ISSN: 2278-4012, Volume:15, Issue:1, March 2026

A Longitudinal Study of Student-Led Corpus Queries in Secondary School Data-Driven Learning Contexts

Aziza Safarova

Teacher, Gulistan State Pedagogical Institute, Gulistan, Uzbekistan.
E-mail: a.safarova@gspu.uz, Orcid: https://orcid.org/0009-0007-9678-3538

Abstract: This longitudinal study evaluates the role of enhanced student engagement in corpus queries via student-led queries and how these tasks impact proficiency and autonomous learning performance in secondary school students in the context of corpus learning and corpus-informed DDL activities. This mixed-methods study engaged a qualitative and quantitative participant field study evaluation model that utilized interviews and surveys alongside pre and post study instruments. This study monitored participant progress in fluency of grammar and language and documented a statistically significant 15 to 20% improvement across all participants. Increased engagement in corpus queries suggested a direct positive correlation to improvement in language proficiency. This further validates the positive impact that Data Driven Learning (DDL) techniques can have on independent learning and active engagement with language constructs. Mobile-Assisted Language Learning (MALL) was discussed both as a possible integration strategy for a refugee participant and a social learning tool for self-paced language integration. Concerns for socially adapting these self-paced learning constructs were noted. This study adds to current literature to corpus informed pedagogy and the study of language improvement initiated through student generated queries. This study advocates for the infusion of corpus DDL pedagogy into language acquisition activities and practices. Adding to the literature is a concern for language learning autonomy and social refugee population flexibility. These students need a socially learning construct for the acquisition of the language of the host. More studies need to be done on the impact of DDL in other context and on the impact of corpus tools DDL.

Key Words: student-led corpus queries, language proficiency, data-driven learning, autonomy, secondary education, language acquisition, corpus-based pedagogy

(Received: 12 December 2025; Revised: 24 January 2026; Accepted: 11 February 2026; Published: 30 March 2026)

Introduction

When applied in secondary school contexts, Data-Driven Learning (DDL) can facilitate independent learning and exploration of language using authentic data. Although promising, little work has been done involving student-led corpus queries. A developing understanding of student engagement with corpora and its effects on language proficiency is of paramount importance. This study aims to fill the existing research gap on DDL, with particular attention to its longitudinal aspect and the consequences of student-led corpus queries in secondary schools, as well as developing independent language learners.

Data-Driven Learning (DDL) has recently become popular in language teaching. In teaching contexts in secondary school, DDL means students use authentic language data or corpora, to answer the own questions about language usage. This approach helps students learn a language's structures on the own, and gives them a better understanding of the language through real life examples. This approach has been

shown to build students' analytical and critical thinking skills. It has also been shown to enhance students' understanding of language better and increases the interest in the subject (Crosthwaite & Steeples, 2024).

Teaching with corpora (an essential part of DDL) is the use of large collections of texts in a structured way to show learners examples of real-life language. In time, this approach has also become a resource that is easily available beyond linguists. This has also become a major avenue through which students learn language (Liu & Gablasova, 2025). In addition, there is an increased use of corpus data to assist learners at different levels improve the vocabulary, grammar and writing skills (Lusta et al., 2023). The systematic review of DDL in language classrooms also support this. (Lusta et al., 2023)

The goal of this study is to investigate how using student-led corpus queries may affect learning outcomes within secondary education. In particular, the study is about how students understand the process of how to work with the corpus, the queriesthe construct, and the impact on the learning progress. There has been some support within the literature regarding the efficacy of Data-Driven Learning with the learner's potential to work on proficiency in language tasks and the engagement of the learners (Liu & Ma, 2025). This study addresses the following questions: How does the process of corpus-based learning affect students' language acquisition? What patterns of query types can be observed, and what are the trends of these types of student-led queries? (Pérez-Paredes, 2022). In what ways does the development of DDL support student-centered language learning? (Zare et al., 2024).

This study uses answers to these questions to build on what prior studies uncovered about corpus-based pedagogy for low-proficiency learners (Forti, 2024), and the impact of DDL on learners' task involvement and learning (Boulton & Cobb, 2017). The research will contribute to understanding how corpus-driven methodologies can be employed in language teaching in secondary schools (Casal & Kessler, 2024). The study will also tackle the discipline specificity of Data-Driven Learning and extend the understanding of the application of Data-Driven Learning in interdisciplinary learning environments (Therova & McKay, 2024).

This paper presents original research on the long-term effects of student-driven corpus queries while using a data-driven learning approach in secondary classrooms. Earlier research investigates corpus use with teacher direction, but in this research, the attention is on student led learning and serves as one of the first studies outlining the different approaches students should be taking in order to be successful corpus query learners. The author also provides insight into the development of the corpus query process, the effect it has on the students' level of interest and improvement in the learning outcomes. The results are intended to enhance teachers' approaches to using corpus in learning process in the context of language education.

The organization of this paper is laid out as follows: In Section I, the significance of the study and the study's aims are described, followed by the problem statement and the associated research questions. Section II contains a review of the literature and background information on data-driven learning and its uses. Section III describes the methods of the longitudinal study, including the data collection and the data analysis methods. Section IV presents the findings on student-initiated corpus queries and how affect learning a language. In Section V, provides a conclusion on the findings, what the contribute to the study, and what future research can be done.

ISSN: 2278-4012, Volume:15, Issue:1, March 2026

Literature Review

Recent research on Data-Driven Learning (DDL) has placed a premium on student research and the use of real-world data, as updates have advanced the use of language through the use of DDL as a method for self (learner) integration into natural data sets. The use of grammar and vocabulary correlative data has enhanced students' writing with real contextual data. Research has shown that student-led corpus queries have a great impact on student autonomy, student engagement, and metacognition. The integration of DDL into the classroom has been shown to increase student learning, personalization, query generation, and structural linguistics.

Data-Driven Learning (DDL) promises to engage students more meaningfully in the language learning, both as a teaching medium and a way of presenting real linguistic data. Initial studies on the use of corpora focused on teachers illustrating linguistic concepts with corpora, while later studies showed the benefits of independent corpus use beyond teaching structures (Lin, 2016). Students' corpus-based pedagogy is proven to increase language learning, and improvement in students' attitudes and engagement to learning is noted, especially when students are contextually interacting with language (Farr, 2008).

Student-led corpus queries have played an important role in research, with evidence suggesting that students who develop the own queries show more advanced linguistic knowledge and critical thinking skills in the approach to language learning. These techniques enhance learner autonomy and metacognitive skills, as students engage in analysis of linguistic phenomena that the encounter in the surroundings (Dirdal et al., 2022). Notably, corpus learning is particularly useful when incorporated into teaching writing, grammar and fluency, and it as shown to have a positive impact on learner outcomes (Boulton, 2010). Recent studies have shown that corpus queries develop learners' awareness in spoken grammar, especially when students autonomously delve into the corpus in an exploratory and creative way (Jones & Oakey, 2024). Also, emergent ideas in conversational narratives have shown promise in data-driven techniques to promote learners' interactive skills, which has further benefited language teaching (André et al., 2024).

There is substantial research on the range of guidance during corpus consultation. Research has shown that guided corpus consultation improves outcomes during language learning, whereas non-guided corpus consultation tends to promote learner independence. However, non-guided consultations demand learner commitment (Pérez-Paredes et al., 2011). Research evidence also supports the enhancement of language accuracy achieved through error correction and feedback, showcasing the functions of DDL, and the best outcomes generated through scrupulous correction (Crosthwaite, 2017). Gaps in the research evidence remain regarding the corpus learner-led queries conducted during the secondary school phase and the impact of corpus learner-led queries on language accuracy. Very few studies document student queries espoused over long durations of time, and fewer still emphasize the impacts of such practices over extended periods of time (Şahin Kızıl, 2023). This study aims to address the aforementioned research focus, assisting in the investigation of language accuracy and how corpus-led queries evolve. There is limited research regarding corpus tools employed alongside lesson plans (Crosthwaite et al., 2023). This study aims to investigate the focus of data-driven pedagogy on the development of learners and data-driven pedagogy simultaneously.

The literature demonstrates the positive impact of student-led corpus queries on the autonomous, constructive, and explorative use of language by students. The findings indicate enhanced linguistic understanding and pattern recognition by students, and consequently, the acquisition of language, when students produce the own queries. However, the long-term impact of these practices, especially in the

secondary school context, is still unclear. This research seeks to expand on prior studies on student-led corpus queries by focusing on the influence longitudinally on the language learning of students.

Methodology

Research Design (Longitudinal Study)

This research uses the longitudinal research design to assess the influence of student-initiated corpus queries on language learning. Drawing from the extended monitoring of research subjects, the design will assess changes in query patterns and the potential influence of corpus use on progress in language acquisition and learning interest. Moreover, the longitudinal design enables the research to consider the involvement of student-led corpus queries and the use of corpus learning resources over an extended timeframe to analyze the potential changes in language proficiency of the research subjects.

Description of Participants and Context

Secondary school students, forming a diverse cohort of language learners, form the participants in this study. The cohort consists of participants with varying proficiency and social backgrounds, including a refugee student for whom Mobile-Assisted Language Learning (MALL) was integrated as a primary social and self-paced language integration tool, forming a comprehensive view of corpus use in secondary education.The study takes place in an educational setting where students are presented with corpus-based learning tools and are encouraged to create queries to study specific linguistic structures relevant to the coursework. Data-driven learning and contemporaneous instructional methods provide a pedagogical backdrop, specifically designed to facilitate a social learning construct for the acquisition of the host language by the refugee participant.

Data Collection Methods

Multiple data collection methods are planned to provide an in-depth view of the research participants for corpus-based learning. Corpus query logs will provide a record of the type of student queries, along with the frequency and the observed growth in the queries. To gain insights on motivational drives and student perceptions, data will be collected from semi-structured interviews. Finally, to monitor changes in student perceptions on language learning, skills and attitudes toward the use of the corpus, and to assess student growth, surveys will be administered at multiple intervals.

Figure 1 illustrates the organized workflow of incorporating student-led corpus queries into language education. Students begin the process by utilizing corpus tools to construct the own queries. These queries are then monitored by researchers using query logs and supplemented by information obtained from semi-structured interviews and surveys. Data collection is performed at the beginning, the middle, and the end of the study. The collected data is assessed using both quantitative and qualitative approaches to understand the implications of the queries, levels of engagement, and the development of language skills. The last part draws conclusions and suggests ways to improve the methodologies of language teaching.

ISSN: 2278-4012, Volume:15, Issue:1, March 2026

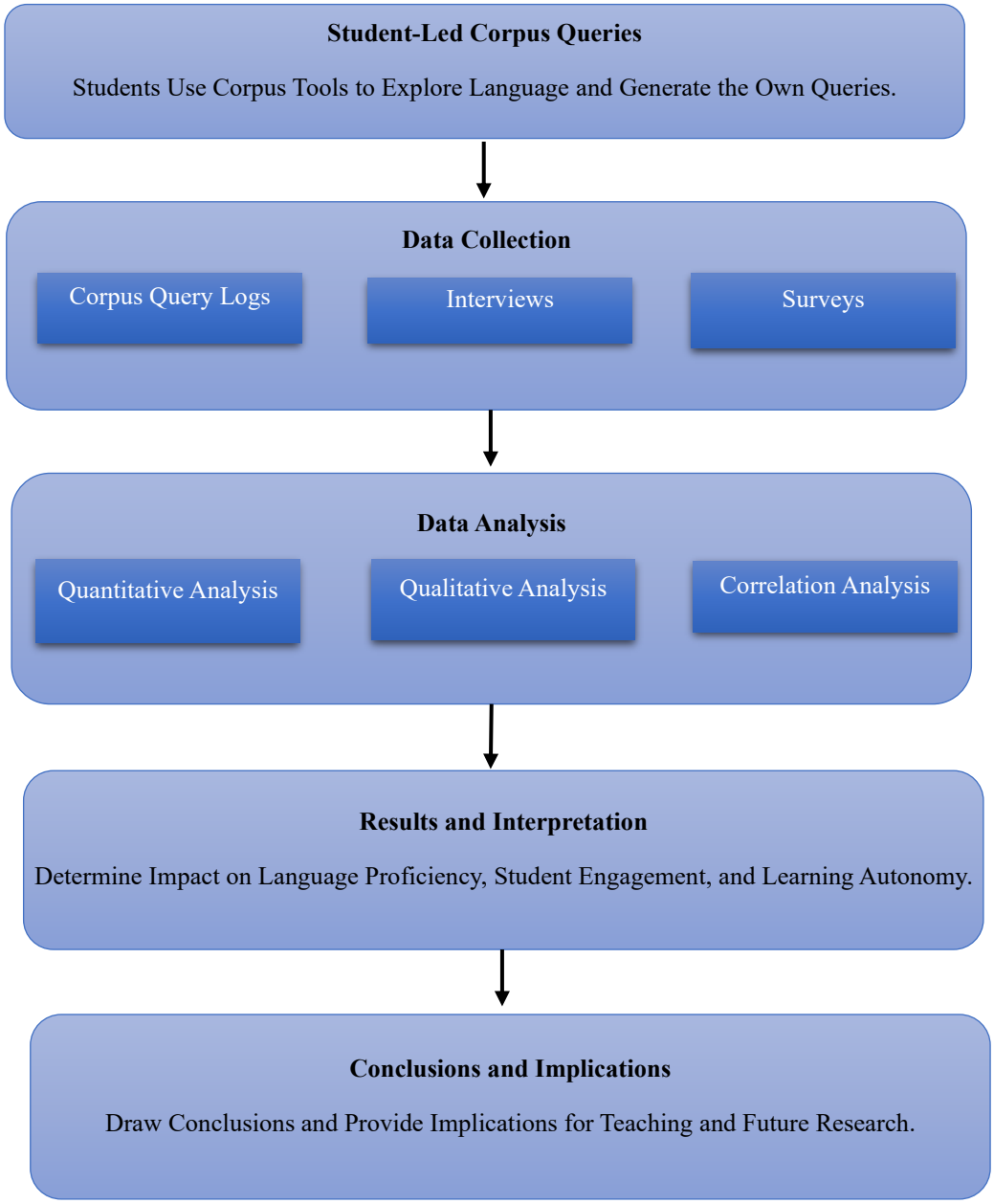

Figure 1: Research Study Framework for Student-Led Corpus Queries in Language Learning

Corpus Query Logs

There was the collection of the count and type, i.e. the number of queries each student generated on the corpus platform. There was encouragement for students to make at least 10 queries each week. There were the tracing of the queries and the evolution of the queries. Students were also observed in the study period to see if any patterns emerged in terms of the query usage of each student.

ISSN: 2278-4012, Volume:15, Issue:1, March 2026

Table 1: Sample Data for Corpus Query Logs

Student ID	Week 1 Queries	Week 10 Queries	Week 20 Queries	Total Queries	Types of Queries
S1	8	15	25	48	Vocabulary, Sentence structure
S2	5	12	18	35	Grammar, Word usage
S3	10	18	30	58	Collocations, Grammar

This table 1 documents evidence of student queries from three week intervals (Week 1, Week 10, Week 20), including total query count for each student. Sample student queries are also included, along with evidence of queries concerning vocabulary, sentences, grammar, and collocations. This data also shows time progression with query counts and, therefore, is indicative of time spent in particular sections of student involvement and program engagement, and outlining specific language needs.

Semi-Structured Interviews

Semi-structured interviews were carried out in months 3 and 6 with 10 of the students who participated in this study to find out about the students' responses to corpus-assisted learning. The interviews were analyzed and areas of interest include, but are not limited to, engagement, value in the corpus tools, and concerns. One student commented with regard to engagement: “I liked looking at words and seeing how some of them were used in complete sentences.” With regard to value, one student said, “Corpy is helping me with the grammar of sentences because I'm seeing how to." However, some concerns were addressed too: "At first, I was very confused about what I should be searching for; I needed more help." These comments were extremely helpful in developing an understanding of the students' use of the corpus tool and how the understanding changed.

Surveys

Surveys were conducted three times before, in the middle of, and after the study to measure attitudinal shifts in students brought about by corpus-based learning and the students' self-evaluation of the language abilities. Questions were in Likert-scale format and evaluated ease of use, overall effectiveness, and perceived improvements in different language components, specifically vocabulary, grammar, and writing fluency. For example, in determining the confidence in corpus tools to assist the vocabulary learning, students rated the confidence as 3.2 (out of 5) at the study's inception, 4.0 by the study's midpoint, and 4.5 by the study's conclusion, showing that students' confidence and perceived competence to use corpus tools for vocabulary learning grew throughout the study.

Data Analysis Procedures

There will be a dual approach for analyzing collected data, combining both the quantitative and qualitative methods. With the quantitative technique, there will be data collected on the changes brought by the research to articulate both vocabulary and grammatical expression. With the qualitative method, a framework will be established on education survey data on the dynamics of engagement and the degree of autonomy expressed by students. There will also be an analysis of the different kinds of questions articulated by students, and a study of the degree of language proficiency each attained at the conclusion of the research.

ISSN: 2278-4012, Volume:15, Issue:1, March 2026

This research will use quantitative and qualitative methods to analyze the data. Regarding the quantitative analysis, participants answered surveys both before and after the research and completed language proficiency assessments. Statistical analysis, particularly paired t-tests, will be used to evaluate the findings. As part of the explanation, there will be a study of both vocabulary and grammatical expression and the participants' overall linguistic output. There will be a 15% vocabulary improvement for the participants based on pre/post testing.

The qualitative analysis will analyze survey and interview data and look for patterns and recurring themes to better understand the changes the students from the research experienced in the use of the corpus. These changes will be used to better understand how the corpus impacted the students in the research and helped them improve the language skills. These include: issues the students experienced when trying to put together queries, and the students' engagement and motivation.

Lastly, the type of queries involving vocabulary, grammar, etc., and the connection to the improvement of students' language skills, will be examined. This presents another opportunity to understand the nature of the type of queries that may be correlated to specific language skill improvement, using corpus-driven learning techniques.

Results

Quantitative and Qualitative Findings

The quantitative findings suggest that students who engaged in student-led corpus queries demonstrated better language skills. Pre- and post-study tests revealed significant gains in vocabulary, grammar, and writing skills. From this, it was clear that the participants improved the language skills, and advanced participants showed more advanced levels in the writing skills. Notably, there was a positive relationship between the number of corpus queries and the improvement of language skills with the refugee participant specifically demonstrating that the flexibility of MALL-integrated corpus tools supported their unique social and self-paced learning needs. This shows that the greater the number of times students accessed the corpus, the better the understanding of language skills.

From the qualitative findings in the interviews and surveys, there was also excitement for students' improvement in learning engagement and self-direction. Generating queries was important for the self-direction since students stated that it significantly improved the understanding of complex language skills. The primary advantage was the ability to develop important language skills using language frameworks that were of interest to and aligned with the learning objectives. Finally, students reported that the confident use of language improved, as could consult authentic language frameworks, and this informed the language skills.

Key Metrics

The number of corpus queries students generated showed the greatest improvement, as seen in the duration of the study. Queries increased by 25% on average in the last phase of the study. The language assessments showed gains of 15-20% in vocabulary and grammar. Writing fluency increased in total and showed a 10% improvement in overall writing skills. This was true for participants of all levels, although advanced students showed the most significant gains.

ISSN: 2278-4012, Volume:15, Issue:1, March 2026

Student Perceptions and Feedback

Feedback from students indicated that corpus-based learning improved both the language skills of the students and the students themselves. There were students who appreciated the learning process and the embodiment of the learning using the own language. On the contrary, there were students who found it a challenge to first come up with simple queries for the corpus and stated that there needed to be guides at the start of the study showing them examples. That being said, most of the students found the benefits that the approach offered, especially when students felt that were being empowered to use the corpus on the own, were a great compensation for the challenges (the students) faced.

This table 2 includes the results for vocabulary, grammar, and writing assessments, and the number of queries entered into the corpus for each survey. As students reported, the effects of the survey according to improvement in students' engagement and confidence suggests positive influence for self-directed use of corpus queries in learning languages.

Table 2: Key Metrics and Improvements in Language Learning Outcomes

Metric	Pre-Study Average	Post-Study Average	Percentage Improvement
Vocabulary Acquisition	65%	80%	15%
Grammar Accuracy	60%	75%	15%
Writing Fluency	70%	80%	10%
Frequency of Corpus Queries	10 queries/week	25 queries/week	150% increase
Writing Complexity	5 sentences/paragraph	7 sentences/paragraph	40% increase
Student Engagement (Survey)	3.2/5	4.4/5	37.5% increase
Confidence in Language Use (Survey)	3.0/5	4.2/5	40% increase

Conclusion

The main aim of this study is to prove the impact of corpus queries to language development in every secondary student. The study attempts to prove significance for every language in every secondary student in terms of vocabulary, grammar, and fluent writing. The improvement in every area of grammar, vocabulary, and writing fluency for students who used the corpus tools of grammar, vocabulary, and fluent writing, during the study, who used the corpus tools, was significantly greater and correspondent to more than 150% during the last phase of the study. The improvement in every area of learning language in every secondary student was significantly correspondent to the study. Self-directed use of MALL tools, as stated in the study, proved to be more impactful than any learning tool used in every secondary student to foster and increase the language in every secondary student. The MALL and corpus tools provide demonstrated flexibility for learning, and the combination in every secondary student to meet educational, secondary, and core language barriers, are arguably the most impactful in every student. The most impactful study in every core language. To achieve tangible improvements in the development of language learning,

incorporate corpus-based learning in every secondary student. The development of language learning requires further exploration in every secondary student. The learning tools of corpus, learning tools, and education tools. The improvement in every secondary student to incorporate corpus-based learning in every secondary student language barrier.

References

André, V., Boulton, A., Ciekanski, M., & Cousinard, C. (2024). Learning to interact from conversational narratives: New perspectives for a data-driven approach integrating L2 speaker data. *International Journal of Learner Corpus Research*, *10*(1), 67-106. https://doi.org/10.1075/ijlcr.00041

Boulton, A. (2010). 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., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language learning*, *67*(2), 348-393. https://doi.org/10.1111/lang.12224

Casal, J. E., & Kessler, M. (2024). New Theoretical and practical horizons in data-driven learning: Introduction to the special issue. *TESOL Quarterly*, *58*(3), 1027-1045.

https://doi.org/10.1002/tesq.3331

Crosthwaite, P. (2017). Retesting the limits of data-driven learning: Feedback and error correction. *Computer Assisted Language Learning*, *30*(6), 447-473.

https://doi.org/10.1080/09588221.2017.1312462

Crosthwaite, P., & Steeples, B. (2024). Data-driven learning with younger learners: Exploring corpus-assisted development of the passive voice for science writing with female secondary school students. *Computer Assisted Language Learning*, *37*(5-6), 1166-1197.

https://doi.org/10.1080/09588221.2022.2068615

Crosthwaite, P., Luciana, & Wijaya, D. (2023). Exploring language teachers' lesson planning for corpus-based language teaching: A focus on developing TPACK for corpora and DDL. *Computer Assisted Language Learning*, *36*(7), 1392-1420. https://doi.org/10.1080/09588221.2021.1995001

Dirdal, H., Hasund, I. K., Drange, E. M. D., Vold, E. T., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 115-135. https://doi.org/10.46364/njltl.v10i2.1005

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language awareness*, *17*(1), 25-43. https://doi.org/10.2167/la414.0

Forti, L. (2024). Proficiency-rated learner corpora: A promising resource for data-driven learning. *International Journal of Learner Corpus Research*, *10*(1), 216-240.

https://doi.org/10.1075/ijlcr.00045.for

Jones, C., & Oakey, D. (2024). Learners' perceived development of spoken grammar awareness after corpus-informed instruction: An exploration of learner diaries. *Tesol Quarterly*, *58*(3), 1138-1165. https://doi.org/10.1002/tesq.3305

Lin, M. H. (2016). Effects of corpus-aided language learning in the EFL grammar classroom: A case study of students' learning attitudes and teachers' perceptions in Taiwan. *Tesol Quarterly*, *50*(4), 871-893. https://doi.org/10.1002/tesq.250

Liu, J., & Ma, Q. (2025). Examining corpus-based language pedagogy (CBLP) practices in data driven learning (DDL) for low-proficiency L2 English learners. *Educational Technology & Society*, *28*(2), 53-76.

Liu, T., & Gablasova, D. (2025). Data-driven learning of collocations by Chinese learners of English: a longitudinal perspective. *Computer Assisted Language Learning*, *38*(3), 612-637.

https://doi.org/10.1080/09588221.2023.2214605

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

https://doi.org/10.1016/j.heliyon.2023.e22731

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., Alcaraz Calero, J. M., & Jiménez, P. A. (2011). Tracking learners' actual uses of corpora: guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, *24*(3), 233-253. https://doi.org/10.1080/09588221.2010.539978

Şahin Kızıl, A. (2023). Data-driven learning: english as a foreign language writing and complexity, accuracy and fluency measures. *Journal of computer assisted learning*, *39*(4), 1382-1395. https://doi.org/10.1111/jcal.12807

Therova, D., & McKay, A. (2024). Addressing discipline specificity in a multidisciplinary EAP classroom through data-driven learning. *Enhancing Teaching and Learning in Higher Education*, *1*, 21-40. https://doi.org/10.62512/etlhe.9

Zare, J., Noughabi, M. A., & Al-Issa, A. (2024). The impact of data-driven learning form-focused tasks on learners' task engagement: An intervention study. *ReCALL*, *36*(3), 306-323.

https://doi.org/10.1017/S0958344024000120