Gilquin, G. & S. Granger. 2026. 'L2 English learner varieties'. In R. Reppen, L. Goulart & D. Biber (eds) *The Cambridge Handbook of English Corpus Linguistics. Second Edition* (pp. 509-528). Cambridge: Cambridge University Press.

Chapter 25: L2 English learner varieties

Gaëtanelle Gilquin and Sylviane Granger

Centre for English Corpus Linguistics

Université catholique de Louvain

1. The state of the art of learner corpus research

In the late 1980s, corpus linguists began to show an interest in learner language and started collecting learner corpora, that is, principled electronic collections of writing and/or speech produced by foreign/second language (L2) learners. This research strand, commonly referred to as learner corpus research (LCR), serves two related but distinct objectives: to gain a better understanding of language learning and to design more efficient pedagogical tools and methods tailored to learners' attested needs. Its focus is on performance, i.e., the use of language by learners in actual production. To investigate learner performance, it applies methods mainly based on the tools and techniques of corpus linguistics, which, thanks to the degree of automation they involve, allow for the study of potentially large samples representative of specific learner populations.

1.1. Learner corpus data

There are many different types of performance data and only some of them qualify as learner corpus data. Unlike the more experimental data types often used in second language acquisition

(SLA) studies, where learners are forced to produce a particular form (as in fill-in-the-blanks exercises or read-aloud tasks), the essence of learner corpus data is that learners should be able to use their own wording. In principle, like any other corpora, learner corpora need to be authentic, i.e., "gathered from the genuine communications of people going about their normal business" (Sinclair, 1996). However, learners, especially those who learn the target language in a country where that language is not used for intranational communication, rarely – if ever – use the target language in their normal everyday activities. The criterion of authenticity therefore needs to be relaxed in the case of learner corpora and different degrees of control should be recognized (see also Tracy-Ventura & Myles, 2015). For example, relatively little control is exerted in the case of essay writing or informal interviews, whereas picture descriptions or translations place more constraints on language production (but still allow learners to choose how to express a certain message).

Learner corpora can be of different types – general or specific, written, spoken or multimodal, cross-sectional or longitudinal – and they can include data produced by learners with different profiles: different mother tongues (L1), proficiency levels, etc. The 'Learner Corpora around the World' list, maintained by the Centre for English Corpus Linguistics (2024), reveals that some types of learner corpora are more common than others (see also Paquot & Plonsky, 2017).

Corpora of learner English are more frequent than corpora of any other target language, which is to be related to the fact that the first learner corpora represented English and that, accordingly, LCR initially dealt only with English as a foreign language. Although the focus of this chapter is on English learner corpora, it should be pointed out that LCR has now opened up to other languages besides English.

In addition, written learner corpora, such as the *International Corpus of Learner English* (ICLE; Granger, Dupont, Meunier, Naets, & Paquot, 2020) or the *Corpus and Repository of*

Writing (CROW; Staples & Dilger, 2018-), by far outnumber spoken learner corpora. This imbalance results from the difficulty of collecting and transcribing oral data produced by learners. While several spoken learner corpora such as the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin, De Cock, & Granger, 2010) and the *Trinity Lancaster Corpus* (Gablasova, Brezina, & McEnery, 2019) have started to redress the balance, it would be advisable to have access to high-quality audio files, in addition to transcripts, to be able to investigate aspects like learner pronunciation and prosody.

Another feature of learner corpora is that many are made up of data produced by advanced learners. This initially stemmed from a wish to counterbalance the general neglect of more advanced proficiency levels in L2 studies and teaching resources. This was also the result of convenience sampling, as most learner corpus compilers were university teachers. Recently, however, less advanced learners have been targeted too (e.g., Hasund & Hasselgård, 2022).

Yet another imbalance is between pseudo-longitudinal corpora, which contain data produced by different groups of learners representing distinct proficiency levels, and longitudinal corpora, which contain data produced by individuals as they progress across proficiency levels. Because the latter are more difficult to collect, since they involve following the same learners over a period of time, they are less widespread. While pseudo-longitudinal corpora cannot give access to individual patterns of development, as longitudinal corpora do, they can nevertheless help characterize learner groups at different proficiency levels and thus provide a complementary view of L2 development (Bestgen & Granger, 2018).

1.2. Linguistic phenomena

LCR has tackled aspects of learner language that tended to be neglected in earlier SLA studies, which traditionally prioritized morphology and syntax (Ortega, 2009: 110). LCR, on the other

hand, is characterized by a strong focus on lexis, lexico-grammar and a range of discourse phenomena. This has been made possible by corpus software tools such as *AntConc* (Anthony, 2023), *WordSmith Tools* (Scott, 2024) or *LancsBox* (Brezina, Weill-Tessier, & McEnery, 2020) (see Chapter 3). Studies using these tools have uncovered a facet of learner language which had hitherto been largely left untouched, i.e., the learner phrasicon, which Granger (2021) defines as learners' stock of multiword units. The units that have been at the forefront of LCR are collocations (Paquot, Gablasova, Brezina, & Naets, 2022) and lexical bundles (Xia, Ai, & Pae, 2022), but other types of multiword units have been investigated too, such as collostructions (Wu & Wang, 2022) and phrase-frames (Tan & Römer, 2022). Discourse features are also strongly in evidence in learner corpus studies, in particular the use of connectors (Gao, 2016), discourse markers (Gilquin, 2016), stance markers (Pyykönen, 2023) and involvement features (Hasund & Hasselgård, 2022). Grammatical features used to be under-researched in LCR but are clearly gaining momentum. Some studies deal with aspects of grammar that can be studied on the basis of raw data, either because they involve one specific linguistic item, such as *what*-clefts (Callies, 2009), or a closed class, such as modal verbs (Li, 2022). Others rely on part-of-speech-tagged data (Lim, Mark, Pérez-Paredes, & O'Keeffe, 2024) or, less frequently, parsed data (Cao, 2023).

1.3. General research orientations

Three main research orientations can be identified in LCR: descriptive, predictive and pedagogical.

LCR studies tend to be descriptive. Many studies aim to describe the distinctive use of a particular linguistic feature by learners in comparison with native/expert writers or speakers. Lee, Bychkovska, and Maxwell (2019), for example, compare informal features in L1 and L2

undergraduate student writing. Other studies compare learners at different proficiency levels (Marin Cervantes & Gablasova, 2017) or across time (Bestgen & Granger, 2018).

This type of research orientation has often been criticized as being overly descriptive but, as rightly observed by Rankin (2017), a thorough description is a prerequisite for both theoretical and applied studies of L2 performance. That said, a growing number of studies ground description in SLA theory so as to gain a rigorous interpretation of learner-corpus-based findings and/or provide a more solid data base for SLA theory. For instance, Akbaş and Dinçer (2021) rely on learner corpus data to revisit the order of acquisition of morphemes established in SLA. Such studies can help bridge the gap between SLA and LCR, two fields which are still quite distant from each other, although recent research shows signs of a rapprochement (Le Bruyn & Paquot, 2021).

Predictive studies are becoming increasingly popular in LCR. These studies investigate linguistic features in learner corpora to predict some performance outcome, such as learners' proficiency level. Kim and Ro (2023), for example, use frequency-based and construction-focused indices to assess their predictive validity in relation to English as a foreign language (EFL) learners' speaking proficiency, while Gaillat et al. (2021) identify which features best predict the writing proficiency levels of EFL learners. Many predictive studies use automated linguistic indices (Abdi Tabari, Johnson & Gao, 2025).

Another important orientation of LCR resides in its strong links with teaching, which were in evidence from the start. Several studies demonstrate how learner corpora can be used to develop pedagogical resources in areas as diverse as the design of reference tools (dictionaries and grammars) and instructional materials (textbooks and computer-assisted language learning (CALL)), language testing and assessment, and classroom practice. The insertion of error notes in monolingual learners' dictionaries was the very first application of learner corpus research (Gillard & Gadsby, 1998). More recently, the Louvain Dictionary of

English for Academic Purposes (Granger & Paquot, 2015) adopted this approach in the form of a web-based writing aid based on both native and learner corpora that illustrates the correct use of academic words and warns against frequent errors. As regards grammar books and instructional materials, several learner-corpus-based resources include warnings to raise learners' awareness of frequent errors and/or error detection and correction exercises (Bunting, Diniz, & Reppen, 2013). Learner corpus data can also contribute significantly to language testing and assessment, for example by refining the descriptors of the Common European Framework of Reference for Languages (CEFR) (Díez-Bedmar, 2018; Thewissen, 2013), informing test development (Gyllstad & Snoder, 2021) and improving automatic assessment systems (Higgins, Ramineni, & Zechner, 2015). Learner-corpus-based classroom practice is still quite rare overall but is slowly becoming more common (Rankin & Schiftner, 2011; Staples, Conrad, Dang, & Wang, 2024).

1.4. Representative studies

1.4.1. Gilquin and Granger (2021)

Gilquin and Granger (2021) study the passive, seen as a lexico-grammatical structure rather than a purely grammatical one, as used by several L1 populations in ICLE and a control group of native students in the *Louvain Corpus of Native English Essays* (LOCNESS). While the frequency of the passive is considered, revealing that several learner populations use the passive significantly less often than native writers, the focus is on lexico-grammatical aspects, including the lexical verbs used in the passive. This approach not only confirms that certain verbs are more likely to be passivized than others, but also shows that such preferences may differ between native and non-native writers. For example, the passive ratio of the verb *describe* is statistically higher in ICLE than in LOCNESS, whereas it is the other way around for the verb

see. It also appears that certain phraseological patterns are more distinctive of one or several L1 populations, and hence possibly transfer-related, e.g., *as far as x BE concerned* for the French-speaking learners. In the study, a distinction is further drawn between ICLE subcorpora that correspond to typical EFL populations and others that are closer to English as a second language (ESL) populations, but the results do not point to a clear separation between the two in terms of the passive use.

1.4.2. Thewissen (2013)

This study makes use of data from ICLE to trace second language developmental trajectories in terms of accuracy. It is based on data sampled at the same point in time from learners at different proficiency levels (pseudo-longitudinal data). The data were submitted to two independent processes: they were rated by testing experts according to the CEFR and fully annotated for errors on the basis of the Louvain error-tagging system (Dagneaux, Denness & Granger, 1998). Errors were counted using potential occasion analysis, which quantifies errors in relation to the number of times they could be produced rather than in relation to the total number of words in the corpus. The study investigates 45 error types across four proficiency levels, from lower intermediate to near-native. The results show that the developmental patterns vary in function of the error category. While many error types (e.g., spelling errors or verb-dependent preposition errors) are characterized by progress followed by stabilization, others (e.g., noun number errors) show steady linear development and others still (e.g., verb tense errors) show no sign of development. The findings generally bring support to SLA claims about the nonlinear nature of language development. They also have major pedagogical implications, in particular for language testing research.

1.4.3. Götz (2019)

Götz (2019) relies on learner interviews from LINDSEI and a multifactorial regression analysis to study the frequency of filled pauses (FPs; e.g., *eh*, *erm*) in the speech of learners from different L1 backgrounds. She first investigates the possible effect of L1, as is common in LCR, and finds out that Greek, Italian, French and Polish learners tend to use FPs more often than expected. She then focuses on variables included in the LINDSEI metadata, disregarding L1 this time, and shows that several of them are significant predictors. These include the number of months spent in an English-speaking country (the longer the stay, the lower the frequency of FPs) and the interviewers' gender (there were more FPs in the learner's speech when the interviewer was female). Some of these variables could potentially account for differences initially attributed to L1. Thus, the high frequency of FPs in the Greek subcorpus could be explained by the fact that it includes no learners who had stayed abroad and that all interviewers were female. Götz's study highlights the importance of going beyond the L1 variable and reveals that other variables, including learning context and sociolinguistic ones, may be equally important.

1.4.4. Cotos (2014)

The objective of the study is to explore the potential of a local learner corpus, that is, a corpus representing the language production of one's own students, to enhance data-driven learning (DDL) pedagogy. The first stage of the study involved a comparison of the use of linking adverbials by learners and native speakers. Based on the results, two types of DDL activities were designed, each submitted to a different group of students. One group completed a set of activities involving native speaker corpus data (NSC) only, the other used materials based on both native and learner data (LDD). To identify the respective impact of the two types of activities, three sets of data were collected: pre-experiment data (argumentative essays), immediate and delayed post-test data. The results show that the impact of the DDL activities

was positive for both groups but was particularly noticeable for the LDD group, who displayed a more marked increase in the frequency of linking adverbials, a more balanced distribution of semantic categories and fewer erroneous uses. Responses to an open-ended post-experiment questionnaire revealed that both groups perceived the DDL activities positively, but the LDD group found them particularly helpful and motivating because they involved their own writing. Altogether, the study shows that combining native and (local) learner corpus data is particularly appealing and conducive to learning and that it should be promoted in future CALL pedagogy.

1.5. Methodological issues for learner corpus research

While the methods generally applied in corpus linguistics can be applied in LCR too, there are a number of methods that have been specifically designed to deal with learner corpus data. We will discuss these, and also say a few words about multimethod approaches.

1.5.1. Contrastive interlanguage analysis

Many learner corpus studies rely (explicitly or implicitly) on contrastive interlanguage analysis (CIA) (see Granger, 1996). This method involves two types of comparison: one between a learner corpus and a reference corpus, and another one between different learner (sub)corpora.

The first type of comparison makes it possible to highlight features that are distinctive for learner language. This includes errors (which were the focus of pre-corpus interlanguage studies), but also cases of under- or overuse, i.e., the use of significantly fewer or more instances of a particular item as compared to the reference corpus. Particularly important in this type of comparison is the choice of the reference corpus, as different reference corpora (e.g., British vs. American English or novice vs. expert writing) may yield different results (see Gilquin, 2022).

The revised version of the method, CIA^2 (Granger, 2015a), emphasizes that the reference need not be native (it could be competent L2 use, for instance) and that it can also be manifold.

The second type of comparison involved in CIA is between (varieties of) interlanguages. Most commonly, this comparison is made between several learner populations with different L1s, which helps identify the possible source of certain non-standard features. Features that are limited to one L1 group (or several groups with L1s from the same language family) are likely to be due to transfer from the mother tongue (interlingual features). On the other hand, features that are shared by a large number of L1 groups are more probably linked to inherent difficulties in acquiring the target language (intralingual features). This type of comparison is crucial to avoid attributing to L1 transfer problems that are common to learners from different L1s or that are caused by other factors (e.g., time constraints in Ädel, 2008). The inclusion of more varied factors such as proficiency, L2 exposure or medium in CIA reflects the growing recognition of the variability that characterizes learner language, including individual variation (see, e.g., Ädel, 2015).

1.5.2. Integrated contrastive model

While comparisons between different L1 populations help make more reliable claims about transfer, learner corpus researchers have sought to develop even more rigorous methods to identify transfer, given its central place in SLA. One such method is the integrated contrastive model (ICM) (see Gilquin, 2000/2001; Granger, 1996), which combines CIA with a (corpus-based) contrastive analysis (CA) between the target language and the learner's L1. CA uncovers the characteristics of the learner's L1 and the differences it presents with the target language. On this basis, one can predict possible occurrences of transfer or provide plausible explanations for certain characteristics of the interlanguage. In Gilquin (2025), collostructional analysis is used to compare (i) non-finite verbs in MAKE causative constructions to non-finite verbs in

equivalent constructions with the lemma FAIRE in French (CA) and (ii) non-finite verbs in MAKE causative constructions produced by learners and native speakers (CIA). This twofold analysis points to certain verbs, such as verbs expressing a change of state or location, as being potentially overused by French-speaking learners because of transfer from their L1.

1.5.3. Computer-aided error analysis

Computer-aided error analysis (CEA) (Dagneaux et al., 1998) relies on error-tagged corpus data, i.e., data in which learners' errors have been identified, tagged and, usually, corrected (Díez-Bedmar, 2021). This method entails a number of difficulties at each of its stages. The stage of identification is very time-consuming because it is normally carried out manually (although there have been some attempts at automatic identification of errors, e.g., Yang, Chen, Liu, & Liu, 2020). Different correctors may have different opinions on what counts as an error and it is therefore advisable to have more than one corrector and high inter-rater reliability. Each identified error is then tagged, using a system of error tags which should be very thoroughly documented so as to ensure consistency in the assignment of the tags (see, for example, Granger, Swallow, & Thewissen's (2022) detailed error tagging manual). The correction stage is problematic too, as it may not always be easy to reconstruct what the learner meant and several corrections may be equally plausible. In this respect, a multi-level error annotation system (cf. Reznicek, Lüdeling, & Hirschmann, 2013) may be useful as it allows for the multiple correction of errors. Despite these difficulties, computer-aided error analysis represents a major improvement over traditional error analysis for several reasons, among which the following two. First, it investigates errors in context – alongside correct uses – rather than in isolation. Second, it embraces the full spectrum of errors, from the most basic word-based ones like spelling and morphology errors to those that extend over longer stretches of text like cohesion and style errors.

1.5.4. MuPDAR

In comparison with the early days of LCR, today's LCR is often characterized by the use of advanced statistics, including multifactorial approaches. In this context, MuPDAR (Multifactorial Prediction and Deviation Analysis with Regressions) is worth mentioning as it was developed specifically for LCR (see Gries & Deshors, 2014). It involves the coding of linguistic features thought to distinguish between alternating words or constructions (e.g., *may* and *can* in Gries & Deshors, 2014). On the basis of how these features behave in native corpus data, predictions are made about what choice a native speaker would make in a specific context, and this choice is then compared with the learner's actual choice, to highlight nativelike and non-nativelike choices. Gries and Deshors (2014), for example, demonstrate that French-speaking learners tend to disprefer *may* in negative clauses and that this is one area in which they differ from native speakers. While MuPDAR provides very detailed information about learners' linguistic choices, it should be pointed out that it is only suited for certain phenomena, in particular linguistic alternations (e.g., the genitive alternation), which limits its applicability.

1.5.5. Multimethod approaches

Learner corpora do not provide information on all there is to know about learner language. They are therefore sometimes combined with other sources of information, and more particularly experimental data, to bring a different (but complementary) perspective on learner language. Such a combination comes with several advantages (see Gilquin, 2021), including the possibility of approaching both performance (i.e., what learners produce) and competence (i.e., what learners know). In Möller (2017), the passive is approached through a combination of learner corpus data and experimental data, in the form of a task involving the passive transformation of active sentences. Differences between two groups of learners that are compared in the study appear to be more marked in the experimental data than in the corpus

data. This could be explained by the fact that learner corpus data allow learners to choose their own wording and hence to deliberately avoid the passive (or other difficult aspects of the language), which is not possible in more constrained experimental data.

1.6. Advances and challenges

The advances in data collection, methodology, linguistic analysis and applications are impressive, but each of these areas still has a number of challenges to overcome before the field can truly come into its own.

1.6.1. Data

One of the main contributions of LCR, if not *the* main contribution, is the amount and variety of learner data that it has made available to the community of researchers interested in L2 acquisition for theoretical or applied purposes. Yet, certain types of data are still too rare. This includes spoken corpora accompanied by (high-quality) sound files as well as multimodal corpora. Perhaps speech-to-text technology will facilitate the transcription process in the future and encourage the compilation of more spoken learner corpora, but attempts at automatically transcribing learner English have so far been disappointing (in Mariko & Kondo (2019), over 50% of learner speech automatically transcribed by IBM Watson had to be corrected manually). Longitudinal corpora are also comparatively uncommon. Although they are difficult to collect and may therefore remain relatively small, they have a real contribution to make to the study of L2 development, since they can bring to light learners' individual trajectories across time (Castello, 2016). LCR would also benefit from the compilation of "multidimensional learner corpora (or perhaps databases)" (Gilquin, 2015: 29) representing different varieties, registers, degrees of control, etc., preferably produced by the same learners, so as to allow for reliable

comparisons. For all these learner corpus types, annotation should be provided whenever possible, in various forms, with the hope that parsing, in particular, can become more accurate and widespread.

In addition to the textual data, metadata should also be considered. The metadata of learner corpora have always been rich and varied, and some of the variables that were missing from the first learner corpora (such as proficiency levels) have now been incorporated into many corpora. However, several other factors, such as aptitude and motivation or degree and type of exposure to the target language, some of which can only be measured indirectly through tests and questionnaires (Larsson, Plonsky, & Hancock, 2022), merit greater attention. Standardization of the metadata should also be aimed for, to ensure comparability across studies (Paquot, König, Stemle, & Frey, 2024).

1.6.2. Methodology

Methods specifically developed for LCR (see Section 1.5) have been used successfully to analyze a wide range of linguistic phenomena and have contributed to some important debates in language acquisition. CIA studies have raised the issue of the norm (native vs. non-native; novice vs. expert), while CEA has introduced a higher degree of standardization at each level of the error analysis process (error identification, annotation, counting methods, interpretation). The ICM has paved the way for more rigorous investigations of transfer, and MuPDAR has opened the door to multifactorial analyses taking various linguistic features into account. The use of adequate statistical measures has also made it possible to consider individual variation in addition to group tendencies.

Over the years, LCR has become more quantitatively oriented and has come to rely increasingly on advanced statistics. This has provided many valuable insights into learner language that could not have been obtained otherwise. However, this quantitative shift has often

been to the detriment of more qualitative designs, and there has been a tendency for linguistic interpretations to be "heavily deprioritised or completely missing" (Larsson, Egbert, & Biber, 2022: 139). One of the challenges in LCR will therefore be to achieve a balance between quantitative and qualitative approaches, and to reinstate the idea, already formulated by Biber, Conrad, and Reppen (1998), that qualitative methods are worthwhile, especially in combination with quantitative methods.

In the future, we may expect that technological developments in natural language processing (NLP) and artificial intelligence (AI) will continue improving the automation of certain methodological aspects of LCR, for increasingly large amounts of data. However, it will remain important to check the quality of such automated output through manual work, and also to ensure that the 'big data' approach does not come at the expense of a differentiated approach that takes into account the variability of learner language according to learners' individual profiles.

1.6.3. Linguistic analysis

While morphosyntax was the prime focus of SLA studies, LCR has triggered a shift of focus to lexis and discourse, a development that has proved to be particularly fruitful for the analysis of advanced interlanguage. Early LCR studies focused primarily on lexically based phenomena, but the increased use of morpho-syntactically annotated learner corpora has enabled the investigation of a wider set of phenomena, such as tense and aspect (Fuchs & Werner, 2018) and noun phrase/verb phrase complexity (Larsson & Kaatari, 2020). However, some aspects of language are still underexplored. Pragmatic features, for example, have been largely neglected, one reason being that the data traditionally collected in LCR do not lend themselves to the analysis of some essential aspects of pragmatics, in particular speech acts (thanks, requests, apologies, advice, etc.). Collecting more diversified learner corpus data (e.g., emails, posts on

social media, discussion forums, study groups) should contribute to filling this gap in the future. Phonetic aspects like pronunciation and prosody are also largely uncharted due to the fact that spoken language is mainly investigated on the basis of transcripts rather than sound files. Studies relying on the acoustic signal are still quite rare (Ballier & Martin, 2015). The influence of the full range of factors that affect learners' behavior should also be investigated, in particular the impact of proficiency, task and topic (Ionin & Díez-Bedmar, 2021; Kim & Lu, 2024).

1.6.4. Applications

Pedagogical applications of LCR are promising but have by no means reached their full potential. Progress is slow and some areas lag seriously behind. While most LCR studies end with a brief section on pedagogical implications, pedagogical applications are still quite infrequent. Some areas fare better than others, however. One area where progress has been particularly notable is that of correction tools and writing aids. The increased use of learner corpus data has made it possible to develop more efficient algorithms and one can expect rapid developments thanks to the use of AI-generated techniques. There is still a large margin of progress, though, as even with the use of AI, results are still poor for text-level errors, i.e., errors that cut across sentence boundaries, such as those pertaining to coreference, cohesion or tense (Fang et al., 2023).

Sadly, progress has been particularly modest in the area where advances had been most ardently expected by early learner corpus researchers, that of textbooks and pedagogical grammars. There are quite a few corpus-informed pedagogical resources but very few are learner-corpus-informed. Exceptions include Bunting, Diniz, and Reppen (2013) as well as McCarthy and O'Dell (2017). The most common approach adopted by these resources is to use learner corpus data to confirm intuition-based problem areas, referred to by Granger (2015b: 488) as the "weak approach". The "strong approach" (Granger, 2015b: 488), which truly takes

the learner corpus as the starting point for the analysis, is seldom used. O'Donnell (2012) is a notable exception to this trend. He shows that selecting and ordering grammatical items according to the needs of a specific L1 population can lead to a major overhaul of the grammar syllabus.

Classroom practice is another area where progress has been quite slow, although a recent survey (Götz & Granger, 2024) reveals that data-driven learning involving learner corpus data, often used conjointly with native corpus data, is positively evaluated by both teachers and learners, especially when learners get to work on their own data.

One of the reasons why LCR findings tend not to be taken up by materials designers and teachers is that they are scattered. There is therefore a pressing need for a synthesis of current findings which could boost the pedagogical outcomes of LCR.

2. Empirical study: Two-word discourse markers in native and non-native speech

Some of the aspects discussed above will now be illustrated by means of a study of two-word discourse markers (DMs) in native and non-native speech. Through this study, we were interested in finding out how EFL learners from different L1 backgrounds use DMs, and how their use compares to that of native speakers (quantitatively and qualitatively), but we also wanted to demonstrate that when doing learner corpus research one should consider individual data in addition to pooled data.

The role of DMs in efficient communication has often been underlined, both in L1 (e.g., Erman, 1987) and L2 speech (e.g., Hasselgren, 2002), and the problems EFL learners encounter when using (or not using) them have been brought to light repeatedly, often on the basis of corpora. However, studies of DMs in LCR tend to be limited in two ways: first they are usually

restricted to one learner population (e.g., German learners in Müller, 2005; Swedish learners in Aijmer, 2011) and second they adopt a global approach that does not take account of individual variation (Götz, 2013 is a major exception to this). Our study departs from these in that it examines data produced by learners from eleven L1s and considers them both as groups and as individual learners. The investigation of several learner populations (rather than just one) makes it possible to determine the influence of the L1 on the use of DMs. As for the analysis at the level of individual speakers, it follows a trend in corpus linguistics that recognizes that "corpora are inherently variable internally" (Gries, 2006: 110) and that comparing "native and non-native corpora as wholes (…) runs the risk of disguising differences between individual texts, and may therefore potentially produce misleading results" (Durrant & Schmitt, 2009: 168).

Two corpora are at the basis of our study, namely LINDSEI, which is made up of informal interviews with higher intermediate to advanced EFL learners (Gilquin et al., 2010), and the *Louvain Corpus of Native English Conversation* (LOCNEC), which is an exact replica of LINDSEI but with native speakers of English (De Cock, 2004). In total, twelve L1 groups are represented: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish (each corresponding to a subcorpus of LINDSEI consisting of some 50 interviews), as well as British English (corresponding to the 50 interviews of LOCNEC). We applied a corpus-driven method to select the DMs. After extracting all bigrams from LINDSEI and LOCNEC, we selected those bigrams occurring with a minimum frequency of 200 occurrences in at least one of the L1 subcorpora. We then identified the DMs among these, for a total of six different DMs, viz. *and so, and then, I mean, in fact, sort of* and *you know*. Finally, we manually excluded the concordance lines where the bigram did not behave as a syntactically non-obligatory string, keeping a sentence like *You can't start doing things like that* ***you know*** but excluding one like ***you know*** *everything about them*. The data were analyzed by means of a contrastive interlanguage analysis, of which we exploited the two types of comparison: a

comparison between learners and native speakers and a within-learner comparison (comparison between different learner populations and different learners).

Considering the corpora as aggregates reveals a general underuse of DMs among learners: *you know*, *sort of*, *I mean* and *and then* are significantly more frequent in LOCNEC than in LINDSEI, while the frequency of *and so* is not significantly different between the two corpora; *in fact* is the only exception, being significantly overused by the learners (see Table 1).

	LOCNEC	LINDSEI
you know	514.9	266.8
sort of	487.2	56.1
I mean	372.7	159.2
and then	291.3	184.3
and so	65.9	79.4
in fact	5.2	55.3

Table 1. Relative frequency per 100,000 words of DMs in LOCNEC and LINDSEI

There is, however, a great deal of variation between the subcorpora, depending on the learners' L1. For instance, *sort of* (Figure 1) is rarely used among most learner groups, but the Swedish (SW) group displays a very high frequency which brings it close to the level of the English native speakers (EN). On the other hand, a DM like *in fact* (Figure 2) shows three main peaks of frequency for French (FR), Bulgarian (BU) and Italian (IT) learners, whereas all the other populations (including the native speakers) display a rather low frequency.

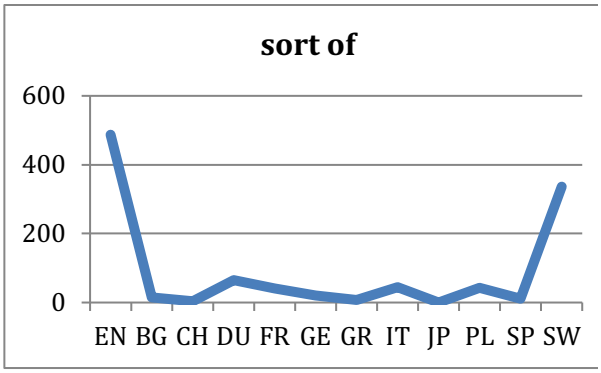

Figure 1. Relative frequency per 100,000 words of *sort of* in the different subcorpora

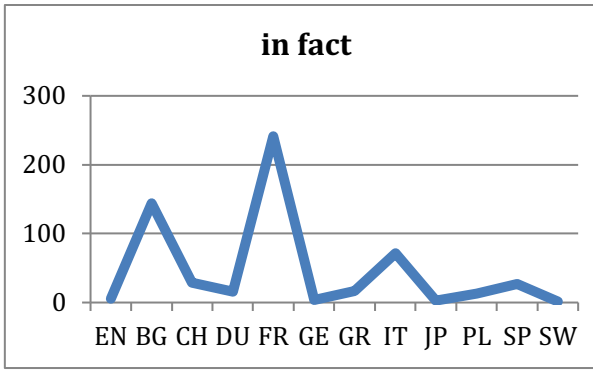

Figure 2. Relative frequency per 100,000 words of *in fact* in the different subcorpora

An ANOVA test reveals the presence of interesting clusters of L1 populations for each of the six DMs, which demonstrates the relevance of the L1 variable. It should be emphasized, however, that the variation is not necessarily a direct reflection of the L1 influence. The overuse of *in fact* among the French-speaking learners, for example, is very probably a consequence of the frequent use of the equivalent French DM *en fait*, but the high frequency of *sort of* among the Swedish learners may simply be due to these learners' high proficiency in speaking skills.

At the last level of analysis, that of the individual interviews, we also found important variation, in the form of speakers' idiolectal preferences for certain DMs. Interestingly, this variability appears in the native corpus too, where it is actually the most marked of all (sub)corpora. While some native speakers have a strong preference for *sort of*, for instance, others clearly prefer *you know*. The same is true of the learners, who may have strong preferences for certain DMs. The main difference between the two groups, however, is that the learners' preferences tend to be exclusive: they show a preference for one or two DMs to the exclusion of the others, which they do not (or at least hardly ever) use; the native speakers, by contrast, may display certain preferences but usually still use the whole range of DMs (except for *and so* and *in fact*, which are less common in LOCNEC).

It may be tempting, in corpus linguistics in general and in LCR in particular, to limit the analysis to a quantitative approach. While quantitative results provide interesting insights into certain aspects of interlanguage, they may also hide some qualitative trends that are worthy of interest. In the case of two-word DMs, we wanted to go beyond the simple equation between high frequency of DMs and fluency (as established, e.g., in Hasselgren, 2002) to see how native-like learners' use of DMs was. With this aim in mind, we selected two groups of learners whose use of *you know* showed different tendencies, namely the French-speaking learners who use relatively few occurrences of *you know* (228 per 100,000 words) and the Polish learners who use many occurrences of it (715 per 100,000 words), and we compared these two populations with the native group (relative frequency of *you know*: 515 per 100,000 words). By making a distinction between cases where *you know* does not interrupt any structure (e.g., *it's just like a curtain **you know** so you've gotta change it*) and cases where it interrupts phrases or closely-knit structures (e.g., *if we **you know** make some something legal*), we noticed that the French-speaking learners' behavior was closer to the native norm than that of the Polish learners. While the French-speaking learners use *you know* in an interrupted structure in 37% of the cases (to be compared with 35% in LOCNEC), the Polish learners interrupt closely-knit structures in 60% of the cases. In other words, the French-speaking learners underuse *you know* but when they do use it, the breakdown of interrupted vs. non-interrupted structures is very similar to that found among the native speakers. On the other hand, the Polish learners use *you know* very often (more, in fact, than the native speakers), but they tend to use it within strings that should in principle have been uttered as uninterrupted chunks. Especially striking are the interruptions between a copula (usually *be*) and the predicative complement (as in *he stopped being **you know** humorous*) and the interruptions between a preposition and the rest of the prepositional phrase (as in *that was some guy from **you know** the upper classes*). Such uses, rather than

contributing to the fluency of the Polish learners' speech, mark it as particularly disfluent and non-native-like, contrary to what the quantitative analysis would have suggested.

While both the quantitative and qualitative analyses should be expanded to provide a full picture of the use of two-word DMs by EFL learners, this study shows the potential of LCR to bring to light linguistic features typical of certain (groups of) learners. The exploitation of a native corpus, used in combination with the learner corpora, makes it possible to see how the learner data are situated in relation to a certain reference norm (without being limited by it), whereas the inclusion of the L1 variable gives a glimpse of the possible influence of the mother tongue. This 'classic' application of CIA is furthermore refined by also considering the speakers individually, which underlines the idiolectal variation that can be found within corpora. Like many other learner-corpus-based studies, however, this one does not examine the wide range of variables that are available in learner corpora like LINDSEI and that may have an effect on the learners' use of DMs (such as the time spent in an English-speaking country or the knowledge of other languages), nor does it take into account the general context in which the learners acquire English (e.g., input-rich vs. input-poor environment). In addition, since we only worked on the basis of the (transcribed) utterances produced by the interviewees, without looking at the interviewers' turns, we could not adopt a more pragmatic perspective, which would have involved investigating how interaction is created between speakers by means of DMs, nor a sociolinguistic perspective, which could for instance have studied whether the interviewer's profile (e.g., male or female, native or non-native speaker of English) has an impact on the presence of DMs in the learner's language. Finally, this study is mainly descriptive and does not seek to establish links with two fields which are close (and complementary) to LCR, namely SLA and foreign language teaching (FLT). An SLA approach would for example favor a more interpretative perspective and a more pronounced focus on competence than is the case in LCR. An FLT approach would look into the ways in which the

corpus findings can be applied for pedagogical purposes; in the case of our study, this could start with the question of whether (and how) DMs should be taught in the classroom. A combined LCR-SLA-FLT approach would thus result in an integrated descriptive/interpretative/applied perspective which is still largely lacking in the field of interlanguage studies.

References

Abdi Tabari, M., Johnson, M. D., & Gao, J. 2025. Using automated indices of cohesion to explore the growth of cohesive features in L2 writing. *International Review of Applied Linguistics in Language Teaching*, 63(3), 2169–2200.

Ädel, A. 2008. Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, and M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research*, 35–53. Amsterdam: Rodopi.

Ädel, A. 2015. Variability in learner corpora. In S. Granger, G. Gilquin, and F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, 401–422. Cambridge: Cambridge University Press.

Aijmer, K. 2011. *Well I'm not sure I think…* The use of *well* by non-native speakers. *International Journal of Corpus Linguistics*, 16(2), 231–254.

Akbaş, E. & Dinçer, Z. Ö. 2021. Accuracy order in L2 grammatical morphemes: Corpus evidence from different proficiency levels of Turkish learners of English. *Studies in Second Language Learning and Teaching*, 11(4), 607–627.

Anthony, L. 2023. AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Ballier, N. & Martin, P. 2015. Speech annotation of learner corpora. In S. Granger, G. Gilquin, and F. Meunier (Eds.), *The Cambridge handbook of learner corpus research,* 107–134. Cambridge: Cambridge University Press.

Bestgen, Y. & Granger, S. 2018. Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, and L. M. Dillmann (Eds.), *Corpora and lexis*, 277–301. Leiden: Brill.

Biber, D., Conrad, S., & Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Brezina, V., Weill-Tessier, P., & McEnery, A. 2020. #LancsBox v. 5.x. [Computer Software]. Available from http://corpora.lancs.ac.uk/lancsbox.

Bunting, J. D., Diniz, L., & Reppen, R. 2013. *Grammar and beyond 4*. Cambridge: Cambridge University Press.

Callies, M. 2009. "What's even more alarming is..." – A contrastive learner-corpus study of *what*-clefts in advanced German and Polish L2 writing. In M. Wysocka (Ed.), *On language structure, acquisition and teaching. Studies in honour of Janusz Arabski on the occasion of his 70th birthday*, 283–292. Katowice: Wydawnictwo Uniwersytetu Slaskiego.

Cao, A. 2023. A study on syntactic complexity and topic difference influence in second language writing based on dependency treebank. *World Journal of Educational Research*, 10(3), 66–82.

Castello, E. 2016. Towards a longitudinal study of metadiscourse in EFL Academic Writing: Focus on Italian learners' use of *it*-extraposition. In E. Castello, K. Ackerley, and F. Coccetta (Eds.), *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment*, 179–197. Bern: Peter Lang.

Centre for English Corpus Linguistics. 2024. Learner corpora around the world. Louvain-la-Neuve: Université catholique de Louvain. Available at https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Cotos, E. 2014. Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202–224.

Dagneaux, E., Denness, S., & Granger, S. 1998. Computer-aided error analysis. *System*, 26(2), 163–174.

De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series, 2, 225–246.

Díez-Bedmar, M. B. 2018. Fine-tuning descriptors for CEFR B1 level: Insights from learner corpora. *ELT Journal*, 72(2), 199–209.

Díez-Bedmar, M. B. 2021. Error Analysis. In N. Tracey-Ventura and M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora*, 90–104. New York: Routledge.

Durrant, P. & Schmitt, N. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157–177.

Erman, B. 1987. *Pragmatic expressions in English. A study of* you know, you see *and* I mean *in face-to-face conversation.* Stockholm: Almqvist & Wiksell.

Fang, T., Yang, S., Lan, K., Wong, D. R., Hu, J., Chao, L. S., & Zhang, Y. 2023. Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. *ArXiv* abs/2304.01746.

Fuchs, R. & Werner, V. 2018. Tense and aspect in second language acquisition and learner corpus research. Introduction to the special issue. *International Journal of Learner Corpus Research*, 4(2), 143–163.

Gablasova, D., Brezina, V., & McEnery, T. 2019. The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, 5(2), 126–158.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. 2021. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2), 130–146.

Gao, X. 2016. A cross-disciplinary corpus-based study on English and Chinese native speakers' use of linking adverbials in academic writing. *Journal of English for Academic Purposes*, 24, 14–28.

Gillard, P. & Gadsby, A. 1998. Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), *Learner English on computer*, 159–171. London: Addison Wesley Longman.

Gilquin, G. 2000/2001. The integrated contrastive model. Spicing up your data. *Languages in Contrast*, 3(1), 95–123.

Gilquin, G. 2015. From design to collection of learner corpora. In S. Granger, G. Gilquin, and F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, 9–34. Cambridge: Cambridge University Press.

Gilquin, G. 2016. Discourse markers in L2 English: From classroom to naturalistic input. In O. Timofeeva, A.-C. Gardner, A. Honkapohja, and S. Chevalier (Eds.), *New approaches to English linguistics: Building bridges*, 213–249. Amsterdam: John Benjamins.

Gilquin, G. 2021. Combining learner corpora and experimental methods. In N. Tracy-Ventura and M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora*, 133–144. New York: Routledge.

Gilquin, G. 2022. One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, 55(1), 87–99.

Gilquin, G. 2025. Transfer of collostructions: The case of causative constructions. *Corpus Linguistics and Linguistic Theory*, 21.

Gilquin, G., De Cock, S., & Granger, S. 2010. *Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, G. & S. Granger. 2021. The passive and the lexis-grammar interface: An inter-varietal perspective. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora*, 72–98. Bristol: Multilingual Matters.

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, and A. C. Simon (Eds.), *Fluency and disfluency across languages and language varieties*, 177–196. Louvain-la-Neuve: Presses universitaires de Louvain.

Götz, S. & Granger, S. 2024. Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, 10(1), 1–38.

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

Granger, S. 2015a. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24.

Granger, S. 2015b. The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, and F. Meunier (Eds.), *The Cambridge handbook of*

learner corpus research, 485–510. Cambridge: Cambridge University Press.

Granger, S. (Ed.). 2021. *Perspectives on the L2 phrasicon. The view from learner corpora.* Bristol: Multilingual Matters.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. 2020. *The International Corpus of Learner English. Version 3*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. & Paquot, M. 2015. Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), 118–141.

Granger, S., Swallow, H., & Thewissen, J. 2022. The Louvain error tagging manual. Version 2.0. CECL Papers 4. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/cecl-papers.html (accessed 30 May 2024).

Gries, S. Th. 2006. Exploring variability within and between corpora: Some methodological considerations. *Corpora*, 1(2), 109–151.

Gries, S. Th. & Deshors, S. C. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, 9(1), 109–136.

Gyllstad, H. & Snoder, P. 2021. Exploring learner corpus data for language testing and assessment purposes: The case of verb + noun collocations. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora*, 49–71. Bristol: Multilingual Matters.

Hasselgren, A. 2002. Learner corpora and language testing. Smallwords as markers of learner fluency. In S. Granger, J. Hung, and S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, 143–173. Amsterdam: John Benjamins.

Hasund, I. H. & Hasselgård, H. 2022. Writer/reader visibility in young learner writing: A study of the TRAWL corpus of lower secondary school texts. *Journal of Writing Research*, 13(3), 447–472.

Higgins, D., Ramineni, C., & Zechner, K. 2015. Learner corpora and automated scoring. In S. Granger, G. Gilquin, and F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, 587–604. Cambridge: Cambridge University Press.

Ionin, T. & Díez-Bedmar, M. B. 2021. Article use in Russian and Spanish learner writing at CEFR B1 and B2 levels: Effects of proficiency, native language, and specificity. In B. Le Bruyn and M. Paquot (Eds.), *Learner corpus research meets second language acquisition*, 10–38. Cambridge: Cambridge University Press.

Kim, M. & Lu, X. 2024. L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic. *The Modern Language Journal*, 108(1), 270–296.

Kim, H. & Ro, E. 2023. Assessment of sentence sophistication in L2 spoken production: Expansion of verbs and argument structure constructions. *System*, 119, 103175.

Larsson, T., Egbert, J., & Biber, D. 2022. On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora*, 17(1), 137–157.

Larsson, T. & Kaatari, H. 2020. Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, 45, 100850.

Larsson, T., Plonsky, L., & Hancock, G. R. 2022. On learner characteristics and why we should model them as latent variables. *International Journal of Learner Corpus Research*, 8(2), 237–260.

Le Bruyn, B. & Paquot, M. (Eds.). 2021. *Learner corpus research meets second language acquisition*. Cambridge: Cambridge University Press.

Lee, J. J., Bychkovska, T., & Maxwell, J. D. 2019. Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System*, 80, 143–153.

Li, L. X. 2022. Developmental patterns of English modal verbs in the writings of Chinese learners of English: A corpus-based approach. *Cogent Education*, 9(1), 2050457.

Lim, J. D. O., Mark, G., Pérez-Paredes, P., & O'Keeffe, A. 2024. Exploring part of speech (POS) tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1), 31–59.

Mariko, A. & Kondo, Y. 2019. Constructing a longitudinal learner corpus to track L2 spoken English. *Journal of Modern Languages*, 29, 23–44.

Marin Cervantes, I. & Gablasova, D. 2017. Phrasal verbs in spoken L2 English: The effect of L2 proficiency and L1 background. In V. Brezina and L. Flowerdew (Eds.) *Learner corpus research: New perspectives and applications*, 28–46. London: Bloomsbury.

McCarthy, M. & O'Dell, F. 2017. *English vocabulary in use. Advanced. 3rd Edition*. Cambridge: Cambridge University Press.

Möller, V. 2017. *Language acquisition in CLIL and non-CLIL settings: Learner corpus and experimental evidence on passive constructions*. Amsterdam: John Benjamins.

Müller, S. 2005. *Discourse markers in native and non-native English discourse.* Amsterdam: John Benjamins.

O'Donnell, M. 2012. Using learner corpora to redesign university-level EFL grammar education. *Revista Española de Lingüística Aplicada*, Extra 1, 145–160.

Ortega, L. 2009. *Understanding Second Language Acquisition*. London: Routledge.

Paquot, M., Gablasova, D., Brezina, V., & Naets, H. 2022. Phraseological complexity in EFL learners' spoken production across proficiency level. In A. Leńko-Szymańska and S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research*, 115–136. Amsterdam: John Benjamins.

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. 2024. The core metadata schema for learner corpora: Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, 10(2).

Paquot, M. & Plonsky, L. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61–94.

Pyykönen, M. 2023. Exploring patterns of lexical variation in the use of epistemic stance markers in written L2 English across task types and levels of proficiency. A corpus-based study. *International Journal of Learner Corpus Research*, 9(2), 215–247.

Rankin, T. 2017. The distribution of reflexive intensifiers in learner English. *International Journal of Learner Corpus Research*, 3(1), 36–60.

Rankin, T. & Schiftner, B. 2011. Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, 16(3), 412–434.

Reznicek, M., Lüdeling, A., & Hirschmann, H. 2013. Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In A. Díaz-Negrillo, N. Ballier, and P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data*, 101–123. Amsterdam: John Benjamins.

Scott, M. 2024. WordSmith Tools (Version 9) [Computer Software] Stroud: Lexical Analysis Software.

Sinclair, J. 1996. Preliminary recommendations on corpus typology, technical report, EAGLES (Expert Advisory Group on Language Engineering Standards). http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (accessed 30 May 2024).

Staples, S., Conrad, N., Dang, A. & Wang, H. 2024. Building language and genre awareness through learner corpus data in a second language writing course. *International Journal of Learner Corpus Research*, 10(1), 146–182.

Staples, S. & Dilger, B. 2018- . Corpus and repository of writing [Learner corpus articulated with repository]. Available from https://crow.corporaproject.org.

Tan, Y. & Römer, U. 2022. Using phrase-frames to trace the language development of L1 Chinese learners of English. *System*, 108, 102844.

Thewissen, J. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77–101.

Tracy-Ventura, N. & Myles, F. 2015. The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), 58–95.

Wu, T. & Wang, M. 2022. Development of the progressive construction in Chinese EFL learners' written production: From prototypes to marginal members. *Corpus Linguistics and Linguistic Theory*, 18(2), 307–335.

Xia, D., Ai, H., & Pae, H. K. 2022. "Please let me know": Lexical bundles in business emails by business English learners and working professionals. *International Journal of Learner Corpus Research*, 8(1), 1–30.

Yang, C. T-Y., Chen, H. H-J., Liu, C-Y., & Liu, Y-C. 2020. A semi-automatic error retrieval method for uncovering collocation errors from a large learner corpus. *English Teaching & Learning*, 44, 1–9.