ACORP special issue: Corpora and AI for inductive learning: Theory and practice

Introduction

The application of language corpora in education has gained increasing prominence over the past three decades, most notably through the development of Data-Driven Learning (DDL)—a pedagogical approach that leverages authentic language databases and promotes inductive learning (Boulton & Vyatkina, 2021; Johns, 1991). Inductive learning has been studied extensively in education research as an umbrella term that encompasses a range of instructional methods including inquiry learning, problem-based learning, project-based learning, case-based learning, discovery learning, and more. What distinguishes inductive and deductive learning is the sequence of instructions. While deductive learning starts with an explanation of general principles followed by learners applying them to specific cases, inductive learning begins with specific observations, problems, and cases, and learners construct, formulate, and discover principles and rules in the process of analyzing or interpreting these specific cases (Prince & Felder, 2006). Inductive learning is rooted in constructivism which holds that learning occurs when individuals construct and reconstruct knowledge through experience. While the first four inductive methods mentioned above are commonly used in engineering and science curricula (see Prince & Felder, 2006, for a detailed discussion), language education has mainly adopted discovery learning especially for grammar instruction (see Ellis & Shintani, 2013, for deductive vs. inductive presentation). Given the *Applied Corpus Linguistics (ACORP)* journal's relevance to language learning and teaching, the term "inductive learning" in this special issue refers to discovery learning (Bruner, 1960), where instances of language use are presented to learners to observe and to identify patterns and to make generalizations of the language, versus rule explanation illustrated with language examples in deductive instruction.

DDL has been consistently linked to inductive and discovery learning and is supported by Schmidt's (1990) noticing hypothesis, which posits that conscious awareness of language features is essential for acquisition. Over the past three decades, DDL has emerged as a powerful pedagogical approach in language education grounded in the use of authentic language corpora. Based on the extent to which students are scaffolded and given direct access to corpus data (i.e., large language databases), three main approaches to corpus-based pedagogy have developed: corpus-informed teaching, integrated corpus-supported teaching, and self-directed DDL (Jablonkai & Csomay, 2022; Szudarski, 2025). In line with findings on inductive language learning (e.g., Cerezo et al., 2016; Leow, 2015), DDL and corpus-based pedagogies have been found to enhance language awareness, autonomy, and critical thinking by encouraging learners to explore real-world language texts, to observe instances of linguistic phenomena, and to discover relevant linguistic patterns (Jablonkai & Csomay, 2022).

Despite the strong potential of DDL, its implementation in classroom contexts often presents instructional, cognitive, and technical challenges that limit its accessibility and effectiveness for both learners and educators. A recent qualitative review of barriers to DDL by

Sun and Mizumoto (2025) identified several inherent and external factors (e.g., tools and approaches, learner capacity and beliefs) that make DDL challenging. A notable barrier is the technical complexity of corpus tools, many of which were originally developed for research purposes rather than classroom use. This mismatch often limits their accessibility for both teachers and learners as empirical studies on DDL often reveal (Pérez-Paredes, 2019). Learners typically require substantial time and scaffolding to use corpora autonomously, and many report feeling overwhelmed by the volume and density of linguistic data. Furthermore, most corpora are better suited to learners with higher proficiency levels, leaving beginners at a disadvantage. Issues such as computer anxiety—or even corpus-specific anxiety—can further inhibit engagement and confidence (Ortega, 1997). Indeed, in a recent meta-analysis on DDL for low-proficiency L2 English learners, Liu and Ma (2025) reported that DDL tends to be more effective in enhancing lexicogrammatical knowledge, when using paper-based approaches (i.e., corpus-informed teaching with no direct interaction with language corpora), and being supported by teacher guidance and feedback.

Beyond technical barriers in DDL, the challenges are largely embedded in the nature of inductive learning, given the heavy demands on working memory during the active and deep cognitive processing (Kirschner et al., 2006). Studies in second language acquisition have yielded mixed findings on the effects of inductive instruction, compared with deductive instruction, on the development of grammar (e.g., Erlam, 2003; Herron & Tomasello, 1992; Moranski & Zalbidea, 2022; Seliger, 1975; Shaffer, 1989), vocabulary (e.g., Kaur & Hegelheimer, 2005; Tsai, 2019) and pragmatic competence (e.g., Qi & Lai, 2017; Takimoto, 2008). The inconclusive findings could be due to the complexity of the target knowledge (Qi & Lai, 2017; Takimoto, 2008), the learner's analytical ability (Erlam, 2005), and the variation in the degree of support provided in inductive instruction. As pointed out by Erlam (2003), prior studies showing inconsistent results differed in the process of inductive instruction, such as whether learners were guided explicitly to look for rules or patterns and whether metalinguistic explanation or feedback was provided in the end. That is to say, for learners who are not good at inferring patterns from observations to learn complex structures, additional support may help avoid cognitive overload which could hamper the learning process (Chen et al., 2023; Gorbunova et al., 2023). Indeed, guided induction as an enhanced version of a pure inductive method, where instructors support the rule formulation process by increasing the salience of the patterns illustrated in materials directing learners' attention to the patterns, and providing hints for induction, has been confirmed to be more effective than a pure deductive method (e.g., Cerezo et al., 2016; Haight et al., 2007; Lai et al., 2020; Smart, 2014; Vogel et al., 2011). Given the challenges of inductive learning and the barriers associated with DDL alongside the demonstrated effectiveness of guided approaches and recent advances in technology (e.g., generative AI) merit closer examination for their potential to provide scaffolding and serve as the "more knowledgeable other" in DDL.

AI and corpora for inductive learning

Recent developments in technology have witnessed the emergence and widespread use of new tools that are based on large language databases. Generative AI tools (such as Large Language Models – LLMs) offer new pedagogical affordances. Their natural language interfaces and adaptive capabilities make them highly accessible and enable the generation of learner appropriate input and guidance. Concerns, however, persist regarding the lack of transparency in data sources, algorithmic bias, ethical implications, and the risk of promoting passive learning through machine-generated content (Crosthwaite & Basia, 2023; Mizumoto, 2023).

Leading scholars in DDL and corpus-based pedagogies (e.g., Crosthwaite & Basia, 2023; Mizumoto, 2023), have advocated for an integration of corpus-informed DDL with generative AI technologies in language education. This integration can be implemented at two main levels: at the level of tools and at the level of pedagogical application. At the tool level, this could mean that corpus analysis software provides access to LLMs (e.g., ChatGPT) or integrate AI-powered functionalities (e.g., CorpusChat in Cheung & Crosthwaite, 2025). The latest version of AntConc (Anthony, 2024) and the Contemporary Corpus of American English tools (Davies, 2025), for instance, incorporate direct access to ChatGPT, enabling users to interact with generative AI within a corpus analysis environment. At the pedagogical level, this integration opens possibilities for sequencing or combining corpus-based and AI-driven tools in language classrooms (e.g., Kartal & Korpuz, 2025; Mizumoto, 2023; Pérez-Paredes et al., 2025). Recent studies collectively emphasize that AI should not be viewed as a replacement for corpus tools or data-driven approaches but as a complementary resource that can enhance the effectiveness of DDL and corpus-based pedagogy. In most cases, AI tools function as a mediating tool, offering scaffolding that supports learners in navigating and interpreting corpus data more effectively. As the field is still in its infancy, further research is needed to explore how AI can meaningfully complement inductive corpus-based language learning, while also examining the evolving roles of teachers, identifying emerging professional development needs, developing robust pedagogical frameworks for effective integration, and addressing the key ethical considerations.

In this ACORP special issue

In this *ACORP* special issue, contributors explore the rapidly evolving intersection of corpus linguistics and generative artificial intelligence (GenAI) in language learning, teaching, and research. As GenAI tools gain prominence alongside established corpus-based approaches, new pedagogical opportunities and challenges emerge for learners, teachers, and researchers alike. The articles in this issue collectively examine how corpora and GenAI can be integrated to enhance language analysis, genre awareness, writing development, and instructional design. Together, they offer critical insights into how these complementary technologies can inform DDL, promote critical digital literacies, and reshape the future of language education and applied linguistics research.

In the first article in the special issue titled *Not so fast? A comparative study of pre-service teachers' language analysis, lesson planning and materials development using corpora*

and generative artificial intelligence, Leńko-Szymańska discusses the incorporation of corpora and GenAI in language teacher education. She contrasts corpus-based approaches that promote DDL and GenAI tools that afford immediate and user-friendly access to linguistic data. In her study, she compares the use of corpora and GenAI by pre-service teachers in pedagogically oriented language analysis, lesson planning, and materials development. She found that, while students preferred AI tools for their efficiency, they preferred corpus-based tools for their pedagogical importance and worth. She concludes that both resources are useful, however, for different purposes, and therefore, they should complement each other rather than replace one another.

In the second article, titled *Raising genre awareness through visualizing language features*, Blake and Mozgovoy introduce an open-access AI-powered tool, called the Feature Visualizer, and report on students using the tool to better understand the rhetorical structures of articles in computer science. The tool "houses an annotated corpus of scientific research articles written by computer science majors and allows learners to explore authentic texts using on-demand visualizations and multimodal explanations". Learners are then able to engage with the texts as they identify recurring language patterns and rhetorical patterns through highlighted and annotated texts (visualization) discovering genre conventions. The authors tested their hypothesis with a group of Japanese undergraduate computer science majors and found that the "tool enhanced learners' awareness of academic writing conventions and genre features".

In the third article, titled *A comparative analysis of AI-generated texts, corpus data, and speaker judgments: subject honorification patterns in Korean*, Jung and Kim explore the extent to which second language pragmatics instruction could be more efficient with corpora and AI tools. While technologies (e.g., GenAI) afford naturally occurring and authentic communication settings, the authors explore pedagogically appropriate language input and their application in existing pedagogies (e.g., DDL). More specifically, they investigate the appropriateness of examples drawn from a corpus housing transcripts of spoken texts and a GenAI tool in the context of Korean pragmatics. They found that ChatGPT favored a subject-verb matched form while corpus data "reflected the highly complex, context-dependent use and variation of honorifics". Results showed that Korean first language speakers' judgments lined up with corpus findings indicating more context-sensitive judgements while second language speakers' acceptability judgments reflected ChatGPT's patterns with no such sensitivity beyond the matched forms.

The next article by Begmatova and Saydazimova titled *Enhancing learners' academic writing skills: A comparative analysis of traditional and AI-assisted instructional approaches* explores the impact of AI-assisted approaches to teaching academic writing, literature review in particular, to first year students in an English Medium Instructional context in Uzbekistan. A learner corpus of student writing was compiled including 306 literature reviews produced in two classes: one following a traditional and another following an AI-assisted instructional approach. The analyses highlighted that the AI-assisted approach resulted in higher semantic complexity, referential cohesion, and lexical diversity in student papers as well as qualitatively different use

of referencing conventions, integration of cohesive devices, and demonstration of argumentation and critical analysis.

The fifth article by Chitez, Csürös, and Rogobete titled *Upgraded literacy: teacher training approaches to integrating corpus data and AI tools for school text readability adaptation* looks at how teachers employed AI and corpus-based tools to simplify texts for better readability for second language learners. These simplifications consisted of lexical substitution, syntactic restructuring, and vocabulary. Results showed that AI tools helped in reducing linguistic load, generating instructional content, and offering readability feedback. "Teachers reported increased confidence in adapting texts and greater awareness of linguistic dimensions influencing accessibility. The findings suggest that structured training in digital linguistic tools can enhance teachers' capacity to develop inclusive, readable texts tailored to diverse learner needs".

Finally, the article by Forchini and Murphy titled *iThink, therefore iCheck: ChatGPT and the role of linguists in a classroom using AI* explored the integration of ChatGPT in a graduate course. Their study focuses on how the tool may be able to perform linguistic analysis similar to a multidimensional analysis (MDA) and how the students evaluate that output through inductive learning. They provide ChatGPT's linguistic analyses of dialogs from two movies and examine how students (with or without linguistic training) interact with that analysis while also evaluating the tool and the learning experience. Their findings "reveal that ChatGPT's ability to perform both general and MDA-based analyses was limited, often inconsistent and inaccurate [...] Students with prior MDA training showed stronger data literacy and more critical engagement with the tool, while untrained students exhibited overreliance and misconceptions regarding ChatGPT's capabilities." They conclude that their results call for "targeted instruction to foster analytical skills and reduce uncritical AI use", arguing for AI to be critically integrated into academic settings.

Two short communication pieces are also included in the special issue. Anthony's piece titled *Concordancing with AI: Applications of word and sentence embeddings* outlines ways in which concordance lines could be enhanced with AI tools and how that helps inductive learning. He proposes a word and sentence embedding model that could enhance concordancing searches and the interpretation of results. First, he explains what embeddings are, then he explains how searches could result in more nuanced and contextually relevant results as they capture semantic and contextual information. Then, with three case studies, he illustrates how these embeddings can be used effectively.

Pang's short communication piece titled *Leveraging large language models to supplement corpus-based inductive learning of Chinese as a second language* discusses how corpus tools may be problematic for Chinese as second language learners, and points to solutions integrating AI tools in support of inductive learning. More specifically, they demonstrate how using Sketch Engine and Claude Opus 4 afford the discovery processes in inductive learning while ensuring accurate linguistic input for the learners.

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

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/10.64152/10125/73450

Bruner, J. S. (1960). *The process of education.* Harvard University Press.

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures: An analysis of learning outcomes and processes. *Studies in Second Language Acquisition, 38*(2), 265–291. https://doi.org/10.1017/S0272263116000139

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review, 35*(2), 63. https://doi.org/10.1007/s10648-023-09782-w

Cheung, L., & Crosthwaite, P. (2025). CorpusChat: integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2025.2506480

Crosthwaite, P., & Basia, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Davies, M. (2025, September 2). English-Corpora.org: new AI/LLM features for corpus analysis. *Corpora Listserv.* Retrieved https://www.english-corpora.org/ai-llms/ on October 31, 2025.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research.* Routledge. https://doi.org/10.4324/9780203796580

Erlam, R. (2003). The effects of deductive and inductive instruction on the acquisition of direct object pronouns in French as a second language. *The Modern Language Journal, 87*(2), 242–260. https://doi.org/10.1111/1540-4781.00188

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9*(2), 147–171. https://doi.org/10.1191/1362168805lr161oa

Gorbunova, A., van Merrienboer, J. J., & Costley, J. (2023). Are inductive teaching methods compatible with cognitive load theory? *Educational Psychology Review, 35*(4), 111. https://doi.org/10.1007/s10648-023-09828-z

Haight, C. E., Herron, C., & Cole, S. P. (2007). The effects of deductive and guided inductive instructional approaches on the learning of grammar in the elementary foreign language college classroom. *Foreign Language Annals*, *40*(2), 288–310. https://doi.org/10.1111/j.1944-9720.2007.tb03202.x

Herron, C., & Tomasello, M. (1992). Acquiring grammatical structures by guided induction. *The French Review, 65*(5), 708–718.

Jablonkai, R. R., & Csomay, E. (2022). *The Routledge handbook of corpora and English language teaching and learning*. Routledge. https://doi.org/10.4324/9781003002901

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *ELR Journal, 4*, 27–46.

Kartal, G. & Korpuz, E. (2025). Enhancing the potential of data-driven learning: overcoming the gatekeeping of concordancing through ChatGPT. *Computer Assisted Language Learning.* 1–27. https://doi.org/10.1080/09588221.2025.2569347

Kaur, J., & Hegelheimer, V. (2005). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, *18*(4), 287–310. https://doi.org/10.1080/09588220500280412

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Leow, R.P. (2015). *Explicit learning in the L2 classroom: A student-centered approach.* Routledge. https://doi.org/10.4324/9781315887074

Lai, C., Qi, X., Lü, C., & Lyu, B. (2020). The effectiveness of guided inductive instruction and deductive instruction on semantic radical development in Chinese character processing. *Language Teaching Research, 24*(4), 496–518. https://doi.org/10.1177/1362168818805265

Liu, J., & Ma, Q. (2025). Examining corpus-based language pedagogy (CBLP) practices in data-driven learning (DDL) for low-proficiency L2 English learners: A meta-analysis. *Educational Technology & Society, 28*(2), 53–76. https://doi.org/10.30191/ETS.202504_28(2).RP04

Mizumoto, A. (2023). Data-driven learning meets Generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics, 3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

Moranski, K., & Zalbidea, J. (2022). Context and generalizability in multisite L2 classroom research: The impact of deductive versus guided inductive instruction. *Language Learning, 72*(S1), 41–82. https://doi.org/10.1111/lang.12487

Ortega, L. (1997). Processes and outcomes in networked classroom interaction: Defining the research agenda for L2 computer-assisted classroom discussion. *Language Learning & Technology, 1*(1), 82–93.

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning, 35*(1–2), 36–61. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., Curry, N., & Aguado Jiménez, P. (2025). Integrating critical corpus and AI literacies in applied linguistics: A mixed-methods study. *Computer Assisted Language Learning,* 1–27. https://doi.org/10.1080/09588221.2025.2569351

Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education, 95*(2), 123–138. https://doi.org/10.1002/j.2168-9830.2006.tb00884.x

Qi, X., & Lai, C. (2017). The effects of deductive instruction and inductive instruction on learners' development of pragmatic competence in the teaching of Chinese as a second language. *System, 70*, 26–37. https://doi.org/10.1016/j.system.2017.08.011

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Seliger, H. W. (1975). Inductive method and deductive method in language teaching: A re-examination. *IRAL, 13*(1), 1–18.

Shaffer, C. (1989). A comparison of inductive and deductive approaches to teaching foreign languages. *The Modern Language Journal, 73*(4), 395–403. https://doi.org/10.2307/326874

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, A. X., & Mizumoto, A. (2025). Exploring the barriers to data-driven learning in the classroom: a systematic qualitative synthesis. *Applied Corpus Linguistics*, *5*(2), 100126. https://doi.org/10.1016/j.acorp.2025.100126

Szudarski, P. (2025). Vocabulary and Data-Driven Learning. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave encyclopedia of computer-assisted language learning* (pp. 1–7). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_76-1

Takimoto, M. (2008). The effects of deductive and inductive instruction on the development of language learners' pragmatic competence. *The Modern Language Journal, 92*(3), 369–386. https://doi.org/10.1111/j.1540-4781.2008.00752.x

Tsai, K.-J. (2019). Corpora and dictionaries as learning aids: Inductive versus deductive approaches to constructing vocabulary knowledge. *Computer Assisted Language Learning, 32*(8), 805–826. https://doi.org/10.1080/09588221.2018.1527366

Vogel, S., Herron, C., Cole, S. P., & York, H. (2011). Effectiveness of a guided inductive versus a deductive approach on the learning of grammar in the intermediate-level college French classroom. *Foreign Language Annals, 44*(2), 353–380. https://doi.org/10.1111/j.1944-9720.2011.01133.x