

UNLEASHING THE POTENTIAL OF DATA-DRIVEN LEARNING (DDL) IN LANGUAGE TEACHING AND LEARNING

Humairah Fauziah

Universitas Negeri Malang, Malang

humairah.fauziah.2102219@students.um.ac.id; humairah.fauziah@gmail.com

Abstract: Data-driven learning (DDL) is a pedagogical strategy that makes use of real-world linguistic data to improve language acquisition. DDL has the potential to equip students with a variety of essential learning abilities, such as a component of discovery learning that makes learning more engaging and encourages students to take an active role in their own learning. DDL enables students to analyze and identify linguistic patterns, enhance their language proficiency, and gain metalinguistic awareness by using corpora and concordancers. In order for learners to actively engage with language data and uncover linguistic patterns and rules, DDL changes the emphasis from standard grammar education to a more learner-centered approach. The article discusses the idea of DDL and its implications for language education, emphasizes the advantages of DDL in students' language development, acknowledges the limitations and difficulties of DDL, and suggests potential solutions to address these difficulties by using examples from previous research studies to show how DDL activities affect language learning outcomes.

Keywords: Data-Driven Learning (DDL), Direct Use of Corpora, Discovery Learning, Indirect Use of Corpora.

INTRODUCTION

Data-driven learning (DDL) is a term coined by Tim John referring to the use of corpus for language teaching and learning, in which it has been originally used as a source of data for linguistic research such as translation, forensic linguistics, and discourse analysis (Lin, 2021). The main idea of bringing corpus in the language teaching and learning is to provide ample opportunities for students to explore the target language data from the corpus and draw their own conclusion based on their observation. In the context of learning English as a foreign language (EFL), exploring the target language in the corpus is beneficial for students to understand how the target language is used within its specific context because corpus contains natural discourse (written texts of spoken language transcripts) in diverse contexts that presents descriptive information on how people use the language in various language functions, contexts, levels of formality, and manner of writing or speaking (Gilquin & Granger, 2022).

While DDL holds immense potential for language learners, it is important to acknowledge the challenges it presents, particularly for students at lower proficiency levels (Boulton, 2019). These students often lack the necessary linguistic competence to effectively analyze language patterns, which can impede their ability to fully benefit from DDL activities. Thus, this paper aims to examine the challenges associated with DDL and explore its pedagogical implications for English as a Foreign Language (EFL) students across all proficiency levels. By addressing the limitations of DDL, this paper seeks to provide examples of DDL activities that can be implemented in EFL classrooms. These activities aim to overcome the challenges posed by low-level students' linguistic competence and foster their active

involvement, autonomy, and motivation in the language learning process. By delving into the potential benefits and practical applications of DDL, this paper aims to shed light on its effectiveness and encourage further exploration in adapting DDL to meet the diverse needs of EFL students.

METHOD

Data-Driven Learning (DDL)

Data-driven learning (DDL) treats language as data and positions students as researchers (Anthony, 2019), moving away from traditional language learning approaches guided by teachers and textbooks (Boulton, 2017). DDL empowers students with essential language skills and encourages curiosity and active learning. Johns (1991) suggested three approaches for corpus based DDL language teaching: 1) identifying and exploring language problems that students want to address, 2) classifying language patterns, and 3) formulating rules based on corpus data, while Boulton (2010) listed four basic DDL processes: exploring data, detecting patterns, formulating hypotheses, and generalizing to other cases. DDL involves identifying problems, observing language, formulating hypotheses, constructing rules, validating them through textbooks, and generalizing to other cases (see Figure 1).

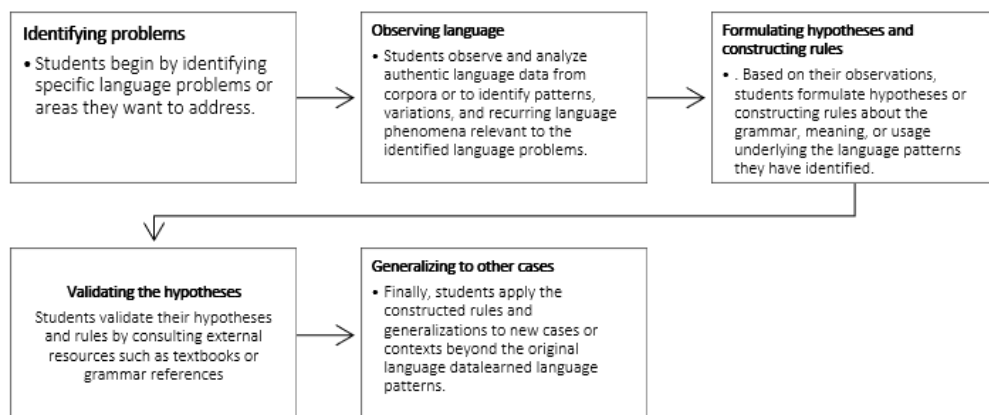


Figure 1. The Steps of Corpus-Based DDL Classroom Activities

In DDL, students can use corpus directly and indirectly (Crosthwaite, 2019). The primary distinction between direct and indirect corpus use in data-driven learning (DDL) is how learners interact with and access language data (see Table 1). Teachers give students immediate corpus access to examine target language patterns using a specific software (Sun & Hu, 2020). Teachers should verify what students find in the corpus since they may misinterpret their study. Indirect corpus use, also known as hands-on concordancing (Boulton, 2012a) and corpus-printouts (Stevens, 1991), involves teachers giving students corpus-printouts with several corpus samples. In this instance, teachers should evaluate students' premise, choose appropriate examples, and remove problematic examples.

Tools Used in DDL

DDL always deals with the use of a tool or concordance software to access the corpus and to present the corpus data (Gilquin & Granger, 2022). These tools help learners access, search, and manipulate the corpus effectively.

1. Concordancer

Concordancer is a software that allows students to search and analyze corpus language data. Its user-friendly interface lets students enter Key Word in Context (KWIC) words, phrases, or grammatical structures and get corpus instances (Kwary, 2018). Concordance software lets

learners see how the KWIC is used in multiple settings, study collocations and patterns, and understand language usage. The software generates concordance lines that show the searched item and its surrounding words, giving learners rich examples to evaluate and comprehend. If students search for the KWIC "teach" in the corpus, the sentences including the word will display as concordance lines in various forms (see Figure 2). The software also allows students to alter and study corpus data through sorting, filtering, and frequency analysis (Lee et al., 2015). Concordance software helps learners discover patterns and make educated linguistic observations in DDL activities.

His ideas of the ability of children has influenced what is **taught** to children, for example division would not be taught to a five-year-old child. The skills involved in road crossing will improve our ability to **teach** children affective road skills. Whilst this study does not follow a certain behavior and operant conditioning " **teaches** " through negative reinforcement, how to prevent the event from occurring. There are many merits to this method. It **teaches** patients how to prevent themselves from being in situations where they are not wanted.

Figure 2. The KWIC *teach* in BAWE Corpus using Sketch Engine (as Concordance Software)

2. Corpus

A corpus refers to a large and structured collection of texts or spoken language data that is used as a representative sample of a particular language or languages. A corpus can consist of written texts from different genres and sources, such as books, newspapers, magazines, websites, and academic papers. It can also include transcriptions of spoken language, such as conversations, interviews, speeches, or recorded dialogue. In the context of classroom instruction, there are four main types of corpora frequently utilized: generalized corpus, specialized corpus, learner corpus, and DIY (do it yourself) corpus (Bennett, 2010). These corpora play a direct role in facilitating the teaching and learning process.

Table 1. Type of Corpus

No	Type of Corpus	Description
1	General Corpus	A generalized corpus consists of a diverse range of texts that represent the language in general. It includes texts from various genres, such as novels, newspapers, websites, and academic papers.
2	Specialized Corpus	A specialized corpus focuses on specific domains or subject areas. It consists of texts that are relevant to a particular field of study or profession (Hyland, 2019).
3	Learner Corpus	A learner corpus comprises texts produced by language learners themselves. It contains samples of learners' written or spoken output at various proficiency levels. A learner corpus provides valuable insights into the language development and errors made by learners (Granger & Tribble 1998).
4	Pedagogic or DIY (Do It Yourself) Corpus	The DIY corpus, also known as a self-constructed corpus, is created by the learners or teachers themselves. It involves collecting and organizing language data from authentic sources that are specifically used for language learning purposes (Chambers, 2019) and relevant to the learners' needs and interests (Chen & Flowerdew 2018).

TASK TYPES IN DDL

In Data-Driven Learning (DDL), various types of tasks can be implemented to engage students in language analysis and exploration. These tasks aim to promote active learning, critical thinking, and the application of discovered language patterns. Here are some common types of tasks used in DDL:

Table 2. Task Type of DDL

No	Type of Corpus	Description
1	Concordance Analysis	Students analyze concordance lines or examples from the corpus to identify language patterns, collocations, or grammatical structures. They can observe word frequencies, examine the context of language use, and draw conclusions about the target language feature.

2	Error Correction	Students use a corpus to rectify their language errors. In this activity, students analyze authentic language examples from the corpus to compare their own language usage with correct and natural language patterns (Crosthwaite & Stell 2020).
3	Error Analysis	Students analyze language data from the corpus to identify and analyze common errors or mistakes made by language learners. They can compare correct and incorrect language usage, identify recurring error patterns, and propose explanations or solutions for the errors.
4	Contrastive Analysis	Students compare language data from their native language with the target language to identify similarities, differences, and potential areas of difficulty in language learning.

FINDINGS and DISCUSSION

Challenges Of DDL for Language Learners

Data-Driven Learning (DDL) presents several challenges that educators and learners may encounter. One of the DDL limitations is the challenge posed by insufficient linguistic competence of learners. Insufficient linguistic competence can present a significant challenge for learners engaging in Data-Driven Learning (DDL). Students with limited vocabulary or grammar knowledge may struggle to make accurate observations or interpretations when directly using the corpus in their research, that refers to 'fake discovery' (O'Keeffe, 2021).

The second DDL limitation deals with the availability of access to resources in which DDL heavily relies on access to appropriate language corpora and technological tools for data analysis (Gilquin & Granger, 2022). However, not all educational institutions or learners may have access to comprehensive and up-to-date linguistic corpora. Limited access to relevant resources can hinder the effective implementation of DDL and limit the range of language data available for analysis. The third concerns the appropriate time and training for both teachers and students (Papaioannou et al., 2020). Educators need to familiarize themselves with corpus linguistics principles and concordancing tools to guide students effectively. Students also need time to develop the necessary skills for analyzing language data and formulating hypotheses.

The fourth is about the contextual relevance of language corpora that may not always align perfectly with learners' specific contexts or needs. While corpora provide valuable language data for analysis, they may not always align perfectly with learners' specific contexts or needs. The language data within corpora can vary in terms of register (formality), genre (text type), or cultural relevance, which can impact the applicability of the findings to learners' real-life communication needs. The corpus data might not accurately reflect the appropriate usage of hedging expressions in scholarly contexts. Lastly, some students may find it challenging to adapt to the self-directed nature of the approach because DDL aims to promote active learner engagement. The openness and freedom offered by DDL may lead to feelings of uncertainty or confusion among learners who are accustomed to more structured instruction. Maintaining consistent motivation and engagement throughout the DDL process can be a challenge for both learners and educators.

Previous researchers have made valuable contributions by offering a range of solutions to address the limitations of Data-Driven Learning (DDL). In the context of using a corpus directly in language learning, one of the challenges that students often encounter is the overwhelming amount of data to analyze. Gilquin and Granger (2020) suggest teachers manipulate the presentation of corpus data. Instead of overwhelming students with the entire corpus, teachers can select specific concordance lines that are relevant to the students' context and learning goals. Another strategy involves asking students to compile their own "personal corpus," which consists of texts specifically selected by the learners themselves. This personalized corpus allows them to focus on language materials that are directly related to their academic or professional interests, enhancing their engagement and relevance in the learning process (Chen & Flowerdew 2018).

Chen & Flowerdew (2018) and Papaioannou et al. (2020) encourage teachers to integrate DDL with their regular language instruction. In this case, teachers can identify specific learning objectives and topics where DDL can provide valuable insights and integrate them into the curriculum accordingly. In DDL, students may struggle with independently navigating and making sense of the corpus data. This is where scaffolding DDL activities can play a crucial role in addressing this limitation (see Corino & Onesti, 2019). Scaffolding refers to the process of providing support and guidance to learners as they engage with new and challenging tasks.

New Direction of DDL In Language Teaching And Learning

In the realm of Data-Driven Learning (DDL), the use of corpora can sometimes be daunting for English as a Foreign Language (EFL) and lower-level students due to its complexity. However, Smart (2014) proposes an alternative, which I prefer to call it as a "softer-DDL" approach, that offers a more accessible and learner-friendly method for language learning. The softer-DDL approach emphasizes the importance of discovery learning and the utilization of authentic sources of language. Instead of relying solely on corpora, the softer-DDL approach recognizes the significance of discovery learning, allowing students to develop their own understanding of language rules and patterns through more accessible sources such as articles, blogs, videos, or real-life texts. These authentic sources expose learners to genuine language use in different contexts, making the learning process more meaningful and relevant. By shifting the focus from complex corpora to a softer-DDL approach, teachers can create a supportive and engaging learning environment for EFL and lower-level students.

In addition, Boulton (2012) suggests an alternative perspective in which the web can be used as the corpus and Google as the concordancer. He acknowledges that traditional corpora and concordancers were initially designed for linguists, which may present a barrier for occasional or non-specialist users. Boulton (2012) argues that the familiarity and accessibility of the internet and Google make them appealing tools for language research and analysis. This approach allows users to leverage the vast amount of data available on the web and utilize Google's search capabilities to explore language patterns and conduct linguistic investigations. However, He also highlights the advantages and disadvantages of using the web and Google as corpus and concordancer tools from a pedagogical perspective. While the web offers a vast bank of data that is familiar to most people, its heterogeneity and lack of control may pose challenges in terms of data reliability and representativeness. Similarly, relying on Google as the concordancer may provide a user-friendly interface, but it may lack the advanced linguistic functionality and customization options offered by dedicated concordancer tools.

By considering the softer-DDL approach advocated by Smile (2014) and the use of alternative sources of language data, teachers can overcome the potential challenges and limitations associated with traditional corpora. This new direction in DDL encourages teachers to explore a wider range of authentic language materials and resources, such as online articles, blogs, videos, or even web searches, as valuable sources of linguistic data.

CONCLUSION

Data-Driven Learning (DDL) encourages students to actively study language trends, hypothesize, and discover linguistic laws. DDL's major features aid EFL students. It improves language comprehension, autonomy, and involvement. Students observe, analyze, and create rules using language as data. DDL tasks must be adjusted for different competence levels. Educators can make DDL accessible and beneficial for all students by meeting the demands of learners at varying skill levels. Language learners can benefit from DDL, while pupils with weaker linguistic skill may struggle. Scaffolding training, offering clear language input, progressive implementation, and integrating DDL with other language instruction approaches might help learners with different linguistic competences overcome DDL's constraints. DDL

exercises can be made more relevant and engaging by selecting and supplementing corpus data, addressing learners' contextual needs, and encouraging cooperation and reflection. However, the use of corpus in DDL can be challenging for some students. Hence, teacher can consider other language resources in DDL such as web as the corpus and Google as the concordance. Further research is needed to determine how this new DDL direction works in practice. Studies can examine how the softer-DDL approach and alternate language sources affect language learning outcomes, motivation, and engagement. Language teaching and learning would benefit from guidelines, tools, and training programs to help teachers integrate this new DDL direction into their pedagogical methods.

REFERENCES

- Anthony, L. (2019). Tools and strategies for data-driven learning (DDL) in the EAP writing classroom. In K. Hyland and L.L.C. Wong (Eds.), *Specialised English: New Directions in ESP and EAP Research and Practice* (179–194). Routledge.
- Bennet, G.R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. University of Michigan Press ELT
- Boontam, P., & Phoocharoensil, S. (2018). Effectiveness of English preposition learning through data-driven learning (DDL). *3L: Language, Linguistics, Literature*, 24(3), 125–141. <https://doi.org/10.17576/3L-2018-2403-10>
- Boulton, A. (2019). Data-driven learning for younger learners: Obstacles and optimism. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp.14-20). Routledge.
- Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, 50(4), 483–506. <https://doi.org/10.1017/S0261444817000167>
- Boulton, A. (2012a). Hands-on/hands-off: Alternative approaches to data-driven learning. In J. Thomas & A. Boulton (Eds.), *Input, process, and product: Developments in teaching and language corpora* (pp. 152–168). Brno: Masaryk University Press.
- Boulton, A. (2012b). What data for data-driven learning?. *The EUROCALL Review. Proceedings of the EUROCALL 2011 Conference*. 20(1), 23-27. <https://doi.org/10.4995/eurocall.2012.16038>
- Boulton, A. (2010). Data-driven learning: On paper, in practice. In T. Harris, & M. Jaén (eds). *Corpus linguistics in language teaching* (pp.17–52). Peter Lang.
- Chambers, A. (2019). Towards the Corpus Revolution? Bridging the Research-Practice Gap. *Language Teaching*, 52(4): 460-75. <https://doi.org/10.1017/S0261444819000089>
- Chen, M., & Flowerdew, J. (2018). 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>
- Corino, E., & Onesti, C. (2019). Data-driven learning: A scaffolding methodology for CLIL and LSP teaching and learning. *Front. Educ.* 4(7), 1-12. <https://doi.org/10.3389/educ.2019.00007>
- Crosthwaite, P. and Stell, A. (2020). It helps me get ideas on how to use my words: Primary School Students' Initial Reactions to Corpus Use in a Private Tutoring Setting', in P. Crosthwaite (ed.) *Data-Driven Learning for the next generation: Corpora and DDL for Pre-Tertiary Learners* (pp. 150-70). Routledge.
- Crosthwaite, P. (2019). Data-driven learning and younger learners: Introduction to the volume. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners* (pp. 1-10). Routledge. <https://doi.org/10.4324/9780429425899-1>
- 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>
- Granger, S. and Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (ed.), *learner English on computer* (pp. 199-209). Longman.
- Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal*, 63(4), 363–372. <https://doi.org/10.1093/elt/ccn056>
- Gilquin, G. & S. Granger. (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). Routledge.
- Hyland, K. (2019). Foreword: Corpora and specialised English in the university curriculum. In P. Crosthwaite and L. Cheung (eds), *Learning the language of dentistry: Disciplinary corpora in the teaching of English for specific academic purposes* (pp. xi-xiv). John Benjamins. <https://doi.org/10.1075/scl.93.for>
- Johns, T. F. (1991). Should you be persuaded: Two examples of data-driven learning materials. *English Language Research Journal*, 4, 1-16. Retrieved from https://lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf
- Kwary, D.A. (2018). A corpus and a concordancer of academic journal articles. *Data in Brief*, 16, 94-100. <https://doi.org/10.1016/j.dib.2017.11.023>
- Lee, J. H., Lee, H., & Sert, C. (2015). A corpus approach for autonomous teachers and learners: Implementing an on-line concordancer on teachers' laptops. *Language Learning & Technology*, 19(2), 1–15. Retrieved from <http://llt.msu.edu/issues/june2015/emerging.pdf>
- Lin, M.H. (2021). Effects of data-driven learning on college students of different grammar proficiencies: A preliminary empirical assessment in EFL classes. *SAGE Open*, 1-15. <https://doi.org/10.1177/21582440211029936>
- O'Keeffe, A. (2021). Data-driven learning: A call for a broader research gaze. *Language Teaching*, 54(2), 259-72. <https://doi.org/10.1017/S0261444820000245>
- Papaioannou, V., Mattheoudakis, M. and Agathopoulou, E. (2020). Data-driven learning in a Greek secondary education setting: The implementation of a blended approach. In P. Crosthwaite (ed.), *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners* (pp. 187-207). Routledge.
- 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>
- Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant authentic text. *English for Specific Purposes*, 10(1), 35–46. [https://doi.org/10.1016/0889-4906\(91\)90014-N](https://doi.org/10.1016/0889-4906(91)90014-N)
- Sun, X., & Hu, G. (2020). Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class. *Language Teaching Research*, 27(3), 1-29. <https://doi.org/10.1177/1362168820954459>