




Data-driven learning beyond English: Insights and implications from three monographs

Luciana Forti^a, Nina Vyatkina^b and Eva Schaeffer-Lacroix^c

^aDepartment of Italian Language, Literature and Arts in the World, University for Foreigners of Perugia, , luciana.forti@unistrapg.it; ^bDepartment of Slavic, German, and Eurasian Studies, University of Kansas, , vyatkina@ku.edu and ^cTeacher Training Department, Sorbonne University, , eva.lacroix@sorbonne-universite.fr.

How to cite: Forti, L.; Vyatkina, N.; Schaeffer-Lacroix, E. (2023). Data-driven learning beyond English: Insights and implications from three books. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik. <https://doi.org/10.4995/EuroCALL2023.2023.16865>

Abstract

This paper stems from the 2023 CorpusCALL SIG Symposium on Data-Driven Learning (DDL) for Languages Other Than English (LOTE). Three monographs on DDL for LOTE were presented and are briefly illustrated in this paper (Forti, 2023; Vyatkina, 2024; Schaeffer-Lacroix, 2019). Despite the growth of DDL in second language (L2) research and education in many languages, its focus has largely been on English. As a result, the knowledge about the applicability of DDL for learning LOTE remains limited. This hinders the validity and generalizability of DDL as a whole, and conceals the important implications related to bridging the research vs. practice gap in such contexts where the focus is on LOTES. The monographs introduced in this paper demonstrate the relevance of DDL for learning LOTES by discussing corpus-based resources, pedagogical applications, and empirical research from three perspectives.

Keywords: *corpus linguistics, data-driven learning, German, Italian, languages other than English.*

1. Introduction

DDL is an increasingly popular field in language education and research. Corpus-informed materials for L2 English have grown significantly. However, LOTE resources remain scarce (Jablonkai et al., 2020). This disparity extends to research on the effectiveness of pedagogical corpus applications, where most studies focus on L2 English. Recent research synthesis on DDL reveals that 89% of empirical studies have been conducted in English teaching contexts (Boulton & Vyatkina, 2021). This limits the validity, generalizability, and practical applicability of research findings in multilingual educational settings, and may perpetuate an undesirable monolingual standard. To obtain a more complete understanding of the field and enhance its practical implications in multilingual pedagogical contexts, more corpus applications for LOTE are needed.

Here, the authors address this gap by discussing their monographs that represent first book-length publications entirely devoted to DDL for LOTE. By doing so, the authors report on multi-year projects with or learning scenarios for university students in three different linguistic and geographical settings. The authors advance the DDL field by discussing available DDL resources, pedagogical applications, and empirical research for Italian and German.

2. A review of three monographs on DDL for learning LOTE

2.1. Corpus use in Italian language pedagogy: Exploring the effects of data-driven learning (Forti, 2023)

The use of corpora in Italian language pedagogy has attracted the interest of scholars for at least 30 years (Forti, 2023; Polezzi, 1993). However, only nine empirical studies over a total of 489 have focused on L2 Italian (Boulton & Vyatkina, 2021), and nearly all of them employed questionnaires to elicit learner attitudes and behaviors while engaging with DDL activities, with little insight into language gains. Furthermore, the corpus-based resources used were mostly unavailable to the public, as were the data elicitation tools, consequently hindering replicability of research.

The author's empirical study makes use of publicly available corpora and contains the data elicitation tools used. It was conducted in eight classes of (native language) L1 Chinese students, enrolled in a foundation language course at an Italian university. The eight classes were randomly assigned to either the experimental condition (DDL activities) or control condition (traditional activities). The experimental group consisted of 62 students (female = 47), with age range 18-27, while the control group consisted of 61 students (female = 38), with age range 18-26. A 1-hour lesson focused on verb-noun combinations was taught once a week in the eight classes for eight weeks. At the end of the lessons, a questionnaire, aimed at eliciting learner attitudes towards collocation learning and DDL, was administered. A test evaluating knowledge of the collocations was administered at four-week intervals. The last administration of the test was conducted four weeks after the end of the lessons to measure retention rates.

In terms of language gains, we found U-shaped learning curves in both groups, with no significant differences between them. This may be due to the fact that the length of the pedagogical intervention consisted of fewer than ten sessions: according to a meta-analysis of DDL studies (Lee et al., 2019), interventions with more than ten sessions tend to have a larger effect on language gains. We also identified a tendency towards higher retention rates in the DDL group. This indicates that the specific traits of concordance-based DDL activities (i.e. being able to sift through multiple contextualized examples of a word combination, and then detect regularities) are likely to produce longer-term learning. Semantic transparency, L1-L2 congruency, and dimensions of collocation knowledge were included in the analysis, providing a more nuanced view of the findings. In terms of learner attitudes, some initial difficulties were reported in relation to working with concordance lines. Nevertheless, they recognized the usefulness of the activities in enhancing their awareness about word combinations.

2.2. Corpus applications in language teaching and research: The case of data-driven learning of German (Vyatkina, 2024)

The systematic review part of the monograph provides an overview of the history of DDL for L2 German language and reveals a rich tradition dating back to the late 19th century, with frequency lists for German emerging in the 1920s. Various pedagogical applications for German, including the influential Routledge Frequency Dictionary series, have since evolved. L2 German DDL resources encompass different corpus interaction methods (direct/indirect, hands-on/hands-off), publication formats (reference books, textbooks, tests, monographs, articles), and linguistic areas (vocabulary, grammar, lexico-grammar, pragmatics). While L2 German DDL research is smaller in scale compared to L2 English, it stands as a prominent LOTE target. Numerous L2 German studies (Vyatkina, 2024, Chapter 3) offer detailed insights into DDL implementations, showcasing their effectiveness when learners receive ample support. However, the field calls for increased methodological diversity, a broader range of targeted language skills, consistent reporting, and long-term studies.

The focus in the empirical study part of the monograph was on efficacy of teaching L2 German collocations to U.S. university students by combining Instructed Second Language Acquisition (ISLA) and DDL approaches. The data were collected from one intact group of high-intermediate proficiency learners in an L2 German course who were exposed to both a paper-based and a computer-based DDL treatment and consented to participate in the study. The study compared pretest-posttest gains in lexical, morphological, and collocational knowledge. The results confirmed the effectiveness of explicit interventions in developing productive L2 collocation knowledge, aligning with the usage-based theory of language acquisition (e.g. O'Keefe, 2021). The study also validated DDL as an

effective method for teaching collocations, with both DDL methods bringing significant knowledge gains with an advantage of the computer-based method for morphological knowledge and because of its efficiency. This research extended the scope of DDL beyond English to inflectional languages, demonstrating its applicability to both lexical and grammatical collocations. It emphasized the integration of best practices from ISLA and DDL research and encouraged cross-disciplinary collaboration (O’Keeffe, 2021).

The pedagogical applications part of the monograph addresses how open-access corpora can serve as Open Educational Resources (OERs). ‘Incorporating Corpora’ (Vyatkina, 2020) is presented as one such OER, which is tailored for English-speaking L2 German learners and teachers. It utilizes DWDS (Digitales Wörterbuch der Deutschen Sprache), an open access German corpus and tool suite, offering interactive online exercises linked to DWDS. This OER includes user-friendly tutorials and instructions for both educators and learners, addressing key DDL issues while adhering to web accessibility guidelines (Meunier, 2022). Successful pilot testing with intermediate-level students that affirmed its instructional value is reported.

2.3. Encounters with German language in use promoted by pedagogical corpora (Schaeffer-Lacroix, 2019)

As mentioned in section 2.1, German is the best represented foreign language in DDL research after English. According to Boulton and Vyatkina's (2021) meta-study on DDL articles written in English, eight out of sixteen articles concern L2 German and are designed for learners whose L1 is English; only two papers represent German L2 learners whose L1 is French. If one expands the list of DDL research on German L2 for learners whose L1 is French to studies written in French, more results can be identified: two reports on the creation of corpora for learning L2 German in France (Trouvain et al., 2013; Wigham & Poudat, 2020), ten out of fourteen DDL studies conducted by Schaeffer-Lacroix ¹, and her research presented in this paper.

Here, the author’s monograph presents a brief overview of five corpus-based scenarios designed for teaching German in France in different learning settings. Their scientific background is inspired by researchers such as Bachelard, Bruner, Chanquoy, Sweller, Tricot, Pekarek Doehler, Piaget, and Vygotsky. Pedagogical concepts like discovery learning, intertextual text production, interaction between expert and novice, and language awareness structure its five scenarios, which cover the whole range between low and high instruction levels. Learners’ perceptions of the effectiveness of DDL were identified with the help of filmed interviews led by an external researcher right after the classroom experiments set up for three out of the five scenarios (for details, see Schaeffer-Lacroix 2016, 2018b, 2022 in the list provided in footnote 1). These perceptions were compared to the learners’ activities tracked through filmed computer screens and audio recordings of pair discussions. The author defends the general idea that small, specialized corpora like those created for her five learning scenarios can support DDL research and practice in an effective way, even if, for statistical reasons, Dodd (1997) considers those datasets which contain fewer than one million tokens do not merit the label ‘corpus’. However, in the eyes of foreign language learners, corpora are big enough if they help them find answers to their questions, and students can even be discouraged by a huge amount of data (Schaeffer-Lacroix, 2009, p. 194 & 200). Crosthwaite and Baisa (2023) warn the DDL community not to ignore the growing importance of artificial intelligence applications, and they recommend turning them into partners instead of considering them as a threat for language education. This is one more reason to use small corpora whose content and quality can be (semi-)manually checked; this makes them relevant for a given task, be it for particular language learning needs and contexts or as training material for machine learning.

3. Discussion and conclusions

Overall, these three monographs offer new insights and resources for researchers, language teaching practitioners, and students interested in corpus-based learning and teaching methods (Table 1). It appears that opening the floor for LOTE, not only with respect to the target language but also with respect to the language of publication, offers the opportunity to share the knowledge and the methods stemming from different cultures, to renew the DDL field

¹ [DDL research on German as a foreign language](#)

and to reaffirm its importance at a time when its existence is challenged by artificial intelligence applications.

Table 1. Summary of the monographs.

	Forti (2023)	Vyatkina (2024)	Schaeffer-Lacroix (2019)
Target language	Italian	German	German
Teaching context	Italian L2 language courses for prospective university students in Italy.	German as a foreign language courses in the USA.	German as a foreign language courses in France.
Main objective	To illustrate the main methodological challenges in researching DDL effects and to demonstrate, by means of an empirical study, how such challenges may be addressed.	To address three existing divides in the DDL field between: 1) English and other languages; 2) DDL and ISLA research; and 3) research and pedagogical practice.	To inform on how to integrate DDL activities in language learning scenarios.
Systematic review	DDL teaching materials and research studies.	DDL teaching materials and research studies.	Comparison of DDL activities to other CALL activities. Available corpora and corpus tools.
Empirical study	Comparing the effect of DDL vs. non-DDL activities aimed at developing phraseological competence.	Comparing the effectiveness of hands-on and hands-off DDL for teaching verb-noun collocations.	The effect of DDL on language awareness (prepositions, verb particles, commas) and on learning to write according to the constraints of the given text genre.
Pedagogical applications	Principles and resources to develop DDL activities.	A suite of open access, interactive DDL activities.	Five teaching scenarios.

Publishing research in LOTE supports multilingualism and allows DDL researchers to stick closely to their audience. Using another language than English may moreover help with designing activities inspired by a non-English cultures and avoid biases introduced by the English-language perspective. The authors support the strengthening of the theoretical grounding of DDL research, the integration of pedagogical DDL applications with more learner-friendly user interfaces, and the enhancement of rigor in study design and reporting.

References

Aston, G. (2002). The learner as corpus designer. In B. Kettemann & G. Marko (Eds), *Teaching and learning by doing corpus analysis* (pp. 9–26). Rodopi B.V.

- Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning & Technology*, 25(3), 66–89. <https://www.lltjournal.org/item/10125-73450>
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 100066. <https://doi.org/10.1016/j.acorp.2023.100066>
- Dodd, B. (1997). Exploiting a corpus of written German for advanced language learning. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 131–145). Longman.
- Forti, L. (2023). *Corpus use in Italian language pedagogy: Exploring the effects of data-driven learning*. Routledge.
- Jablonkai, R.R., Forti, L., Castelló, M. A., Iguenane, I. S., Schaeffer-Lacroix, E., Vyatkina, N. (2020). Data-driven learning for languages other than English: the cases of French, German, Italian, and Spanish. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thoučny (Eds.), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 132-137). Research-publishing.net. <https://doi.org/10.14705/rpnet.2020.48.1177>
- Lee, H., Warschauer, M., & Lee, J. H. (2019). 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>
- Meunier, F. (2022). Revamping DDL: Affordances of digital technology. In R. R. Jablonkai & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 344–360). Routledge.
- O’Keeffe, A. (2021). Data-driven learning: A call for a broader research gaze. *Language Teaching*, 54, 259–272. <https://doi.org/10.1017/S0261444820000245>
- Polezzi, L. (1993). Concordancing and the teaching of ab initio Italian language for specific purposes. *ReCALL*, 5(09), 14–18. <https://doi.org/10.1017/S0958344000004067>
- Schaeffer-Lacroix, E. (2009). *Corpus numériques et production écrite en langue étrangère. Une recherche avec des apprenants d’allemand* [Electronic corpora and learning to write in a foreign language] [PhD thesis, Sorbonne nouvelle]. HAL. <https://theses.hal.science/tel-00439095>
- Schaeffer-Lacroix, E. (2019). *Corpus et didactique de l’allemand – La langue à bras-le-corps* [Encounters with German language in use promoted by pedagogical corpora]. Lambert-Lucas.
- Trouvain, J., Laprie, Y., Möbius, B., Andreeva, B., Colotte, V., Fauth, C., Fohr, D., Mella, O., Jügler, J., & Zimmerer, F. (2013). Designing a bilingual speech corpus for French and German language learners. *Proceedings of Corpora and Tools in Linguistics, Languages, and Speech*, 32–34. https://www.coli.uni-saarland.de/~trouvain/Trouvain_et_al_2013.pdf
- Vyatkina, N. (Ed.). (2020). *Incorporating corpora: Using corpora to teach German to English-speaking learners* [Online instructional materials]. University of Kansas, Open Language Resource Center. <https://corpora.ku.edu>
- Vyatkina, N. (2024). *Corpus applications in language teaching and research: The case of data-driven learning of German*. Routledge.
- Wigham, C. R., & Poudat, C. (2020). Corpus complexes et standards : Un retour sur le projet CoMeRe [Complex corpora and standards: a review of the CoMeRe project]. *Corpus*, 20. <https://doi.org/10.4000/corpus.4736>