EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16959

Learner corpora and the design of data-driven learning activities

Luciana Forti[a]

[a]Department of Italian Language, Literature and Arts in the World, University for Foreigners of Perugia, , luciana.forti@unistrapg.it

How to cite: Forti, L. (2023). Learner corpora and the design of data-driven learning activities. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik. https://doi.org/10.4995/EuroCALL2023.2023.16959

Abstract

This paper seeks to continue the conversation on an underused resource in Data-Driven Learning (DDL), namely learner corpora. It first explores the potential of learner corpora in data-driven learning activity design, focusing on the advantages claimed by a number of scholars throughout the years and mainly associated with the field of Learner Corpus Research (LCR). It then illustrates the status that learner corpus use has in DDL activity design, on the basis of data drawn from the most recent and comprehensive review of DDL research, covering a timespan of 30 years. After describing the main qualitative and quantitative design features of a learner corpus of Italian (i.e. the CELI corpus*), the paper shows how learner corpora containing texts produced at different proficiency levels can be used as graded corpora for both target-oriented and error-oriented activities. The sample illustrated activities can thus cater to learners at different proficiency levels, including lower-intermediate and intermediate levels, which are still under-represented in DDL research. Some of the main pedagogical and scientific advantages of using level-specific DDL materials in a paper-based format are also presented and briefly discussed.*

Keywords: *Learner corpora, data-driven learning, Common European Framework of Reference for Languages, Italian.*

1. Introduction

Learner corpora have seen major developments in the past few decades. These developments have concerned not only the techniques used to build them, their accessibility, and the learner languages they represent, but also the methodological sophistication with which the linguistic data they contain is analysed. We, thus, have increasingly greater insight into interlanguage features and dynamics, and this can invaluably enhance our understanding of second language acquisition theories. Learner corpus use remains, however, quite limited in pedagogical settings, such as those involving Data-Driven Learning (DDL).

In this paper, we review the potential and status of learner corpus use in DDL, showing how there is a gap between calls for increased learner corpus use in DDL and its actual use in DDL activities. We then introduce and describe the CELI corpus, a learner corpus of Italian, containing written texts produced in the context of language certification exams across four balanced proficiency levels (from B1 to C2). Finally, we present some ways in which a learner corpus such as the CELI can be used in designing DDL activities, and then briefly discuss some of the related advantages.

2. Potential and status of learner corpora in DDL

In DDL, "[…] the task of the language teacher is to provide a context in which the learner can develop strategies for discovery - strategies through which he or she can 'learn how to learn'" (Johns, 1991, p. 1). While originally DDL developed mainly with reference to the use of L1 corpora, the adjacent and partially overlapping field of Learner Corpus Research (LCR) has argued that even "[…] learner corpora can be extremely useful for form-focused instruction, because they present students with typical interlanguage features" (Gilquin & Granger, 2022, p. 433). In particular, learner corpora can provide data on typical errors, as well as examples representing good language use in cases where no errors are present. In these latter cases, learner corpora with good examples of the target language can be particularly useful when there is a need to match the level of language difficulty to a certain proficiency level.

When used in conjunction with a reference L1 corpus (e.g. Ackerley, 2017), the learner corpus plays an important role in the observation of over- and under-use phenomena, the detection of false friends, the detection of non-idiomatic expressions and the identification of errors. By means of systematic and guided comparison between the two corpora, learners are placed within a context in which they have the tools to discover differences between learner-produced and L1-speaker-produced language. This comparison can be even more effective when the learner corpus is a *local* learner corpus (Seidlhofer, 2002). Local learner corpora contain the texts produced by the same learners who will then analyse them in the context of DDL activities. The advantage of this kind of corpus exploration is that learners can analyse features of their own interlanguage through the various data types offered by a corpus (concordances, frequency lists, dispersion graphs, etc.), and this can provide a considerable boost for motivation, since the learners are actively involved in creating a corpus from their own pieces of writing. The many potential and advantageous uses of learner corpora in DDL are very much present in the literature and in the conference presentations, especially in contributions from recent years (Gilquin, 2023; Götz, 2022).

But to what extent are learner corpora actually used in DDL? To answer the question, we can filter the data contained in Boulton & Vyatkina (2021)'s 30-year review on DDL research according to the type of corpus used. The findings show that only 21 of the total of 489 papers collected by the authors involve the use of a learner corpus, which corresponds to a mere 4.29%. Furthermore, in most cases, the learner corpus seems to be a local learner corpus and its use is implemented in conjunction with a reference corpus.

We will now describe the features of an off-the-shelf learner corpus of Italian, and its consequential potential uses in DDL, in the hope of further shedding light on the usefulness of learner corpora in DDL activity design, despite their limited presence in DDL research and pedagogy so far.

3. The CELI corpus

The CELI corpus (Spina et al., 2022, 2023)[1] is a learner corpus of Italian. It consists of written texts collected from the Italian language certification exams known as CELI, *Certificati di Lingua Italiana*[2]. The texts were collected from passed language certifications exams at levels B1, B2, C1, and C2. Texts from lower levels of proficiency were not collected as they were very few, very brief, and the product of highly guided tasks, which would have made it difficult to isolate the actual independent contribution of the learner. Each text, originally written on paper, was manually transcribed on a digital platform. Formal aspects which could have potentially hindered the quality of the post-tagging were normalised. The total number of texts is 3,041, which amount to about 600,000tokens. Each proficiency level contains a comparable quantity of tokens. The nationality backgrounds of the test subjects are varied, with Greek, Spanish, and Romanian being the three most represented ones. The search interface allows selection of not only the proficiency level and the nationality background, but also the age group (from 10-14 to

[1] The corpus can be accessed at the following two webpages: https://lt.eurac.edu/cqpweb/, https://apps.unistrapg.it/cqpweb/. (last accessed: 31/07/2023).

[2] The CELI exams are developed and administered all around the world by the CVCL – Centro per la Valutazione e le Certificazioni LInguistiche, based at the University for Foreigners of Perugia. More information about the CELI language certification system may be found at the following webpage: https://www.unistrapg.it/node/457 (last accessed: 31/07/2023).

75-79), the exam centre location (Abroad/Italy), the task number, the sex, the text genre, and the text type. The CELI is one of the very few learner corpora balanced according to proficiency level, and in which text attribution to a certain level is based on a sound framework, which in this case is provided by the *Common European Framework of References for Languages* (CEFR) (Forti, 2023). Table 1 summarises the main qualitative design features of the CELI, while Table 2 summarises its main quantitative features.

Table 1. Qualitative design features of the CELI corpus (based on classification by Tono, 2003).

Language-related features	Task-related features	Learner-related features
Mode: written.	Data collection: pseudo-longitudinal.	Internal-cognitive: age. Internal-affective (motivation/attitude): n/a.
Genre: mixed (article, blog, e-mail, essay, letter, report, story).	Elicitation: passed language certification exams.	L1 background: by approximation (i.e. nationality; n. 104, mostly Greek, Spanish, Romanian)
Style: mixed (argumentative, descriptive, narrative and mixed).	Use of references: no.	L2 environment: L2/FL.
Topic mixed (leisure, current affaires, etc.)	Time limitation: fixed.	L2 proficiency: yes, based on certification exam.

Table 2. Quantitative features of the CELI corpus

CEFR level	Number of texts/learners	Number of tokens
B1	1,212	156,612
B2	840	152,251
C1	585	149,859
C2	404	149,892
TOTAL	3,041	608,614

4. Using the CELI corpus for graded DDL activities

This section contains some examples of how a balanced learner corpus, based on the CEFR, can provide useful data for both target-oriented and error-oriented DDL activities. The former may be designed as guided-discovery activities, aimed at guiding the learners toward the observation of form-meaning patterns in the concordance lines. The latter may be designed to aim for the learner to detect the error within a set of concordance lines. In both cases, the examples within the concordance lines which learners will engage with will be suitable for the particular proficiency level at which they are.

One area where learners of L2 Italian struggle with is *andare* (‘to go’) + PREPOSITION | ARTICLE + NOUN costruction. Most errors are found in the choice of the preposition and/or article that needs to be used between the verb and the noun. By extracting the data related to this construction from the CELI corpus, we are able to obtain good examples of error-free sentences containing this construction, as well as errors concerning the construction. Appendix 1 provides a table containing a line of concordance lines extracted from the CELI corpus. The table is divided into four sections, one for each proficiency level. Within each proficiency level, we see a section with good examples and one with an erroneous example. With the former, we may ask the learners to try and detect any regularities that are typical when the target construction is used. Each learner will be able to engage with concordance lines that are suitable for their level. The example containing the error, on the other hand, may be used in activities where the learners are presented with a set of concordance lines, within which one example contains the error. The task for the learners will be to identify which concordance line contains the error. In trying to reach this goal, the learners will apply the meta-cognitive strategies that are typical of DDL, such as scanning, comparing, and making hypotheses.

Other kinds of activities can be developed, such as matching multiple split sentences or filling the gap in a provided set of concordance lines. In each of these cases, the activities may be uploaded online by using one of the many applications that are available and suitable for this purpose. However, these activities can also be paper-based (Boulton, 2010). This choice carries with it at least four major advantages: 1) once the activities have been created, they can be re-used indefinitely, so the initial time spent developing them will be time saved in the future; 2) though relying on data extracted from corpora, these activities will not require the use of computers on the learners’ part, and this can be particularly useful with large groups or in contexts where access to computers may still be challenging; 3) the fact that concordance lines are pre-selected by a teacher means that even corpora that do not have a learner-friendly interface can be used for DDL activities; and 4) re-usable DDL activities facilitate replication in empirical investigations aimed at evaluating the effects of DDL in terms of language gains, learner and teacher attitudes, and learning processes that they activate.

5. Final remarks and conclusions

In this short paper, we aimed at continuing the conversation on the potential of learner corpora in DDL activities. After illustrating the potential and status of learner corpus use in DDL research, we described the main design features of the CELI corpus, a learner corpus of Italian. We then briefly presented some ideas on how a corpus such as the CELI corpus may be used in DDL.

The potential of a learner corpus in DDL activity development is inextricably linked to its design. The variety of designs on which learner corpora are based determines not only the potential but also the limitations that a learner corpus will have with regard to its applicability to DDL activity development. A corpus containing balanced subcorpora of different proficiency levels has the specific advantage of potentially catering for learners at those different levels, thus offering the opportunity to have graded input both in relation to good examples of language, as well as in relation to the typical errors that may be found in the language production of learners at different proficiency levels.

We look forward to further discussion on the topic of learner corpus use in DDL activity design and its many connections to other relevant topics in the field, such as the need for grading DDL activities, the pedagogical and scientific advantages of paper-based DDL materials, and the ‘learner-friendliness’ of corpus-based resources.

References

Ackerley, K. (2017). Effects of Corpus-Based Instruction on Phraseology in Learner English. *Language Learning & Technology*, *21*(3), 195–216.

Boulton, A. (2010). Data-Driven Learning: Taking the Computer Out of the Equation: Data-Driven Learning. *Language Learning*, *60*(3), 534–572.

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

Forti, L. (2023). *Exploring the affordances of CEFR-based learner corpora in Data-driven learning* [Plenary presentation], Japan Association for English Corpus Studies (JAECS) Spring Forum 2023, JAECS SIG on DDL, online, May 13, 2023.

Gilquin, G. (2023). Written learner corpora to inform teaching. In R. R. Jablonkai & E. Csomay (Eds), *The Routledge Handbook of Corpora and English Language Teaching and Learning* (pp. 281–295).

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe & M. J. McCarthy, *The Routledge Handbook of Corpus Linguistics* (2nd ed., pp. 430–442). Routledge. https://doi.org/10.4324/9780367076399-30

Götz, S. (2022). *Learner corpora and DDL: A Promising Synergy?* [Paper presentation in Symposium], EUROCALL CorpusCALL SIG symposium, Online.

Johns, T. (1991). Should you be persuaded – two examples of data-driven learning materials. *Classroom Concordancing, English Language Research Journal 4*, 1–16.

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* (Vol. 6, pp. 213–234). John Benjamins. https://doi.org/10.1075/lllt.6.14sei

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il Corpus CELI: Una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, *14*(1), 116–138. https://doi.org/10.54103/2037-3597/18161

Spina, S., Fioravanti, I., Forti, L., & Zanda, F. (2023, in print). The CELI Corpus: Design and linguistic annotation of a new online learner corpus. *Second Language Research*. https://doi.org/10.1177/02676583231176370

Tono, Y. (2003). Learner corpora: Design, development and applications. *Paper Presented at the Corpus Linguistics 2003 Conference (CL 2003) Lancaster*, 800–809.

Appendix 1. CELI corpus data for cline of DDL activities focused on the *andare* (‘to go’) + PREPOSITION | ARTICLE + NOUN costruction.

B1		
Target-oriented examples	Prima abbiamo partecipato alla cerimonia e dopo **siamo andati al ristorante**. **Sono andata nel parco** vicino alla mia università. Adesso posso **andare al lavoro** in bicicletta e non devo andare a piedi. Nel tempo libero, mi piace **andare nel centro storico**. Penso **di andare alla fiera** anche il prossimo anno.	Main features: short sentences; literal meanings.
Error-oriented example	Sono **andata sulla cerimonia** di premiazione.	Error type: wrong selection of preposition.
B2		
Target-oriented examples	Ho deciso di **andare all'estero** per studiare. Mi è venuta una grande depressione, e insomma, sono dovuta **andare dallo psichiatra e dalla psicologa** per aiuto. Il giorno dopo la separazione **sono andata al lavoro**, ma c'era solo il mio corpo. Tutti i lunedì **andavamo al mercato** insieme. Avevo deciso di sposarla, in futuro. Ho visitato la città di Bangkok e **sono andato sull'isola** Kho Tao.	Main features: literal meanings.
Error-oriented example	Ho provato a continuare le cose che amavo fare, **come andare nelle montagne**, incontrare gli amici.	Error type: wrong selection of preposition.
C1		
Target-oriented examples	Una foto qui un'altra là e **via vanno la privacy e il contenuto** delle storie dietro una foto. Abbiamo preso il tram **e siamo andate ai magazzini**. Abbiamo deciso con mio marito di **andare alla laguna** dei sette colori che si chiama Bacalar in Messico. La mia lettera successiva **andrà al Presidente**. Per commentare la gente doveva scrivere su un foglio, **andare nell'ufficio postale** e finalmente inviare la lettera.	Main features: some non-literal meanings.
Error-oriented example	Puoi **andare sui passi** di Dracula ed incontrarlo magari nei tuoi sogni.	Error type: non-idiomatic expression.
C2		
Target-oriented examples	Si rese conto che doveva spiegare alla sua miglior amica **come erano andate le cose** l' altra sera . Non appiattarsi **e lasciarsi andare all’ozio** ma smuovere le risorse del cervello. **Siamo andati dal Giudice**, io e l' altro autista. Avevo deciso di **andarci col treno** nonostante i numerosi cambiamenti tra Genova e la Sicilia. Con tutte le tasse che stiamo pagando e una percentuale dovrebbe **andare alle casalinghe**.	Main features: literal and non-literal meanings; idiomatic expressions.
Error-oriented example	Questa nuova abitudine di fidarti di un algoritmo automatizzato piuttosto che **andare l'avventura** - sia amicale o romantica - nella vita reale.	Error type: error in idiomatic expression.