Gilquin, G. 2026. Data-driven learning. In H. Nesi & P. Milin (eds) *International Encyclopedia of Language and Linguistics. Third Edition. Volume 9* (pp. 627-630). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-323-95504-1.00845-0

Data-driven learning

Gaëtanelle Gilquin

Abstract
Data-driven learning (DDL) is a pedagogical approach which involves letting students use corpus data to help them gain knowledge of a target language. In this article, various ways of doing DDL are presented, depending on whether it is carried out in a direct or indirect manner, what corpora are used and what types of searches are performed. Research on the evaluation of DDL is also reported on, as well as future directions, including the place of DDL in relation to generative artificial intelligence.

Keywords
Corpus-based teaching, corpus linguistics, direct DDL, evaluation of DDL, expert corpora, generative artificial intelligence, indirect DDL, foreign language learning, learner corpora, multimodal corpora, native corpora, parallel corpora, pedagogical practice, second language acquisition.

Key points
- DDL can take various forms, being direct or indirect and relying on different corpora and different types of searches.
- DDL usually produces positive learning outcomes, including gains in language proficiency, but also develops more general skills.
- DDL is a thriving field of research, but improvements in terms of classroom implementation are needed for DDL to become a fully-fledged teaching practice.

Introduction

According to most current theories of language acquisition, learning a language requires some data to learn from, either in the form of language input that learners get exposed to (in the case of implicit learning) or in the form of rules that learners need to study and apply (in the case of explicit learning). In this sense, 'data-driven learning' could be used to describe any form of language acquisition. However, in the field of corpus linguistics, the term 'data-driven learning' – often abbreviated DDL – has a more restricted meaning. It refers to the use of corpus data by learners as a resource to learn (about) a language. While DDL is relevant to native speakers (e.g., to learn how to write academic texts), the focus here will be on the use of DDL for foreign/second language (L2) learning.

Corpora, as repositories of naturally-occurring language, are valuable resources for L2 learning, because they show how the target language is actually used. They also give access to potentially large amounts of data, and hence to many instances of various linguistic phenomena (words, syntactic structures, types of discourse, etc.), which often amount to many more instances than learners could encounter in a typical language learning situation. Because corpora come in many different forms, they can expose learners to a wide range of language varieties and registers, possibly in accordance with their individual needs or wishes.

Direct and indirect DDL

DDL involves letting learners use corpus data. In this respect, a distinction can be drawn between direct and indirect uses of corpus data (Yoon & Jo, 2014). In direct uses, learners get to explore a corpus themselves. Having access to a whole corpus and to different tools to search it means that learners can probe into many aspects of language, including those that they have particular difficulty with. On the other hand, the task may be daunting for learners who have limited experience with corpora. In this case, guidance from the teacher is essential, for example in the form of specific steps that learners should follow one by one to find what they are looking for.

In indirect uses of corpus data, teachers' help is more substantial, as they query the corpus themselves and give learners material with some preselected corpus data, for example a handout with concordance lines, that is, a list of occurrences of an item in context taken from the corpus, and possibly some tasks related to these data. The preselection by the teacher may reduce the degree of authenticity (if only some instances of a word are kept in the concordance lines, for example, the set of concordance lines may no longer reflect the actual distribution of uses of the word), but it can help make the material better adapted to the learners, e.g., by removing concordance lines that would be too difficult for them to understand.

Corpora used in DDL

DDL mostly relies on written corpora, usually representing native, or at least expert, writing. Native or expert corpora tend to serve as models for learners. They help learners determine what is acceptable, what is frequent, what is typical of certain registers, etc. In Chen & Flowerdew (2018a), for example, PhD students in Hong Kong were given access to corpora of research articles, from which they could learn, among other things, about the use of certain words typical of research articles or about the structure of research articles and the features of the different sections (abstract, introduction, literature review, etc.). The students were also encouraged to build their own corpus, representing their specific (sub)discipline, so as to meet their personal needs.

Learner corpora, which are made up of data produced by language learners, can be used as a basis for DDL too. They constitute a source of information about L2 production, including incorrect and unidiomatic forms but also cases of overuse and underuse (i.e., the use of significantly more or fewer instances of an item than in a reference corpus). Students can thus discover what is likely to be problematic in their own production. Provided they are given access to an error-tagged version of the learner corpus (in which errors have been identified and corrected) or to a native/expert corpus with which they can compare the learner corpus, students should also be able to find out how to correct or improve L2 production. In order to avoid exposing students to errors that they would be unlikely to make (see Flowerdew, 2001), it is recommended that the selected learner corpus should represent a population of learners whose profiles are similar to those of the students participating in the DDL activity (same mother tongue background, same proficiency level, etc.). Better still, one can collect and exploit a local learner corpus (Seidlhofer, 2002), which includes data produced by the students themselves, so that the knowledge gained from the DDL activity is directly relevant to them, showing them how they (mis)use the target language. In Cotos (2014), for example, students were required to examine the use of English linking adverbials in a local learner corpus

containing their own writing compared to use in a native corpus and, on that basis, to discover – and correct – inappropriate uses of linking adverbials.

DDL can also make use of bilingual corpora, and more precisely parallel corpora, which include texts in one language and their translations into another language. Such corpora make it possible for students to compare their native language with the target language and to become aware of similarities and differences between them. In Reynolds (2016), Taiwanese students were given access to parallel corpora in Chinese (their native language) and English (the target language). The task required the students to self-edit their academic essays on that basis, with a focus on verb-noun collocations, which tend to be expressed differently across languages. Parallel corpora can also be particularly useful for students in translation, to help them uncover different translation strategies and, more generally, to learn about translation practices from professional translations, which act as models for students, in the same manner as the native/expert corpora discussed previously.

Recently, a plea for more multimodality in DDL has been made (see, e.g., Meunier, 2020). Using speech and video corpora, one can extend DDL to additional aspects of language, including pronunciation and gestures. Hirschi & Kang (2024) offer a good illustration of how DDL can be applied to the learning of pronunciation. In their study, learners of English were able to use a speech corpus to practise their perception and production of lexical stress and prominence of academic words and multiword expressions. Particularly interesting is the fact that the corpus included data produced by highly proficient L2 speakers, validated for their intelligibility, which points to the potential of L2 corpus data as a suitable model for learners.

The availability of a wide range of corpora, representing different languages and registers produced by different populations of language users, means that the corpora exploited in DDL can be chosen according to the students' profiles. For secondary school learners, for instance, it may be pedagogically more relevant to use a corpus of teenage language, as suggested by Pérez-Paredes (2020). For university students learning academic writing, a corpus of scientific publications in their own discipline will help them find out about the conventions that are specific to their discipline. If a suitable corpus is not available, one could be compiled by the teacher or the students.

Types of corpus searches for DDL

Corpora can be queried by means of special software, which usually allows for different types of searches. The most often used function in DDL is concordancing. A concordance search allows learners to make inferences about the uses of the search item on the basis of the surrounding context (its main meanings, the words with which it occurs, the syntactic structures in which it is found, etc.). This is illustrated by Hirata & Thompson (2022), a study in which all DDL activities were based on concordances. More precisely, curated concordance lines were proposed to the learners, in which only "those lines with the highest potential for students to engage with" were kept (Hirata & Thompson, 2022, p. 359). Concordance lines can also be used to prepare gap-fill exercises (e.g., Boulton, 2009), simply by removing the search item from the lines.

Sometimes, DDL relies on frequency lists, which show the words from a corpus in descending order of frequency. In Ackerley (2017), frequency lists stemming from an expert corpus and a learner corpus were examined by the students to determine what words were overused or

underused in the learner corpus as compared to the expert corpus. An automatically generated list of keywords (i.e., words distinctive of a corpus when compared to a reference corpus) could be used to the same effect.

Lists of collocates (i.e., words typically occurring close to a search item) or n-grams (i.e., recurrent sequences of words) can also be extracted automatically with most corpus query tools. They can serve as a basis for DDL activities that focus on phraseological aspects of language (e.g., Daskalovska, 2015; Lay & Yavuz, 2020).

Evaluation of DDL

According to Dong et al.'s (2023) bibliometric analysis of DDL studies, "effectiveness of DDL" is the most common research theme, based on co-citation. Indeed, many empirical studies on DDL include some sort of evaluation, which can concern the participants' attitudes, their practices or the efficiency of DDL (see Boulton, 2008a). In addition, several reviews and meta-analyses have evaluated the evolution of the field, considering its strengths and limitations, and have sought to bring to light the factors that affect the efficiency of DDL (e.g., Boulton & Cobb, 2017; Boulton & Vyatkina, 2021; Chen & Flowerdew, 2018b; Pérez-Paredes, 2019).

What these evaluations reveal is that DDL is not always successful. Some students have difficulty dealing with corpus data and they do not necessarily appreciate this form of pedagogy. They may not learn from DDL activities, or at least not more than with traditional teaching methods. In addition, certain issues may arise during DDL, such as the problem of understanding truncated concordance lines or the time needed to carry out DDL tasks (Chen & Flowerdew, 2018b).

Yet, a majority of DDL studies underline the overall efficiency of DDL. While it may work better in certain contexts than in others (e.g., with students specializing in languages rather than in other areas or with local corpora rather than public corpora; see Boulton & Cobb, 2017), it usually results in some positive learning outcomes. Importantly, these learning outcomes are not limited to gains in language proficiency. It is widely recognized that DDL helps students develop other skills, such as noticing, hypothesizing, interpreting or inferring (e.g., O'Sullivan, 2007).

In addition, DDL has the advantage that it leads to students' active participation, which "corresponds closely to current thinking in educational research in general, and in language-learning pedagogy in particular, in a constructivist paradigm where the active role of the learner is of crucial importance" (Chambers, 2021, p. 417). It also allows for tailor-made activities, possibly adapted to each individual student, which opens the door to differentiated teaching practices and more inclusive pedagogy (see Gilquin, 2024). Moreover, it is compatible with other current pedagogical approaches, such as communicative language teaching (Hirata & Thompson, 2022), task-based language teaching (Wicher, 2020) and translanguaging (Gilquin, 2022).

Interestingly, it has been claimed that DDL could be "reaching the parts other teaching can't reach" (Boulton, 2008b, p. 38). Gaskell & Cobb (2004, p. 301), for example, argue that "[s]entence-level writing errors seem immune to many of the feedback forms devised over the years" and suggest that DDL could help students notice correct uses of such sentence features.

Nesselhauf (2004) points out that learner corpora, used in combination with native corpora, can provide a solution to fossilized errors, while Fuster-Márquez & Gregori-Signes (2018, p. 164) note that controlled activities involving learner corpora "may cover certain learners' needs not found in textbooks". This shows that DDL has a unique contribution to make and that it can serve as a useful complement to more traditional teaching methods.

Directions for the future

Despite the progress that DDL has made over the years and the increasing number of publications devoted to the topic which have turned DDL into a proper field of research, a number of developments are needed for DDL to become a fully-fledged teaching practice. Dong et al. (2023) mention teacher training and classroom implementation as two of the most urgent needs, with the former being a prerequisite for the latter. While most DDL studies so far have been carried out by researchers and have taken the form of teaching interventions in (quasi-)experimental settings, the ultimate aim is for DDL to be widely adopted in classrooms, and possibly even outside classrooms (Gilquin & Granger, 2022). Other areas for improvement have been identified in the literature, including the need for more DDL involving languages other than English, activities better suited for younger and lower-proficiency learners, open-access DDL exercises and corpus user guides (see Vyatkina, 2020).

Another issue that will have to be addressed in the near future is the possible impact of generative artificial intelligence (GenAI) on DDL. Crosthwaite & Baisa (2023) argue that GenAI will not necessarily spell the end of DDL. On the contrary, they claim that the two could complement each other and compensate for their respective shortcomings. They give as an example the fact that DDL relies on authentic texts whose origins and features are known and that it allows for replication of queries, unlike GenAI. On the other hand, they point out that GenAI is generally more user-friendly than concordancers and does not require any technical knowledge, and that it can directly translate linguistic findings into teaching materials. Crosthwaite & Baisa (2023, p. 4) therefore suggest combining DDL and GenAI. They even go so far as to claim that "*not* leveraging GenAI risks leaving DDL practitioners and DDL as an enterprise behind, given the current depth of interest in GenAI for education" (emphasis original).

Conclusion

DDL comes in many different forms, depending on whether it is applied in a direct or indirect manner, what corpora are used and what types of searches are performed. While certain forms of DDL may be more or less successful with a given learner population, empirical studies have demonstrated the overall efficiency of this pedagogical approach, which can improve students' knowledge of different aspects of the target language, but also develop some of their more general skills.

Substantial progress has been made over the last few years in DDL research, which has gone through what Dong et al. (2023) call the conceptualizing stage, the maturing stage and, more recently, the expansion stage. It is now time to work on the next stage, which will involve turning DDL into a widely used pedagogical practice, fully integrated into the teaching curriculum. DDL has a unique contribution to make among current teaching approaches, so this last step should make it possible to reap all the benefits DDL has to offer.

References

Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology*, *21*(3), 195–216. https://doi.org/10125/44627.

Boulton, A. (2008a). Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et Corpus*, *3*, 37–46.

Boulton, A. (2008b). DDL: Reaching the parts other teaching can't reach? In A. Frankenburg-Garcia (Ed.), *Proceedings of the 8th teaching and language corpora conference* (pp. 38–44). Lisbon: Associação de Estudos e de Investigação Cientifica do ISLA-Lisboa.

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

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

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

Chambers, A. (2021). What is data-driven learning? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 416–429). London: Routledge. https://doi.org/10.4324/9780367076399-29.

Chen, M., & Flowerdew, J. (2018a). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes*, *50*, 97–112. https://doi.org/10.1016/j.esp.2017.11.004.

Chen, M., & Flowerdew, J. (2018b). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, *23*(3), 335–369. https://doi.org/10.1075/ijcl.16130.che.

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

Crosthwaite, P., & Baisa, 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.

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

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

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 363–379). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.5.21flo.

Fuster-Márquez, M., & Gregori-Signes, C. (2018). Learning from learners: A non-standard direct approach to the teaching of writing skills in EFL in a university context. *Innovation in Language Learning and Teaching*, *12*(2), 164–176. https://doi.org/10.1080/17501229.2016.1142549.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319. https://doi.org/10.1016/j.system.2004.04.001.

Gilquin, G. (2022). Translanguaging and data-driven learning: How corpora can help leverage learners' multilingual repertoires. *Teaching English as a Second Language Electronic Journal (TESL-EJ)*, *26*(3). https://doi.org/10.55593/ej.26103a22.

Gilquin, G. (2024). Learner corpora to meet learners' individual needs. *Babylonia*, *2*, 14–19. https://doi.org/10.55393/babylonia.v2i.386

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics, second edition* (pp. 430–442). London: Routledge.

Hirata, Y., & Thompson, P. (2022). Communicative data-driven learning: A two-year pilot study. *ELT Journal*, *76*(3), 356–366. https://doi.org/10.1093/elt/ccab066.

Hirschi, K., & Kang, O. (2024). Data-driven learning for pronunciation: Perception and production of lexical stress and prominence in academic English. *TESOL Quarterly*, Early View. https://doi.org/10.1002/tesq.3302.

Lay, K. J., & Yavuz, M. A. (2020). Data-driven learning of academic lexical bundles below the C1 level. *Language Learning & Technology*, *24*(3), 176–193. https://doi.org/10125/44741.

Meunier, F. (2020). A case for constructive alignment in DDL: Rethinking outcomes, practices, and assessment in (data-driven) language learning. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 13–30). London: Routledge.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–152). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.11nes.

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

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. (2020). The pedagogic advantage of teenage corpora for secondary school learners. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 67–87). London: Routledge.

Reynolds, B. L. (2016). Action research: Applying a bilingual parallel corpus collocational concordancer to Taiwanese medical school EFL academic writing. *RELC Journal*, *47*(2), 213–227. https://doi.org/10.1177/0033688215619518.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.6.14sei.

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

Wicher, O. (2020). Data-driven learning in the secondary classroom: A critical evaluation from the perspective of foreign language didactics. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 31–46). London: Routledge.

Yoon, H., & Jo, 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 and Technology*, *18*(1), 96–117. https://doi.org/10125/44356.