Larsson, T., & Gilquin, G. (2026). Corpus analysis of learner language. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*, 2nd ed. Wiley-Blackwell.

Corpus Analysis of Learner Language

Tove Larsson[1] & Gaëtanelle Gilquin[2]
[1]Northern Arizona University, [2]UCLouvain

Abstract
Researchers are increasingly applying corpus techniques to study learner language. This article provides an overview of such techniques and research designs, focusing specifically on those that are directly relevant to the study of learner production. We also cover domains within which most learner corpus research has been carried out. Such domains include lexis, phraseology, grammar, and linguistic complexity.

Keywords: Learner corpus research, corpus linguistics, language learners, study design, research domains.

Introduction

Learner corpus research relies on the analysis of corpora to study features of learner language. While it started "as an offshoot of corpus linguistics" (Granger, Gilquin, & Meunier, 2015, p. 1) and has continued to be influenced by the developments in corpus linguistics, it can be said to constitute a field of its own, with specific areas of interest and distinctive methods. It is also characterized by a high degree of interdisciplinarity, because of its clear links with other fields such as second language acquisition and foreign language teaching.

In what follows, we provide a brief overview of the field. We start off with a discussion of common study designs and techniques. We then outline the main domains of study. Finally, we provide an outlook for the field.

Study Design

This section considers important aspects of the design of studies in learner corpus research. It focuses on some central methods and designs.

Methodological Frameworks Specific to Learner Corpus Analysis

Most methods used in corpus linguistics can be applied to the analysis of learner corpora. Some methodological frameworks, however, have been designed specifically to deal with learner language. An example of this is contrastive interlanguage analysis (CIA; Granger, 1996; Granger, 2015). It relies on (a) the comparison of learner language (L2) with one or several reference varieties (native language, expert writing, etc.) and/or (b) the comparison of several learner language varieties. The first type of comparison makes it possible to establish whether learners use the target language in a similar way to or differently from the reference variety/varieties. Differences may be qualitative (e.g., using *on the contrary* with the meaning and functions of *by contrast*) or quantitative (e.g., using fewer instances of the passive voice or using *however* more often in sentence-initial position). Thanks to the second type of comparison, we can identify differences and similarities across groups of learners. Most

studies using this framework compare learners from distinct mother tongue (L1) backgrounds, with a view to highlighting cases where transfer from the L1 might be at work.

Another example of methodological framework specific to learner corpus analysis is computer-aided error analysis (CEA; Dagneaux, Denness, & Granger, 1998). Studies wishing to apply this method would start from an error-tagged version of a learner corpus, in which errors have been identified and classified according to error type. The analysis then involves looking at these errors in context and comparing them with correct uses of the same linguistic items. Quantifying errors and determining the proportion of correct vs. incorrect uses can help us identify areas that learners struggle with.

Automated Tagging and Manual Coding

Most studies of learner language involve some kind of automated tagging or manual coding of data. That is, we add information to a full text or a text excerpt to be able to carry out further linguistic analysis. To automatically tag corpora, we typically run a computer program to add information about, for example, the part-of-speech of all the words in a corpus. However, while computers are always reliable in that they make the same decision every time when presented with the same situation, they are not always accurate: the computer could be doing the wrong analysis with perfect reliability. Therefore, we want to make sure that we test and report accuracy using two measures: precision (the proportion of the cases that were correctly classified into category X) and recall (the proportion of the actual cases that belong to category X that were correctly classified as category X).

We can use either a tagged or an untagged corpus to add further information to it through manual coding. We may, for example, go through and classify instances of adverbs into categories based on where in the clause they occur (e.g., clause-initial, clause-medial), or we may go through all occurrences of *clear* produced by a group of learners to select for further analysis only the ones that are used for emphasis (i.e., where *clear* means *evident* rather than *transparent*). To ensure that our manual coding is as reliable as possible, we need to (a) use well-developed coding schemes where all the categories are defined and described in detail and (b) test for inter-rater and/or intra-rater reliability to be able to identify problematic categories or sets of tokens and then use this information to further refine our coding scheme (see, e.g., Hober, Dixon, & Larsson, 2023).

Metadata

Learner corpora tend to come with rich metadata that provide detailed information about the learner (age, L1, proficiency level, stays in target language countries, etc.) and the task (prompt, time limit, access to reference tools, etc.). These metadata can be used to select a corpus sample that fits our specific research purpose (e.g., to extract texts written with no access to reference tools in order to determine what learners can produce based solely on their knowledge of the language) or simply to make the corpus sample more homogeneous.

One or several variables stored as part of the corpus metadata can also be studied in relation to specific linguistic features of interest. We could, for example, study the effect of proficiency on lexical sophistication, testing the hypothesis that advanced learners will use more sophisticated vocabulary than intermediate learners. Another example that involves

several learner and task variables is Götz's (2019) study that looked at the effect of factors such as the learners' L1, gender, and their relationship with the interviewer on the use of filled pauses by learners of English.

Cross-Sectional vs. (Quasi-)Longitudinal Designs

Many studies in learner corpus research adopt a cross-sectional design, meaning that they use data produced by different groups of learners at one point in time (or within a limited time period). Typically, these groups of learners differ according to one or several variables and the study involves comparing the groups and establishing the possible effect of these variables (e.g., learners who spent time in a target language country and learners who did not).

Some study designs are longitudinal: they rely on data produced by the same learners over a period of time, so-called longitudinal learner corpora. Such corpora include texts collected at regular intervals (e.g., once every other month or every year) from a certain group of learners. Longitudinal designs make it possible to investigate language development by comparing learners' language use at a certain point in time with their language use at a later point in time. Biber et al. (2020), for instance, investigate the development of grammatical complexity among a group of learners of English tracked over a period of two years.

Longitudinal development can also be studied on the basis of a cross-sectional learner corpus, provided the corpus includes data produced by learners representing various proficiency levels. In this case, development is tracked by comparing groups of distinct learners with different proficiency levels rather than comparing the same learners at different proficiency levels. Such designs are referred to as quasi- or pseudo-longitudinal designs.

Quantitative and Qualitative Analyses of Learner Data

Studies of learner data tend to involve both quantitative and qualitative analysis. For the quantitative analysis, we often use corpus tools (e.g., concordancers) to extract data and statistical programs or environments (e.g., *R*) to run descriptive and inferential statistics. Commonly used statistical techniques in the field include those that enable us to compare groups (e.g., Chi-square tests, *t*-tests), to look at associations among variables (techniques from the regression family), to investigate the predictive power of a set of variables (e.g., random forests), and to look at association among words and phrases (e.g., mutual information). There are also techniques that were developed specifically for CIA-type analyses, most notably MuPDAR (Gries & Deshors, 2020).

While statistical techniques remain valuable resources, for example, to help us identify patterns and see to what extent they can be generalized to a population of interest, they can never (and should never) replace careful and thorough linguistic description. That is, we should strive to move beyond numbers and always return to the language of the texts analyzed. For this step, qualitative analysis plays an important role. As part of this step, we can look more closely at the language that makes up the corpus to see what might explain the quantitative results.

Domains of Study

This section presents some of the main domains that are investigated in learner corpus research: lexis and phraseology, grammar, linguistic complexity, discourse and pragmatics, phonology and other speech-related aspects, and translation. It briefly describes each domain and highlights some key findings.

Lexis and Phraseology

Lexis and phraseology (i.e., the co-occurrence of lexical items) are among the most researched domains in learner corpus research. One reason for this is that vocabulary can provide valuable insights into learner language development. Another, more practical reason is that lexical items – in isolation or in combination – can be retrieved relatively easily from corpora. Concordances, keyword analysis, and lists of words, of collocations, or of lexical bundles are all techniques that are well-suited for the analysis of lexis and phraseology.

Learner corpus studies have revealed a tendency among learners to exhibit overreliance on a small set of high-frequency words (e.g., *think*, *people*), which Hasselgren (1994) refers to as 'lexical teddy bears'. It also appears that, while learners use recurrent combinations of words, these combinations sometimes diverge from those found in expert texts (e.g., using *on the other side* where experts would use *on the other hand*). In some cases, a plausible explanation for such divergent uses is L1 influence, whereby learners translate an expression verbatim from their L1.

Grammar

Another area that has received a fair bit of attention is grammar. Studies have covered topics such as the use of individual grammatical categories (e.g., modal verbs, adverbs, and motion verbs) as well as grammatical patterns and constructions (e.g., the causative construction, extraposition, and genitive alternation). In addition to extralinguistic factors (e.g., L1 background), studies in this domain often also consider the impact of linguistic factors such as the context in which the grammatical feature was produced (e.g., length of subject, semantic category of the head noun). This type of study design can be further subdivided into studies that investigate the impact of extralinguistic or linguistic variables on the use of a grammatical category or pattern in isolation, versus those that use the extralinguistic or linguistic variables to predict a specific variant among two or more competitors (e.g., to answer questions such as "what factors impact the choice of the *of*-genitive over the *s*-genitive?"). As noted for lexis and phraseology above, a common finding in studies of grammatical features is that L1 transfer is relatively widespread, especially in less proficient learner speech or writing (see, e.g., van Vuuren, 2017).

Linguistic Complexity

Studies of learner language have focused on three main aspects of linguistic complexity: grammatical complexity (e.g., Biber, Larsson, & Hancock, 2023), lexical complexity (e.g., De Clercq, 2015), and phraseological complexity (e.g., Paquot et al., 2022). While complexity has been studied in relation to variables such as register and task, most studies use

these concepts to predict and/or describe language development, with the expectation being that as someone's language proficiency improves, they are likely to (a) use more complex grammatical, lexical, and phraseological features and (b) better gage what features are considered appropriate for a given context (e.g., speech vs. writing). Studies of grammatical complexity have found that beginners tend to start out primarily relying on main clauses with no subordination and simple noun phrases (i.e., without any pre- or post-modification) and then move on to producing embedded clauses (e.g., relative clauses) and pre-modifying adjectives. In terms of lexical complexity, research has shown that beginners tend to primarily use high-frequency words with simple morphological make-up, whereas more advanced users rely on a broader repertoire of low-frequency, specialized words. Finally, phraseological complexity covers the study of the variety and sophistication of linguistic co-occurrences (e.g., adjective + noun combinations, verb + direct object). Based on previous research, we may expect lower-proficiency learners to make use of a more restricted set of lexical items and primarily use high-frequency co-occurrences, whereas more advanced learners can be expected to have access to a more diverse set of lexical items and combinations thereof.

Discourse and Pragmatics

While less commonly studied than lexis and grammar, there is a substantial body of research on discourse in the context of learner language. Studies in this subfield focus on language use 'beyond the sentence', meaning that they focus on the larger context in which the language is produced to understand how it affects the meaning of the sentence and how it is perceived. Topics include metadiscourse (i.e., the commentary made on a text by the author; e.g., *frankly*, *to sum up*), discourse markers (e.g., *well*, *I mean*, *so*), and computer-aided conversation analysis more broadly. Previous research has, for example, shown that learners tend to make very frequent use of metadiscoursal features, thus being more explicitly visible in argumentative writing than native and expert writers.

One, in some ways, closely related topic that has received far less focus is pragmatics. Studies in this line of research have tended to apply a word-based approach where the researcher relies on a predefined list of words that have been found to be associated with specific pragmatic functions. Vyatkina and Cunningham (2015) give the example of *please* as a proxy for the speech act 'request'.

Phonology and Other Speech-Related Aspects

Speech in learner corpus research is mostly investigated on the basis of transcripts which reproduce the words uttered by the speakers and which are analyzed for the same aspects as written learner corpora (lexis, grammar, etc.). However, there is also research on acoustic features such as vowel elongation, resyllabification, or stress placement. Such studies often rely on programs such as Praat or ELAN, which allow for tiers of annotation to be added (e.g., syllables, phonemes) and aligned with the acoustic signal when sound files are available. These tools usually also provide automatic measures (e.g., duration of pauses) and visualizations (e.g., spectrograms) that can facilitate the analysis of speech-related aspects.

Different types of speech-related phenomena have been analyzed. Prominent among these is fluency. The analysis of multiple features (e.g., speech rate, filled and unfilled pauses,

repeats) helps establish learner fluency profiles which can then be compared across proficiency levels, with native speakers, etc. (see Götz, 2013). Phonetic and phonological features, both at the segmental level (e.g., pronunciation of phonemes) and the supra-segmental level (e.g., intonation), have also been investigated, although for this type of analysis, it is particularly crucial that the sound files are of very high quality, which may exclude several of the existing spoken learner corpora.

Translation

An emerging domain is learner translation corpus research, which examines translations produced by learners of the source language (i.e., the language from which a text is translated) or the target language (i.e., the language into which a text is translated). Learner translation corpora are more constrained than most learner corpora, because the language production is determined by the contents of the text to be translated. However, they allow for the linguistic study of a specific task – translation – that is central for many language learners.

Learner translation corpus research has focused on two main aspects (see Granger & Lefer, 2023), namely the study of translation features, such as explicitation or simplification, and the assessment of translation quality (including through error analysis) for teaching applications. Yet, any feature of language (collocations, complexity, etc.) can in principle be investigated in learner translation corpora. Speech-related aspects can be explored on the basis of the few interpreting learner corpora that exist to date.

Looking Ahead

While there is overlap between the tools and techniques used by corpus linguists who do not do research on learner language and those who do, there are, as we have seen above, also techniques that have been developed specifically with learner corpus data in mind. Similarly, while there is overlap between topics covered in other research traditions (e.g., second language acquisition) that do not typically use corpus linguistic methods and those commonly investigated in learner corpus research, there are topics that are primarily analyzed in the latter tradition. Researchers interested in corpus analysis of learner language can thus draw from a rich source of knowledge – from research from inside the field as well as from neighboring fields.

The field of learner corpus research is still developing. Moving forward, it would seem beneficial if we could keep advancing the field by striving to improve our methods and topic knowledge based on existing work inside and outside the field, while always maintaining a strong focus on language use and research questions of strong linguistic relevance.

References

Biber, D., Larsson, T., & Hancock, G. (2023). The linguistic organization of grammatical text complexity: Comparing the empirical adequacy of theory-based models. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2023-0016

Biber, D., Reppen, R., Staples, S., & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English

university students. *International Journal of Learner Corpus Research*, *6*(1), 38–71. https://doi.org/10.1075/ijlcr.18007.bib

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

De Clercq, B. (2015). The development of lexical complexity in second language acquisition: A cross-linguistic study of L2 French and English. *EUROSLA Yearbook*, *15*(1), 69–94. https://doi.org/10.1075/eurosla.15.03dec

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.

Götz, S. (2019). Do learning context variables have an effect on learners' (dis)fluency? Language-specific vs. universal patterns in advanced learners' use of filled pauses. In L. Degand, G. Gilquin, L. Meurant & A. C. Simon (Eds.), *Fluency and disfluency across languages and language varieties* (pp. 177–196). Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research – past, present and future. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1–5). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9781139649414.001

Granger, S., & Lefer, M.-A. (2023). Learner translation corpora: Bridging the gap between learner corpus research and corpus-based translation studies. In S. Granger & M.-A. Lefer (Eds.), Learner translation corpus research. Special issue of the *International Journal of Learner Corpus Research*, *9*(1), 1–28. https://doi.org/10.1075/ijlcr.00032.gra

Gries, S. Th., & Deshors, S. (2020). There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, *44*, 69–96. https://doi.org/10.2478/icame-2020-0003

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Hober, N., Dixon, T., & Larsson, T. (2023). Toward increased reliability and transparency in projects with manual linguistic coding. *Corpora*, *18*(2), 245-258. https://doi.org/10.3366/cor.2023.0284

Paquot, M., Gablasova, D., Brezina, V., Naets, H., Lénko-Szymańska, A., & Götz, S. (2022). Phraseological complexity in EFL learners' spoken production across proficiency levels. In A. Leńko-Szymańska & S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research* (pp. 115–136). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.104.05paq

van Vuuren, S. (2017). *Traces of transfer? Pragmatic development in the use of initial adverbials in the interlanguage of advanced Dutch learners of English*. LOT Publications. Available from: https://repository.ubn.ru.nl/handle/2066/168734

Vyatkina, N., & Cunningham, D. J. (2015). Learner corpora and pragmatics. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 281–305). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9781139649414.013

Further reading

Brezina, V., & Flowerdew, L. (Eds.). (2018). *Learner corpus research: New perspectives and applications*. London: Bloomsbury.

Díaz-Negrillo, A., Ballier, N., & Thompson, P. (Eds.). (2013). *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

Contributor bios

Tove Larsson is Assistant Professor of Applied Linguistics at Northern Arizona University. She specializes in corpus linguistics, learner corpus research, L2 writing, grammatical complexity, and research methods. Her work appears in journals such as the *International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, and *System*. She serves on several editorial boards (e.g., *Journal of English for Academic Purposes, Journal of Second Language Writing*) and is the incoming co-Editor-in-chief for the *International Journal of Learner Corpus Research*.

Gaëtanelle Gilquin is Full Professor of English Language and Linguistics at UCLouvain and co-director of the Centre for English Corpus Linguistics. Her research interests include corpus linguistics, learner corpus research, English varieties, and writing processes. She is one of the editors of *The Cambridge Handbook of Learner Corpus Research* and the coordinator of several learner corpus projects, including the *Louvain International Database of Spoken English Interlanguage*.