

Different Registers, same Learners: Towards a Multi-Register Corpus of Learner English

Sylvie De Cock
Centre for English
Corpus Linguistics
UCLouvain, Belgium
sylvie.decock
@uclouvain.be

Gaëtanelle Gilquin
Centre for English
Corpus Linguistics
UCLouvain, Belgium
gaetanelle.gilquin
@uclouvain.be

Sylviane Granger
Centre for English
Corpus Linguistics
UCLouvain, Belgium
sylviane.granger
@uclouvain.be

Pauline Jadoulle
Centre for English
Corpus Linguistics
UCLouvain, Belgium
pauline.jadoulle
@uclouvain.be

Magali Paquot
Centre for English
Corpus Linguistics
UCLouvain, Belgium
magali.paquot
@uclouvain.be

Abstract

This paper presents a learner corpus compilation project carried out within the frame of the CLARIN Knowledge Centre for Learner Corpora. The corpus includes data representing different registers (both written and spoken) produced by the same group of learners. This specificity makes it possible to control for individual styles when investigating the effect of register on the linguistic features of learner language, unlike most register studies in learner corpus research which rely on different registers produced by different groups of learners. When completed, the learner corpus will be made available through the Corpora@UCLouvain platform, which is recorded in the CLARIN Virtual Language Observatory.

1 Introduction

Register variation is a crucial aspect of language production. Depending on the context in which it is used and the communicative purposes that it serves, language tends to display distinctive characteristics, which may have to do with lexis, but also phraseology or syntax, among others (see, e.g., Biber 2012). Corpora, in particular, have raised awareness of register variation and have provided insights into linguistic patterns associated with different registers. Specific methods based on corpora have also been developed to study register variation, most notably multi-dimensional analysis (Biber 1988). This paper focuses on the corpus study of register variation in learner language. It considers the limitations of register studies in learner corpus research and describes the design of a learner corpus meant to represent multiple registers produced by the same learners and compiled within the frame of the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA).

2 Register Studies in Learner Corpus Research

While register is important for any type of language, it is particularly relevant to learner language, because learners may not show the same register awareness as native or expert writers/speakers (cf. Gilquin & Paquot 2008). Learner corpus research has taken register into account in the sense that studies are carried out on the basis of learner corpora representing certain registers (e.g. telecollaborative discourse in Vyatkina 2012). However, studies comparing learner language registers are still relatively rare. One

reason for this is that, until recently, learner corpora represented only a small range of registers, most notably argumentative essays for writing (as in the International Corpus of Learner English; Granger et al. 2020) and interviews for speech (as in the Louvain International Database of Spoken English Interlanguage; Gilquin, De Cock, & Granger 2010). Over the last decade or so, however, learner corpora representing various registers have started to become available, including language for specific purposes learner corpora (e.g. the Active Learning of English for Science Students Learner Corpus; Allen 2012) and learner translation corpora (e.g. the Multilingual Student Translation corpus; Granger & Lefer 2020). Some learner corpora have also appeared that combine several text types, although the range tends to be rather limited (e.g. only written texts in the Tracking Written Learner Language corpus; Hasund, Drange, & Torjusen 2022).¹

Using such learner corpora, a few studies have compared two or three registers, such as speech and writing in Fuchs, Götz, & Werner (2016) or argumentative essays and research papers in Larsson, Paquot, & Biber (2021). A couple of studies have also sought to situate a learner language register among various native/expert registers by comparing their linguistic features (cf. Aguado-Jiménez, Pérez-Paredes, & Sánchez 2012). Importantly, when learner corpus researchers have compared different registers, it has mainly been on the basis of texts produced by different learners (e.g. argumentative essays produced by one group of students and interviews produced by another group). A possible issue with this method is that individual writers'/speakers' styles may affect the comparison of registers. In other words, differences between registers may appear simply because different writers/speakers use language differently.

3 A Multi-Register Learner Corpus

As part of a project on register variation and learner language, we have been compiling a learner corpus made up of texts from different registers produced by the same learners. This corpus will make it possible to compare registers while controlling for individual styles, and to investigate the effect of register on the linguistic features of learner language.

The learner data are collected at UCLouvain among (mainly) French-speaking learners of English who are students in their second year of English major studies. These students are required to produce written and spoken texts in English representing different registers. The written registers include (i) a career readiness essay in which students discuss the career they have chosen as well as its potential benefits and drawbacks, (ii) a cover letter which students write in response to a (student) job advertisement, and (iii) a persuasive essay which requires students to propose a solution to a problem they have chosen and persuade readers to agree with their solution. The spoken registers include (i) a non-interactive, monologic task in which students have to explain where they see themselves in 5 years' time, professionally speaking, (ii) an interactive task which takes the form of a job interview with the teacher, and (iii) an interactive task in pairs of students which involves debating the topics they chose for their persuasive essays. It will be noticed that the written and spoken registers are closely linked to each other: the career readiness essay is related to the monologue on students' professional future, the cover letter is the written equivalent of the job interview, and the persuasive essay corresponds to the debate in pairs of students. This should enhance the comparability of the written and spoken registers.

We have also made special efforts to collect rich metadata about the learners and the tasks. Our starting point has been Frey et al.'s (2023) and Paquot et al.'s (2024) Core Metadata Schema for Learner Corpora, whose aim is to increase the standardization of learner corpus metadata. We have included information about, among others, learners' knowledge of languages, their exposure to English in different situations, but also their literacy. Detailed information about the tasks (e.g. instructions, time constraints and use of language reference tools for writing) and the registers (e.g. communicative purposes, settings, number of addressees) is also provided. A learner ID makes it possible to link together all texts produced by one and the same student.

¹ See CLARIN's L2 Learner Corpora resource (<https://www.clarin.eu/resource-families/L2-corpora>) or the Centre for English Corpus Linguistics' Learner Corpora around the Word webpage (<https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>) for a list of existing learner corpora.

So far, data have been collected among one cohort of students, representing about 60 learners. The data collection will continue among future cohorts of students, until a sufficient amount of data has been collected. Once completed, we aim to make the learner corpus available to the research community, using a FAIR (Findable, Accessible, Interoperable, and Reusable) approach (see Lindström Tiedemann, Lenardič, & Fišer 2018 and König, Frey, & Stemle 2021). The data will be released in open access format through the Corpor@UCLouvain platform, which is recorded in the CLARIN Virtual Language Observatory (VLO).

This learner corpus is one of the first corpus compilation projects at UCLouvain since CKL2CORPORA was officially recognized in November 2022. The project brings together several members of CKL2CORPORA. It thus gives us the opportunity to carry out some of CKL2CORPORA's missions, such as testing and improving Paquot et al.'s (2024) Core Metadata Schema for Learner Corpora and establishing best practices for data formats (including speech transcriptions).

4 Conclusion

The compilation of our multi-register corpus of learner English, carried out within the frame of CKL2CORPORA, is meant to provide researchers with the necessary resources to study how learner language varies according to register. Not only does it represent an improvement on the resources that are usually exploited for the investigation of registers in learner language, including different registers produced by the same group of learners rather than by different groups of learners, but it also contributes to the establishment of best practices in aspects such as metadata standardization and data formats.

Acknowledgements

We acknowledge the financial support of UCLouvain's Hoover Seed Fund for this project.

References

- Aguado-Jiménez, P., Pérez-Paredes, P., & Sánchez, P. (2012). Exploring the use of multidimensional analysis of learner language to promote register awareness. *System*, 40(1), 90-103.
- Allen, D. (2012). Active Learning of English for Science Students (ALESS): A personal introduction. *Proceedings of New Approaches to English Language Education for Students of Science and Engineering in Japan, June 10, 2010*. Published March 26, 2012, pp. 33-37. Waseda University.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9-37.
- Frey, J.-C., König, A., Stemle, E. W., & Paquot, M. (2023). Core Metadata Schema for L2 data. Paper presented at the 2nd Conference of the European Second Language Association (EUROSLA), University of Birmingham, 30 August 2023 – 02 September 2023.
- Fuchs, R., Götz, S., & Werner, V. (2016). The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In Werner, V., Seoane, E., & Suárez-Gómez, C. (eds) *Re-assessing the Present Perfect*, pp. 297-338. De Gruyter.
- Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain.
- Gilquin, G. & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1), 41-61.
- Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.
- Granger, S. & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: A resource for translation teaching and research. *Language Resources and Evaluation*, 54, 1183-1199.
- Hasund, I. K., Drange, E.-M., & Torjusen, H. M. H. (2022). Background, context and studies of the TRAWL corpus. *Nordic Journal of Language Teaching and Learning*, 10(2), I-XVIII.

- König, A., Frey, J.-C., & Stemle, E. W. (2021). Exploring reusability and reproducibility for a research infrastructure for L1 and L2 learner corpora. *Information*, 12, 199.
- Larsson, T., Paquot, M., & Biber, D. (2021). On the importance of register in learner writing: A multi-dimensional approach. In E. Seoane & D. Biber (eds) *Corpus-based approaches to register variation*, pp. 235-258. John Benjamins.
- Lindström Tiedemann, T., Lenardič, J., & Fišer, D. (2018). L2 learner corpus survey : Towards improved verifiability, reproducibility and inspiration in learner corpus research. *Proceedings of CLARIN Annual Conference 2018, Pisa, Italy, 8-10 October 2018*, pp. 146-150.
- Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, 10(2).
- Vyatkina, N. (2012). Applying the methodology of learner corpus analysis to telecollaborative discourse. In Dooly, M. & O'Dowd, R. (eds) *Researching online foreign language interaction and exchange: Theories, methods and challenges*, pp. 267-303. Peter Lang.