Tema

LEARNER CORPORA TO MEET LEARNERS' INDIVIDUAL NEEDS

Les corpus d'apprenants, comme d'autres corpus, ont surtout permis de faire des généralisations sur des populations entières. Ils peuvent cependant être exploités à des fins pédagogiques de manière plus différenciée et inclusive, en montrant comment des apprenant·es avec un profil spécifique utilisent (ou sont susceptibles d'utiliser) la langue cible. Une telle approche peut s'appuyer sur les métadonnées de corpus d'apprenants existants ou sur des données de corpus recueillies parmi ses propres étudiant·es. Les résultats issus de l'analyse de ces corpus peuvent aider à développer du matériel et des activités pédagogiques sur mesure pour répondre aux besoins de groupes d'apprenant·es particuliers ou d'apprenant·es individuel·les, y compris des activités d'apprentissage sur corpus ***(data-driven learning),*** grâce auxquelles les étudiant·es peuvent faire des découvertes sur leur propre utilisation de la langue cible.

Gaëtanelle Gilquin
| UCLouvain

Gaëtanelle Gilquin is Professor of English Language and Linguistics at the University of Louvain, Belgium. She has co-edited the Cambridge handbook of learner corpus research and is the coordinator of several learner corpus projects, including the Louvain International Database of Spoken English Interlanguage.

Learner corpora: generalizing trends

Learner corpora started to be collected in the 1990s, with the aim of providing linguists (including lexicographers) with large electronic databases of authentic language produced by second/foreign language (L2) learners. One of the earliest learner corpora, the International Corpus of Learner English (ICLE), was first published in 2002 and contained some 2.5 million words of L2 English written by learners from 11 different mother tongue (L1) backgrounds (Granger et al., 2002).

Since then, learner corpora have kept growing. The EF-Cambridge Open Language Database (EFCAMDAT), for example, another learner corpus of written English, is currently made up of some 50 million words and represents almost 200 nationalities (based on Shatz, 2020). Spoken learner corpora tend to be smaller, which reflects the time and effort needed to collect and transcribe speech, but the largest spoken learner corpora now come close to 5 million words (4.2 million words for the Trinity Lancaster Corpus (TLC), see Gablasova et al., 2019).

The increasingly large size of learner corpora is usually seen as a welcome development, since large learner corpora are likely to better represent certain learner populations and include more instances of specific linguistic phenomena than small learner corpora. Generalizations made on the basis of large learner corpora therefore tend to be more reliable. This is an important feature for many studies, because learner corpora, similarly to other corpora, have mostly been used to establish what is frequent in language and common to a majority of writers/speakers.

Gilquin et al. (2007), for example, is a guide included in the second edition of the *Macmillan English Dictionary for Advanced Learners*, which is meant to help learners of English produce better academic and professional writing. In this guide, distinctive learner usage is only

The across-the-board approach usually adopted in learner corpus research may not be fully satisfactory to teachers who aim to meet learners' individual needs.

mentioned if it is frequent in the learner corpus used (ICLE) and typical of a majority of the learner groups represented in the corpus (the groups being defined by the learners' L1). Thus, the guide includes a 'Be careful' note about *of course*, because many learners from different L1 groups overuse *of course* in academic writing. By contrast, the overuse of *in fact* is not mentioned in the guide, because it is a feature that mainly characterizes French- and Italian-speaking learners.

Generalizations can be made at more specific levels than that of 'all learners of a target language'. For instance, O'Keeffe & Mark (2017), focusing on grammatical structures used correctly in learner English, adopt several criteria to ensure the widespread use of the structures: they should be frequent in the learner corpus, spread across a range of learners from several L1 families, occur in different registers/tasks, etc. However, a distinction is drawn between learners with different proficiency levels, so that the "grammatical competence statements" (O'Keeffe & Mark, 2017: 457), which show what learners are usually able to do, apply to learners with a given proficiency level, e.g. "Can use the affirmative form of the past perfect simple" (O'Keeffe & Mark, 2017: 476) for learners with a B1 level according to the Common European Framework of Reference for Languages (CEFR).

Such generalizing trends have provided many valuable insights into learner language (see, e.g., Granger et al., 2015) and have led to useful teaching applications. The analysis of the TLC, for example, has helped develop classroom activity worksheets available via the TLC Hub (https://cass.lancs.ac.uk/trinity-lancaster-corpus/). These worksheets highlight strategies – some successful, others less successful – which are often employed by learners in the TLC. Thus, it is shown that successful C1-C2 speakers (along the CEFR scale) use *I don't agree* and *I agree but...* more often than *I disagree* or *I can't agree* to express disagreement (Brezina, 2017). However, such an across-the-board approach may not be fully satisfactory to teachers who aim to meet learners' individual needs.

Differentiated and inclusive instruction with existing learner corpora

Learner language is known to be very heterogeneous, being affected by a large variety of factors such as the learner's L1, knowledge of additional languages, exposure to the target language, but also task, timing, access to reference tools, etc. In a language classroom, learners are therefore likely to use the target language in distinct ways. This is all the more so in mixed classrooms, which have become more common recently, partly as a result of educational policies. Mixed-age classrooms, mixed-ability classrooms, ethnically mixed classrooms, etc. bring together learners from various backgrounds, with distinctive characteristics, diverse aspirations, and so on. If teachers want to offer differentiated and inclusive instruction, as they are often encouraged to do, they need to adapt to the diversity of the classroom and seek to cater for the individual needs of each of their students.

Learner corpora, many of which are publicly available,[1] can help with this, because they can provide information about the language behaviour of learners with specific profiles. If a classroom includes students from a range of L1 backgrounds, different existing (sub)corpora can be exploited that represent each of these L1 backgrounds. Even if teachers are not familiar with these L1s and therefore not aware of the particular challenges that speakers of these L1s may face when trying to learn the target language, they will be able to find out about these in the learner corpora.

[1] See https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html for a list of learner corpora (Centre for English Corpus Linguistics, 2024).

For example, most learners of English find it hard to use high-frequency verbs (such as *make*, *take*, or *give*) appropriately, because these verbs are highly polysemous and in some cases the choice of the verb is largely arbitrary (compare *give a talk* and *make a comment*). However, learners with different L1s tend to produce distinct non-standard combinations with these verbs. Huiping & Yongbing (2014) observe that, in the International Corpus of Crosslinguistic Interlanguage (ICCI), Chinese learners often use *make* with the collocate *party* (e.g. *make a birthday party*). They explain this by the fact that the equivalent of *make* in Chinese, *zuo*, can be used with the meaning of "causing (a birthday party, banquet, etc.) to take place" (Huiping & Yongbing, 2014: 268). Austrian learners, by contrast, produce no such collocations (ibid.). Based on an analysis of *make* in the French and Swedish ICLE subcorpora, Altenberg & Granger (2001: 180) point to combinations that also seem to be caused by transfer from the L1 and hence tend to be specific to these L1 groups, e.g. *make a poll* (instead of the more standard *carry out/conduct a poll*) for French-speaking learners and *make harm* (instead of *do harm*) for Swedish learners.

Learner corpora can help teachers offer differentiated and inclusive instruction by providing them with information about the language behaviour of learners with specific profiles.

Such findings can help provide students with pedagogical activities that address their language specificities (according to their L1 or some other feature), for instance in the form of L1-influenced collocations to be corrected. These targeted activities can stimulate interesting classroom discussions, where students explain how the equivalent word or structure is used in their L1, so that everybody can learn about each other's L1 and become more aware of crosslinguistic variation and also more respectful of differences.

Customizing pedagogical materials and activities according to the diversity of the classroom on the basis of existing learner corpora is made possible by the rich metadata that most learner corpora contain. In the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al., 2010), for example, each learner interview is described in terms of 23 variables, including learners' L1, how long they have been learning English, and how much time they have spent in an English-speaking country.

Each learner corpus comes with its own set of variables, some of which may be particularly relevant in certain teaching contexts. Thus, the Process Corpus of English in Education (PROCEED; Gilquin, 2022) is a learner corpus made up of argumentative essays as well as data showing the process through which these essays were composed (keylog files and screencast videos). Its metadata include information about learners' possible neurodivergence, e.g. whether they were diagnosed with dyslexia or ADHD. Using data from this corpus, it would for example be possible to discover the successful strategies of high-functioning dyslexics writing in L2 English (see Radar & Gilquin, forthcoming) and present these as potential models to dyslexic students who struggle with L2 writing tasks.

While the first compiled learner corpora were mostly made up of written English produced by advanced learners, over the years learner corpora have diversified, representing more varied target languages, L1 backgrounds, proficiency levels, etc. (see Gilquin & Granger, forthcoming). Yet, not all possible student profiles will have their corresponding learner corpora. Less commonly taught languages and less typical learners (including heritage language learners or learners with special educational needs), in particular, are not well represented among existing learner corpora.

In addition, if teachers want to target a very specific profile, they may end up with a small sample, even if they use a large learner corpus to start with (see Callies, 2015: 52). Thus, although the current version of ICLE contains over 5.5 million words, of which almost 500,000 were written by close to 1,000 Chinese-speaking learners, there is only one text produced by a Chinese-speaking learner with Italian as an attested L2, corresponding to 455 words. When existing learner corpora do not provide sufficient or sufficiently relevant data, teachers can turn to data collection among their own students.

Individual tailoring with local learner corpora

Most learner corpora exploited for teaching purposes are "learner corpora for delayed pedagogical use" (Granger, 2009: 21). This means that they are compiled among a certain learner group (usually by academics or publishers) and exploited later among a different learner group (e.g. through the use of a textbook which was written with the help of the learner corpus data). However, teachers can also rely on "learner corpora for immediate pedagogical use" (ibid.), that is, corpora that are collected among the learners who will benefit from their pedagogical exploitation. Such corpora are called "local learner corpora" (Seidlhofer, 2002), because they represent a local learner group, typically the students of the teacher who is in charge of the corpus compilation.

Local learner corpora are usually collected as part of the day-to-day teaching activities. Whenever the students have to complete a writing task, the texts can be added to the corpus. The same is true of spoken tasks, although the time necessary for the transcription of speech may be an obstacle to the compilation of spoken local learner corpora. Provided they are made in a principled way, the teacher's corrections can be integrated into the corpus too, having the function of 'error tagging' and making it possible to retrieve certain error types automatically (e.g. all incorrect uses of the auxiliary *can*) and generate statistics (e.g. learners' progress in spelling over the weeks).

Learner corpora collected from one's own students are usually quite small, but they are truly representative of the students' language production. These corpora can be analyzed to bring to light patterns that are characteristic of the students' L2 usage. These observations can then be turned into pedagogical materials or activities that target the students' specific needs. Rankin & Schiftner (2011), for instance, show how the analysis of a local learner corpus of German-speaking learners of English revealed an overuse and predominantly non-standard use of the marginal preposition *concerning* (e.g. *Alberto showed no real progress concerning grammar*). On the basis of this finding, they prepared targeted exercises, including sentences taken from the corpus in which the students were required to find an alternative to the non-standard uses of *concerning*.

An individual local learner corpus reveals the linguistic features typical of a learner's idiolect and makes it possible to offer tailor-made feedback and instruction to the learner.

Each learner has their unique way of using the L2, which is the result of multiple factors, such as the type and amount of input that they have received, their capacity to remember and reproduce words or structures that they have been exposed to, or their creativity in applying language patterns. Using learner corpus data produced by learners with a similar profile, even from the same classroom, can therefore only provide an approximation of a learner's language system. For many purposes, this approximation will be good enough – and, in any case, better than across-the-board generalizations. However, in some pedagogical contexts, it may be desirable to get to the uniqueness of each learner. This can be done by compiling local learner corpora made up of data produced by individual students.

An individual local learner corpus comprises texts in L2 produced by one and the same learner. The examination of such a corpus reveals the linguistic features typical of the learner's idiolect. This information makes it possible to offer tailor-made feedback and instruction to the learner. Importantly, this approach avoids exposing students to learner language features (including errors) that may not apply to them – one of the main criticisms levelled at the use of learner corpora in pedagogy (see, e.g., Flowerdew, 2001). Thanks to individual local learner corpora, students can also situate themselves in relation to the whole group, find out what aspects of language they already master, and see what progress they have made. Although traditional graded assignments may allow for this too, texts brought together in the form of a corpus can be queried using the tools and techniques of learner corpus research, which can facilitate analysis and give access to more precise and accurate information than would otherwise be available, such as word frequencies, collocations, or keywords (that is, words distinctive of the corpus at hand as compared to a reference corpus).

Using learner corpora in different contexts

Learner corpora can be used by different educational actors and for different functions. They can help language testers develop tests that are suited to learners with special characteristics (e.g. learners with speech disorders) and set fair and achievable standards for them. Teaching materials writers can produce resources that take better account of learners' realities (what they can already do, what they have difficulty with, etc.) and offer contents that are adapted to certain learner groups (e.g. beginners or Spanish-speaking learners). Teachers, provided they have received training in corpus linguistics, can also integrate learner corpora into their teaching routine. Given their students' profiles, they can select the most relevant information and cater for the learners' specific needs by providing them with tailor-made activities. Students can also be given direct access to learner corpus data and be encouraged to make discoveries about learner language themselves, through so-called data-driven learning (see Gilquin & Granger, 2022). With the right type of learner corpus (which could be a corpus of their own language production or a corpus of language produced by learners with similar characteristics), they can embrace their own individuality and become responsible for their differentiated learning.

The way learner corpora can be exploited for pedagogical purposes will vary according to the context. With beginners, for example, highly rated texts from learner corpora can be a source of relatively simple sentences and an achievable target to be used as a model. With more advanced students, learner corpora compared with corpora of expert language can help raise awareness of fossilized errors (Nesselhauf, 2004). In data-driven learning, the teacher's role may be more or less prominent depending on the students' degree of autonomy and their observation and abstraction skills.

Despite the value of (local) learner corpora in highlighting what exactly each learner needs and hence promoting pedagogical differentiation and inclusion, one should not underestimate the difficulty of the endeavour. Offering individualized teaching to one's students clearly requires more time and effort than one-size-fits-all teaching, and the more one seeks to take the diversity of the classroom into account (that is, the closer one aims to get to the uniqueness of learners' language systems), the more work will be involved. Having to carry out extensive analyses on existing learner corpora or compiling learner corpora for this purpose may put too heavy a burden on the shoulders of teachers who are often already overburdened. However, even modest incursions into the realm of learner corpora and pooling of learner-corpus-based resources among teachers whose students have similar profiles should, when combined with other differentiated teaching practices, contribute to more inclusive classrooms, in which all learners feel respected and valued for their differences.

Through data-driven learning, students can embrace their own individuality and become responsible for their differentiated learning.

References

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173-195.

Brezina, V. (2017). Pragmatic functions: Expressing disagreement. Learning from Assessment: CEFR level C1 – Activity worksheet 1. Available at: https://www.trinitycollege.com/resource/?id=7979.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35-55). Cambridge University Press.

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.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 363-379). John Benjamins.

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.

Gilquin, G. (2022). The *Process Corpus of English in Education*: Going beyond the written text. *Research in Corpus Linguistics*, *10*(1), 31-44. Available at: http://ricl.aelinco.es/first-view/174-Article%20Text-1066-1-10-20210407.pdf.

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., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics. Second edition* (pp. 430-442). Routledge.

Gilquin, G., & Granger, S. (Forthcoming). L2 English learner varieties. In R. Reppen, L. Goulart, & D. Biber (Eds.), *The Cambridge handbook of English corpus linguistics. Second edition*. Cambridge University Press.

Gilquin, G., Granger, S., & Paquot, M. (2007). Improve your writing skills (Writing sections). In M. Rundell (Editor in chief), *Macmillan English dictionary for advanced learners. Second edition* (pp. IW1-IW29). Macmillan Education.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). John Benjamins.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International Corpus of Learner English. Handbook and CD-ROM.* Presses universitaires de Louvain.

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

Huiping, Z., & Yongbing, L. (2014). A corpus study of most frequently used English verbs by Chinese beginner learners from a conceptual transfer perspective. *International Journal of Corpus Linguistics*, *19*(2), 252-279.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). John Benjamins.

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457-489.

Radar, L., & Gilquin, G. (Forthcoming). A corpus study of dyslexic university students' L2 writing processes. *Revue de linguistique et de didactique des langues*.

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

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213-234). John Benjamins.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 220-236.