching and Research in the Context of Information Technology [J]. Foreign Languages and Their Teaching, 2019, (5): 1-9+147.