



This work is licensed
under a Creative
Commons Attribution
4.0 International
License.

Doing DDL right: Principles and practical tips for the classic use of corpora in the language classroom

 **Brian Inglis**

University of St Andrews, UK
bri1@st-andrews.ac.uk

The original, classic conception of data-driven learning (“DDL”), learners’ use of corpora to access exemplars of words or constructions, has been convincingly shown to be effective, but remains relatively uncommon in classrooms. Recent research shows a shift towards teacher training in corpora and DDL, a trend with the potential to stimulate wider adoption and move DDL closer to mainstream status among practitioners. This project endeavours to support the effort to make DDL more practical for teachers by examining effective and transparent implementations of it and identifying underlying principles contributing to success. Using a systematic literature search and thematic analysis, the project investigates the research questions: (1) What are the characteristics of DDL implementations reported as successful in primary studies? (2) What aspects of the implementations can their success be attributed to, and what aspects can be observed to be disadvantageous? and (3) What other themes emerge from the analysis? Based on analysis of major themes, some core characteristics of DDL and factors impacting its success are proposed. Finally, connecting these principles to practice is encouraged by reviewing tools and activities for implementing DDL confidently, efficiently and with minimal upfront effort.

Keywords: DDL, data-driven learning, corpus linguistics, corpus-based instruction

Introduction

Since Johns pioneered the field with his image of the language learner as “a Sherlock Holmes” (1998, p. 101), data-driven learning (“DDL”) has evolved considerably, but the classic style of guiding learners to learn vocabulary and constructions inductively by directly engaging with a collection of text (corpus) compiled by linguists for linguistic analysis remains influential. Now, DDL can refer to virtually any application of corpus linguistics to language learning, but it commonly involves use of keyword-in-context (KWIC) concordance lines, where each line shows the target

word, phrase, or construction in context from authentic texts. Boulton and Cobb's meta-analysis found that "DDL works pretty well in almost any context where it has been extensively tried" (2017, p. 386), but most practicing teachers would likely agree that it is relatively uncommon in classrooms. Recent research, such as Bennett and Dhonnchadha (2023) has focused on teacher training to stimulate wider uptake, and Dong et al.'s (2022) bibliometric analysis suggests that this area represents a potentially transformative route forward for the field. This introduces questions of how to effectively train teachers, and what pedagogical and learning design principles can inform the efforts. This study aimed to address these questions and support the effort to make DDL accessible by asking the research questions: (1) What are the characteristics of DDL implementations reported as successful in primary studies? (2) Which aspects of the implementations can their success be attributed to, and what aspects can be observed to be disadvantageous? and (3) What other themes emerge from the analysis?

Methods

A qualitative research synthesis (Chong & Plonsky, 2021) was chosen for its theoretical flexibility and, in combination with thematic analysis (Clarke & Braun, 2014), would allow for an exploratory, inductive investigation. A list of journals that have been influential in shaping DDL was chosen by triangulating between those identified as bibliometrically significant (Dong et al., 2022) and those identified by authoritative researchers (Lee et al., 2018; Boulton & Cobb, 2017; Boulton, 2017). This ensured that possible issues with influential researchers' subjective perceptions and with misleading bibliometric data would be alleviated, while also ensuring focus on studies that have been impactful in shaping DDL practices without being unduly influenced by excessively obscure or unconventional implementations. The journals (*Computer Assisted Language Learning*, *Language Learning*, *Language Learning and Technology*, *ReCALL*, and *System*) were searched for the keywords corpus, corpora, concordance, concordancer, concordancing, KWIC, "data-driven learning" (with a hyphen), "data driven learning" (without a hyphen), or DDL. Inclusion criteria were publication in or after 2011, identified by Dong et al. (2022) as the beginning of a mature stage in DDL research, primary studies using the classical DDL of interest to this study, written in English, and inclusion of detailed procedures that could be reproduced or approximated (allowing for analysis of the procedures and selection of studies that could directly inform practice). Thematic analysis (Clarke & Braun, 2014) was conducted on the final included studies.

Results

Fifteen studies (listed in the Appendix) remained for inclusion in the analysis, and the analysis resulted in the identification of four major themes.

Theme 1: Learning activities and materials

All studies gave some consideration to the relationship between DDL and other classroom activities, both conventional and DDL-specific. Sharing or presenting, though only present in a few studies, seemed very effective and was favoured by students. DDL specific frameworks were common, with processes being structured in stages of activities or projects. Allocation of time showed extremes, with most implementations either very brief, with short activity durations in single sessions, or protracted, included in projects spanning weeks or months.

Theme 2: Instructor facilitation

Practical skills training, a necessary component, demonstrated a wide variety of approaches, including short demonstrations, multiple session workshops, guides, or videos. Most studies showed ongoing feedback or language instruction alongside the activity, while some deliberately limited instructor involvement.

Theme 3: Psychological and mental factors

Cognitive factors received attention in all studies, with inductive learning, cognitive activation or processing, and attention/noticing as the most common characteristics. DDL was also observed to promote autonomy, which was implied to be mutually reinforcing with motivation. Four studies demonstrated anxiety or frustration, while seven showed learners reporting enthusiasm and a desire to continue using the new tools.

Theme 4: Other drawbacks and challenges

High workload and inefficiency, or the perception thereof, was a common observation. Two studies described inadvertent plagiarism, but there is nothing about DDL specifically that would encourage this more than other resources.

Discussion

The most important implications under each research question can be used to inform design principles/decisions or for further research.

RQ1: Common features of successful implementations

Some common features suggested theoretical connections that are not often foregrounded in DDL research (Boulton 2017). Cobb (1999) suggested that DDL is naturally compatible with constructivism, which is reflected in the inductive, process-oriented frameworks found in many studies, such as Geluso & Yamaguchi (2014) and Wu (2021). Further exploration of explicitly linking DDL and constructivism may produce clearer, more intuitive procedures for teachers to deploy. Additionally, many studies observed or inferred cognitive states and processes, suggesting further, more explicit application of concepts like Schmidt's (1990, 2001) noticing hypothesis

and Sheen's (2002) focus-on-form approach. These theoretical connections could contribute to the development of design principles or further research.

RQ2: Aspects associated with success

DDL seems to be especially successful when used in concert with sharing and presenting activities and made personally relevant to learners. Sharing and presenting activities were associated with exceptionally strong results (Daskalovska, 2013) or subjective experiences (Geluso & Yamaguchi, 2014). Personal relevance can be achieved by using customized or targeted learner corpora (Cotos, 2014), but simply using corpora to process feedback on writing (Yoon & Jo, 2014) or choosing their own language focus (Liu & Gablasova, 2023) was also enough to confer a sense of personal relevance. Designing DDL to centre around these activities is likely to enhance it.

RQ3: Emergent themes

The major emergent themes that could support DDL were minimalism and DDL's relation to surrounding activities. While DDL is often viewed as intimidating or labour-intensive, there seems to be an early point of diminishing returns, and it is best to prioritize simplification, streamlining, and focus. With respect to surrounding activities, these are often viewed as incidental to a DDL project, but the findings suggested that pairing DDL with particular activities had significant impact. Learning design and further research may benefit from direct attention to simplicity and the effect of integrating DDL with particular activities.

Implications and practicalities

Teachers interested in implementing DDL must first familiarize themselves with a software option, like AntConc (Anthony, 2014), which is excellent for KWIC concordance lines and relatively easy to learn, or lextutor.ca (Cobb, n.d.), which is extremely powerful but has an intimidating interface. Any option will likely take a computer-literate teacher several hours to learn, after which point the above principles will become relevant, and additional time will be needed to prepare student training activities to allow them to use the tools independently. As a workaround for greater efficiency, students can be instructed to use generative AI to view concordance lines. The output will not have the academic rigor of a corpora expertly compiled by a linguist for linguistic purposes and will likely have other potentially significant errors, so caution is required. Nonetheless, teachers are encouraged to experiment with generative AI tools to try to more efficiently achieve output similar to corpus linguistic tools. A starting prompt could be similar to the following: "Please act as a corpus tool. Show me concordance lines [Optional: insert "from academic/literary/scientific/ engineering/etc. texts"] for the word '[WORD]'. Display 20 words on each side and focus on frequent collocations and common patterns." In addition to experimentation by teachers using generative AI in this way, as these tools evolve rapidly, additional research on using generative AI for DDL or analogous learning activities is encouraged.



JALTCALL

Trends

Vol. 2 No. 1
2025 JALTCALL
Conference Papers
Special Issue

Conclusion

The keys to success appear to be a simple approach, attention to practical training, thoughtful pairing with other learning activities, and a clear idea about the constraints or expansiveness of the project. Corpus tools do not necessarily make this easy, and some upfront effort is naturally required by the teacher. Generative AI offers a workaround to some practical hurdles, but with costs to quality and rigor. The attractiveness of using AI this way points to one of the biggest problems with DDL: the user-friendliness of tools. The field would likely benefit greatly from new, streamlined tools designed specifically for language learning.

References

- Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Bennett, C. & Dhonnchadha, E. (2023). Becoming corpus literate: An in-service EFL teacher education framework for integrating corpora into EFL teaching. *Applied Corpus Linguistics*, 3(1), 1–10. <https://doi.org/10.1016/j.acorp.2023.100048>
- Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, 50(4), 483–506. <https://doi.org/10.1017/S0261444817000167>
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393. <https://doi.org/10.1111/lang.12224>
- Chong, S. W., & Plonsky, L. (2021). A primer on Qualitative Research Synthesis in TESOL. *TESOL Quarterly*, 55(3), 1024–1034. <https://doi.org/10.1002/tesq.3030>
- Clarke, V., Braun, V. (2014). Thematic analysis. In: Teo, T. (eds) *Encyclopedia of critical psychology*. Springer, 1947–1952. https://doi.org/10.1007/978-1-4614-5583-7_311
- Cobb, T. (n.d.). Compleat Lexical Tutor [Internet application]. Available at <https://www.lex tutor.ca/>
- Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research and Development*, 47(3), 15–31. <https://doi.org/10.1007/bf02299631>
- Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202–224. <https://doi.org/10.1017/S0958344014000019>
- Daskalovska, N. (2013). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130–144. <https://doi.org/10.1080/09588221.2013.803982>
- Dong, J., Zhao, Y., & Buckingham, L. (2022). Charting the landscape of data-driven learning using a bibliometric analysis. *ReCALL* (2022), 1–17. <https://doi.org/10.1017/S0958344022000222>
- Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 26(2), 225–242. <https://doi.org/10.1017/S0958344014000044>
- Johns, T. (1998). Contexts: the background, development and trialling of a concordance-based CALL program. In Johns, T., *Teaching and language corpora*. Longman, 100–115. <https://doi.org/10.4324/9781315842677-9>

- Lee, H., Warschauer, M., & Lee, J. G. (2018). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721–753. <https://doi.org/10.1093/applin/amy012>
- Liu, T., & Gablasova, D. (2023). Data-driven learning of collocations by Chinese learners of English: a longitudinal perspective. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2023.2214605>
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524780>
- Sheen, R. (2002) “Focus on form’ and ‘focus on forms.” *ELT Journal* 56(3), 303–305. Oxford University Press. <https://doi.org/10.1093/elt/56.3.303>
- Wu, Y. (2021). Discovering collocations via data-driven learning in L2 writing. *Language Learning & Technology*, 25(2), 192–214. <http://hdl.handle.net/10125/73440>
- Yoon, H., & Jo, J. H. (2014). Direct and indirect access to corpora: An exploratory case study comparing students’ error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1). <http://hdl.handle.net/10125/44356>

Appendix

List of studies analysed

- Chang, J. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26(2), 243–259. <https://doi.org/10.1017/s0958344014000056>
- Chen, H. H., Lai, S. L., Lee, K. Y., & Yang, C. T. (2021). Developing and evaluating an academic collocations and phrases search engine for academic writers. *Computer Assisted Language Learning*, 36(4), 641–668.
- Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202–224. <https://doi.org/10.1017/s0958344014000019>
- Daskalovska, N. (2013). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130–144.
- Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 26(2), 225–242. <https://doi.org/10.1017/s0958344014000044>
- Gordani, Y. (2013). The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners’ vocabulary development. *Computer Assisted Language Learning*, 26(5), 430–445.
- Karras, J. N. (2015). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL*, 28(2), 166–186. <https://doi.org/10.1017/s0958344015000154>

- Larsen-Walker, M. (2017). Can data driven learning address L2 writers' habitual errors with English linking adverbials? *System*, 69, 26–37.
<https://doi.org/10.1016/j.system.2017.08.005>
- Lee, P., & Lin, H. (2019). The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention. *System*, 81, 14–25.
<https://doi.org/10.1016/j.system.2018.12.011>
- Liu, T., & Gablasova, D. (2023). Data-driven learning of collocations by Chinese learners of English: a longitudinal perspective. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2023.2214605>
- Mizumoto, A., & Chujo, K. (2016). Who is data-driven learning for? Challenging the monolithic view of its relationship with learning styles. *System*, 61, 55–64.
<https://doi.org/10.1016/j.system.2016.07.010>
- Mueller, C. M., & Jacobsen, N. D. (2015). A comparison of the effectiveness of EFL students' use of dictionaries and an online corpus for the enhancement of revision skills. *ReCALL*, 28(1), 3–21. <https://doi.org/10.1017/s0958344015000142>
- Tsai, K. (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>
- Wu, Y. (2021). Discovering collocations via data-driven learning in L2 writing. *Language Learning & Technology*, 25(2), 192–214.
<http://hdl.handle.net/10125/73440>
- Yoon, H., & Jo, J. H. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1).
<http://hdl.handle.net/10125/44356>

Author biodata

Brian Inglis is a university English instructor based in the Kansai region of Japan and a doctoral student with the University of St Andrews. His research interests include psychological theories of self-regulation and self-efficacy and how they can be applied to EFL classrooms and second language acquisition.