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

Written learner corpora to inform teaching

Gaëtanelle Gilquin

ORCID: 0000-0001-7747-6443

1. Introduction

This chapter deals with how written learner corpora can be used – and have been used – to inform teaching, and in particular English language teaching (ELT). Learner corpora are corpora made up of the production of second or foreign language learners in the target language. They can consist of written, spoken, or multimodal data. However, for reasons that are partly related to the ease of compilation, written learner corpora have the lion’s share. At the time of writing this chapter, the list “Learner Corpora around the World”, maintained by the Centre for English Corpus Linguistics (2020), includes 64% of learner corpora exclusively made up of written data and 12% of learner corpora including both written and spoken data. When learner corpora are used to inform teaching, they are therefore more likely to contain written learner language. In fact, many of the pedagogical resources described in this chapter present themselves as being based on learner corpora, with no specification of the register, but they turn out to rely on written learner corpus data only. Very often, they aim to inform general English language teaching rather than specific aspects of writing.

While almost any aspect of learner writing can be investigated on the basis of a written learner corpus and can potentially be of interest for teaching purposes, it is the difficulties which learners experience with the target language that tend to be seen as the most pedagogically relevant. Among these, errors are particularly prominent. They can be retrieved from raw learner corpus data (provided they are lexically-based, e.g. *informations* or *dependent of*), ideally with the help of the teacher's or the researcher's experience of what learners usually get wrong. However, the retrieval of learners' typical errors is facilitated by prior error-tagging of the learner corpus, that is, the annotation of each error found in the corpus, often accompanied by a description of the error type and a possible correction (Lüdeling & Hirschmann, 2015). This allows for the automatic extraction of all errors, or all instances of a specific type of error (e.g. errors with modal verbs or errors with *can*).

Another way of finding out about learners' difficulties with the target language is by comparing the frequency of linguistic items in the learner corpus with that in a comparable native corpus. This makes it possible to discover cases of underuse or overuse, i.e. items that are used significantly less often or significantly more often, respectively, by learners than by native speakers. Although not wrong in the strict sense, underused or overused items contribute to the non-native character of learner language and can, thus, inform teaching, especially at more advanced levels.

By identifying misuses, underuses, and overuses on the basis of learner corpora, one can provide a more reliable *description* of learner language. De Cock and Granger (2005), thus, argue that intuition-based lists of errors like Swan and Smith (1987) include misuses that are unlikely to be found among learners. The identification of problematic areas in learner corpora can also help with the *selection* of the most relevant items to include in the teaching syllabus or in pedagogical resources. Thewissen (2015, p. 201), for example, shows that punctuation errors are frequent and "improvement-resistant" in the learner corpus she investigates, which is a signal that more time could be devoted to punctuation in the English as a Foreign Language (EFL) classroom. On the other hand, learner corpora can be approached from a more positive perspective, looking at what learners get right. In Thewissen's (2015) study, adjective order errors are very infrequent, for instance, which suggests that it is probably not necessary to emphasise this aspect more in ELT. Learner corpora, therefore, provide "useful information … on the areas of language which should be reinforced or de-emphasised" (Granger, 2015, p. 488).

Usually, learner corpora come with rich metadata about learners (e.g. their mother tongue or the number of years they have been learning the target language). This is valuable information for teaching, because it can lead to *customisation*, i.e. the adaptation of teaching or materials to the target learner population (e.g. Chinese-speaking learners, beginners). Gilquin and Granger (2021), thus, show that the phrase *as far as x BE concerned* is particularly common in corpus data produced by French-speaking learners, who often use it as a topic introducer at the beginning of the sentence (e.g. *As far as dreams are concerned, the same could be said*). Such a finding could lead to a pedagogical intervention specifically targeted at French-speaking learners. Metadata about learners' proficiency levels can also help with sequencing, i.e. "deciding the order in which linguistic items should be presented" (Granger, 2015, p. 488). This is the aim of the English Profile project (https://www.englishprofile.org/), which uses learner corpus data to investigate what learners at different proficiency levels of the Common European Framework of Reference for Languages (CEFR) can do. On this basis, it is possible to design pedagogical materials that take account of learners' proficiency.

Using written learner corpora to inform teaching can be done in a direct or indirect way, a distinction first introduced by Leech (1997) for corpora of native language, but later applied to learner corpora too (Granger, 2015). A direct approach means that teachers and/or their students consult learner corpora themselves to find out how learners (possibly the students themselves) use the target language. An indirect approach, on the other hand, involves researchers, pedagogical materials writers, software developers, or publishers relying on learner corpora to create resources (dictionaries, textbooks, computer programs, etc.) that are then used by teachers and/or their students. This chapter considers both direct and indirect pedagogical approaches to written learner corpora.

2. Review of current state of research

From the early days of corpus linguistics, it was possible to use corpora of native language to inform teaching. However, the compilation of the first learner corpora, among which the International Corpus of Learner English (ICLE; Granger, 1993) and the Longman Learners' Corpus, together with the development of the field of learner corpus research (see Granger et al., 2015), certainly contributed to a growing interest in the pedagogical applications of corpora. This is due not only to the availability of a new type of corpus, but also to the fact that learner corpus research has always had strong links with the field of language teaching (Granger, 2009) and has, thus, helped draw the attention of pedagogues and pedagogically-oriented researchers to corpora and corpus linguistics. Yet, it is fair to say that learner corpora (written or otherwise, in English or in other languages) have not played a central role in language teaching so far. Flowerdew (2012, p. 207) mentions learner corpora among the "under-represented corpora for pedagogy", noting that "there is little evidence that learner corpora have had much impact on syllabus and materials design to date" (Flowerdew, 2012, p. 210). However, she also underlines "the acknowledged value in integrating learner corpora into language teaching" (Flowerdew, 2012, p. 210). In what follows, we will review some attempts at using written learner corpora directly or indirectly for ELT purposes. For the latter approach, we will consider the contribution of learner corpora to dictionaries (Section 2.1) and to other pedagogical resources, including grammars and textbooks (Section 2.2). For the direct approach, we will see how learner corpora have been used in data-driven learning (Section 2.3).

2.1 Learner dictionaries

One area that has exploited learner corpora successfully is lexicography. De Cock and Granger (2005, p. 72) point out that "[t]he addition of this new resource to the lexicographer's workstation constitutes a new departure in pedagogical lexicography". The exploitation of learner corpora makes it possible to identify learners' main problems with the target language and to then warn them about these potential pitfalls. Error dictionaries (e.g. Turton & Heaton, 1996) and so-called "error notes" in learner dictionaries illustrate errors regularly made by learners and show how they can be corrected so as to reflect native usage, e.g.

(1) ✗ She gave me a good advice.

✓ **She gave me some good advice.**

✗ It is full of good advices on healthy eating.

✓ **It is full of good advice on healthy eating.**

(Turton & Heaton, 1996, p. 9)

The second edition of the *Macmillan English Dictionary for Advanced Learners* (MEDAL2; Rundell, 2007) includes 30 central pages that are meant as a guide to produce better academic writing and professional reports (Gilquin et al., 2007a). The contents, which are organised around a number of rhetorical functions (such as "expressing possibility and certainty", "quoting and reporting", or "summarizing and drawing conclusions"), are entirely based on the close analysis of native and learner corpora. In addition to "Get it right" boxes, learners can find "Be careful" notes that point, among others, to words or expressions that are under- or overused by learners as indicated in Example (2). These notes also give information about words or expressions that tend to be misplaced in the sentence as with the word *therefore* discussed in Example (3), or choices that are not stylistically appropriate as shown in

Example (4). Some of these notes are illustrated by means of a bar chart comparing the frequency of the items in learner writing and native writing/speech.

(2) Learners often use the preposition *in spite of*. However, *despite* is much more frequent. (Gilquin et al., 2007a, p. IW20)

(3) Learners often use *therefore* at the beginning of a sentence. This use is correct, but it is much less frequent than the use of *therefore* inside the sentence. (Gilquin et al., 2007a, p. IW13)

(4) Although, in informal style, *that* can be left out after very frequent reporting verbs such as *suggest*, *suppose*, and *think*, this is less frequently the case in academic writing and professional reports. (Gilquin et al., 2007a, p. IW27)

As suggested by Walter (2010, p. 440), items that are overused by learners can also point to good candidates for the insertion of thesaurus-type information in dictionaries, which would provide learners with good alternatives to replace these overused items. In addition, learner corpora can help identify words that are likely to be familiar to learners and that, included in the defining vocabulary of a dictionary, would result in definitions that are understandable to most learners (Gillard & Gadsby, 1998, p. 163).

Thanks to the metadata in learner corpora, and with the support of electronic lexicography, it has also become easier to customise dictionaries according to learners' specific needs. Granger (2018) shows how the analysis of learner corpus data distinguished by learners' mother tongues (L1) can help develop a bilingualised electronic dictionary which includes L1-specific error notes. She also argues for the inclusion of such L1-specific error notes in bilingual dictionaries, where she claims they would be particularly useful given learners' preference for bilingual dictionaries over monolingual dictionaries (Granger, 2018).

2.2 Other pedagogical resources based on learner corpora

In addition to dictionaries, other pedagogical resources have also started to exploit learner corpora in different ways.[1] Information derived from learner corpora can help grammarians or textbook writers decide what to focus on. Thus, the corpus-informed *Cambridge Grammar of English* includes an A-Z in which the words were selected for inclusion because, among other reasons, they are "known to be difficult for learners of English and often lead to errors" (Carter & McCarthy, 2006, p. 21). In the *Objective IELTS Intermediate Workbook*, some of the exercises centre around words that learners have difficulty with, for example prepositions (Black & Sharp, 2006a, p. 53), as attested by a learner corpus.

The types of error notes that have by now become quite common in learner dictionaries have also progressively made their appearance in other pedagogical resources. This is the case, for instance, in the *Common Mistakes at…* series (e.g. Moore, 2005), the *Cambridge Grammar of English* (Carter & McCarthy, 2006), and the *Viewpoint* (e.g. McCarthy et al., 2012) or *Grammar and Beyond* (e.g. Reppen, 2012) textbooks, as illustrated below:

(5) [*Crucial*, *vital*, *essential*, *fundamental*] are 'limit' adjectives and are not normally used with *very / quite / more*, etc. To emphasise the adjective, you can use *absolutely vital / essential*, etc.:
It's absolutely essential to start off with a good business plan. (not ~~***it's very essential***~~) (Moore, 2005, p. 42)

(6) Do not start a sentence with *Whereas* to contrast ideas with a previous sentence.
An online profile is for friends. ***However,*** *a résumé is for employers.* (NOT ~~*Whereas*~~...) (McCarthy et al., 2012, p. 18)

In addition, textbooks based on the analysis of learner corpora can include exercises related to learners' typical errors. Such exercises can take different forms. They can consist of sentences containing one or several errors that students have to correct, e.g. *Yesterday was much more funnier than the first day* (*Interactive*, Hadkins et al., 2011, p. 39). In *Objective PET* (Hashemi & Thomas, 2010), such exercises are called "corpus spots" and they focus on specific topics, for instance pluralisation (e.g. *I look after the childs when their parents are working*, Hashemi & Thomas, 2010, p. 21), *–ing* vs *–ed* forms in adjective positions (e.g. *She was amazed / amazing by the shops and restaurants*, Hashemi & Thomas, 2010, p. 42), or verb + noun collocations (e.g. *I hope you don't do the same mistake as me*, Hashemi & Thomas, 2010, p. 56). Some exercises go beyond isolated sentences by presenting long extracts or full texts taken from a learner corpus. This is the case in the *First Certificate Trainer* (May, 2010). Students are, for example, shown a letter written by a First Certificate candidate and asked to carry out some specific tasks:

(7) Find and correct the following (1–3):

1 poor layout. Where should it be divided into paragraphs?

2 two informal expressions, four contracted forms and four uses of informal punctuation. Change these to more formal language.

3 two mistakes each in verb forms, spelling and capital letters. Correct these. (May, 2010, p. 25)

Interestingly, some learner texts are also presented as models for students. An application letter "written by Felipe, a very strong First Certificate candidate" (May, 2010, p. 26), for instance, is accompanied by notes highlighting the positive features of the candidate's letter (e.g. "correct structure for current job", "formal linking expressions", "polite to the employer") and encouraging students to write their own letter on the basis of this model.

As is the case with dictionaries based on learner corpora, customisation of other pedagogical resources is possible thanks to metadata. The *Common Mistakes at…* series mentioned above uses certain portions of the Cambridge Learner Corpus to identify errors made by learners at a certain proficiency level. *Objective PET* focuses on errors typical of the level of the Preliminary English Test (Hashemi & Thomas, 2010). Customisation according to learners' L1 is still relatively rare, but Cambridge University Press offers several of its learner-corpus-based textbooks in "English for Spanish Speakers" editions. *Objective First*, for example, has a booklet entitled *100 Writing Tips for Cambridge English: First*, whose contents "have been informed by a study of B2 level Spanish speakers' data in the *Cambridge Learner Corpus*" (Capel, 2014, p. 1). *English in Mind, Italian Edition* (e.g. Puchta & Stranks, 2007) is informed by learner corpus data produced by Italian-speaking learners.

2.3 Data-driven learning and learner corpora

Data-driven learning (DDL), which involves the use of corpora in the language classroom (see Section 4 in this handbook), has mostly relied on native corpora, helping students discover how the target language should be employed. Despite a very early demonstration of the potential of learner corpora for DDL (Granger & Tribble, 1998) and several calls since then to integrate learner corpus data into DDL (e.g. Gilquin & Granger, 2010; Seidlhofer, 2002), only a handful of DDL studies so far have ventured to use learner corpora. Thus, in Chen and Flowerdew's (2018) review of empirical studies on DDL in the English for Academic Purposes classroom published between 2000 and 2017, only five studies out of thirty-seven include learner corpus data.

Doing DDL with written learner corpora means that the typical features of learner writing can be examined. Most of the time, learner corpora are used in combination with native corpora, so that students can compare the features of learner writing with those of native writing. As is the case with pedagogical resources based on learner corpora, the focus of learner-corpus-based DDL is often a problematic aspect of learner writing. Nesselhauf (2004), for example, shows how the comparison of a concordance of the verb SUGGEST in the German component of ICLE and in the Louvain Corpus of Native English Essays (LOCNESS) can make students aware of the non-standard use of SUGGEST followed by a *to*-infinitive.

While off-the-shelf learner corpora like ICLE can serve as a source of information about learner writing for DDL, most researchers or teachers who have integrated learner corpora into DDL activities have actually collected learner corpus data themselves (or sometimes asked their students to do so) in the form of a "local learner corpus" (see Seidlhofer, 2002). A local learner corpus includes texts produced by the students for whom the DDL activities are meant to be designed and implemented. Moon and Oh (2017), for example, collected writing from Korean learners of English at the beginning of the school year. The analysis of this home-made learner corpus revealed learners' tendency to overgeneralise *BE*, producing sentences such as *He* ***is*** *dance very well* (Moon & Oh, 2017, p. 52). Some of the students who contributed to the learner corpus then took part in DDL activities which required them to compare concordances of *is* in the learner corpus and in a native corpus of graded reader texts, so that they could "unlearn the overgenerated *be*" (Moon & Oh, 2017, p. 54). The students, thus, examined learner corpus data produced by themselves and their classmates, and they focused on a feature that had proved problematic for them. The use of a local learner corpus has the advantage of making the DDL activities particularly relevant to the learners and arguably more motivating for them. In Lee and Swales (2006), the relevance is even more obvious since each participant collected a learner corpus of their own writing and compared it with a corpus of expert writing in their field of study. The learner language features that they discovered by comparing the two corpora were, therefore, features which they had produced themselves and which they could correct thanks to the model of the expert corpus. As Lee and Swales (2006, p. 68) underline, the comparison also brought to light the relative lack of variation in students' writing, which could encourage them to look for alternative expressions or patterns in the expert corpus. The usefulness of local learner corpora in DDL is demonstrated by Cotos (2014), who compared DDL activities combining a local learner corpus and a native corpus with DDL activities involving native corpus data only, and who showed that the former are more efficient than the latter.

3. Critical issues

The literature review in the previous section has established that written learner corpora can inform teaching either by contributing to the production of better pedagogical resources more suited to learners' needs, or by giving learners access to authentic learner data, from which they can make their own discoveries. However, this pedagogical potential has not been exploited to the full, although, as suggested by Granger (2015, p. 487), "there are clear signs that the tide is turning, albeit slowly". Some of the issues associated with the use of learner corpora for pedagogical purposes may explain their relative lack of uptake in foreign language teaching (Meunier, 2012).

The first issue has to do with the controversy surrounding the exposure of learners to negative evidence, that is, to erroneous forms of the target language (Flowerdew, 2001). It has been claimed that this may have a detrimental effect on learners by reinforcing (or sometimes even creating) problems in their language production. However, Fuster-Márquez and Gregori-Signes (2018, p. 164), dealing with the use of learner corpora in DDL, show that, provided the data are "authentic and highly specific learner data obtained from a reliable ad hoc learner corpus", "direct exposure to these data through controlled activities may cover certain learners' needs not found in textbooks". In addition, the potentially detrimental effect of negative evidence can be counterbalanced by positive evidence coming from a native corpus or from the learner corpus itself and by exercises that consolidate the standard form (Nesselhauf, 2004).

Another issue concerns the teacher. Despite their central role, "[t]he language teacher is an often neglected figure in learner corpora projects", as Urzúa (2015, p. 99) stresses. Meunier (2012, p. 211) notes that many teachers "are not aware of the possibilities offered by (learner) corpora and of the changes that corpus methods have brought to materials that they are using". She also points out that most teachers receive no or very little (pre- or in-service) teacher training in the use of corpora. If they have not been trained in corpus linguistics, they are unlikely to fully benefit from learner corpora or make their students benefit from them. Urzúa (2015) describes a project in which teachers are involved in the compilation of a local learner corpus of written academic English. Thanks to the training in corpus linguistics received within the framework of this project, they are able to analyse the learner corpus data and interpret their findings on the basis of what they know about their students. By examining their students' actual usage, they are also led to re-evaluate the curriculum and rethink some of the writing tasks they give to their students. Getting teachers involved in learner corpus projects may thus have beneficial consequences, both for them and for their students. Providing them with ready-made materials (e.g. downloadable DDL worksheets) which they could use with their students in the classroom could arguably encourage them to get involved.

A third possible reason for the relative lack of uptake of learner corpus information in foreign language teaching is the conservative position of many educational publishers, who are thought to be, as Harwood (2005, p. 152) puts it, "far more comfortable with rehashes of what has gone before than with something different (and refreshing)". Fortunately, more and more publishers have now integrated corpus data into their pedagogical materials, so that "what has gone before" is increasingly likely to be based on learner corpora. However, as Harwood (2005, p. 152) also reminds us, "[m]arketability rather than pedagogical effectiveness is said to be the publishers' main concern". Thus, for commercial reasons rather than pedagogical ones, the publication of learner-corpus-based customised materials, in particular, L1-specific materials, is less probable simply because the market for such materials is smaller (Gilquin et al., 2007b). In this respect, it is to be hoped that, in the same way as electronic lexicography has started to open new doors for the customisation of learner dictionaries, the increasing use of electronic resources in the classroom will lead educational publishers to propose innovative ways of providing students with tailor-made materials based on learner corpora.

4. Recommendations for practice

If we want written learner corpora to play a more prominent role in teaching, we need to address the critical issues outlined in the previous section, but we also need to make sure we adopt good practices when we use learner corpora with pedagogical aims in mind. This section offers some recommendations related to the choice of a norm (Section 4.1), the representativeness of the results (Section 4.2), and the transparency of information derived from learner corpora (Section 4.3).

4.1 Norm

When a written learner corpus is used for teaching purposes, this is often in combination with a native corpus, which makes it possible, by comparison, to identify non-native features in the learner corpus. The notions of under- and overuse, for example, necessarily imply a statistically significant difference in frequency between the learner corpus and a native corpus. As for DDL, it was emphasised earlier that native corpus data are important because they provide students with an indication of what is right in the target language. Different types of corpora can be used to represent the target language. In fact, it need not even be a native corpus: an expert corpus consisting of articles published in scientific journals (and not necessarily written by native speakers), as in Lee and Swales (2006), can constitute an excellent target for students in an English for academic purposes course. Just confining our attention to native corpora, however, there are several choices available, and one's choice may have an impact on the results (e.g. whether an item turns out to be overused or not by learners; Gilquin, 2021a).

The most important principle that should guide this choice is comparability: the native corpus should be written, like the learner corpus, but it should also correspond to the genre of the learner corpus as closely as possible. In this regard, Nesselhauf's (2004) decision to compare ICLE to LOCNESS makes perfect sense, since both corpora include argumentative and literary essays. The fact that LOCNESS includes data produced by university students, like ICLE, can be both a strength and a weakness. It is a strength in the sense that the writers are about the same age as the students represented in ICLE and they are supposed to have reached a similar cognitive development. On the other hand, as novice writers rather than expert writers, native students tend to produce language that learners would not necessarily want to imitate, as Leech (1998) already recognised.

In the MEDAL2 project described earlier (Gilquin et al., 2007a), the initial decision had been to compare the ICLE data with LOCNESS. However, the first analyses made it clear that presenting students with a model primarily based on LOCNESS would not be pedagogically sound; instead, it was decided to use a corpus of academic English by expert native writers as the main reference, despite the difference in genre. When the objective involves bringing learners closer to the (ideal) target, it is important that they are mainly presented with this target and that other models are used cautiously, if at all, to avoid leading learners to commit errors that they might not have committed otherwise (e.g. the confusion between *it's* and *its*, which is quite common in novice native writing, but rarely occurs in non-native writing).

4.2 Representativeness

Recommendations can also be made with respect to the representativeness of the findings. If one aims to cater for a specific group of students, through the use of customised materials, the learner corpus data should represent the writing of learners whose profiles resemble those of the students as much as possible (same mother tongue background, same acquisitional context, etc.). Local learner corpora are the ultimate candidate for representativeness, since the learners represented in the corpus are also the beneficiaries of the corpus-based lessons or materials.

If one does not aim for customised materials, on the other hand, one should provide information that is as relevant as possible to the largest number of learners. Talking about error notes, Granger (2015, pp. 492-493) rightly points out that "[i]t is counterproductive and, indeed, potentially detrimental to include warnings that are only relevant for a limited group of learners". She gives the example of an error note in the *Cambridge Grammar of English* (Carter & McCarthy, 2006) which warns against the over-passivisation of copular verbs (e.g. **A teacher is been by her*), while this turns out to be a relatively rare phenomenon in learner corpus data, mainly found among populations with certain mother tongue backgrounds (most notably Asian languages).

In the MEDAL2 project, whose aim was to produce a resource that would be helpful to learners worldwide (and not to a specific group of learners), non-native features that were highlighted had to be both frequent and widespread. The latter criterion was operationalised by specifying that the features had to appear in at least ten L1 learner populations out of the sixteen that were investigated. In effect, this also corresponded to a spread of the features across different language families. This criterion explains why, for instance, *in fact* was not mentioned among the overused items. Although it was heavily overused by certain learner populations, the number of populations was too limited, also in terms of language families (mainly Italian- and French-speaking learners overuse *in fact*), to make it worthwhile including a “Be careful” note in the dictionary about this item.

O’Keeffe and Mark (2017), who describe methodological issues related to the English Profile project, and more particularly the English Grammar Profile, do the opposite of the MEDAL2 project: instead of looking for what learners get wrong in the learner corpus, they look for what learners get right, in order to arrive at a number of so called “grammatical competence statements”, which indicate “what learners can do with grammar at each level of the CEFR based on what they have written in Cambridge exams” (O’Keeffe & Mark, 2017, p. 464). Yet, the criteria they apply are quite similar to those of the MEDAL2 project, including “frequency of use”, “rate of correct uses”, and “spread of first language families” (O’Keeffe & Mark, 2017, pp. 469-470). In addition, they recommend considering whether the usage is distributed across a range of individual learners, whether it is distributed across a range of contexts (e.g. letters, informative texts, essays, reports), and whether it is affected by a task (e.g. the high frequency of the pattern *Would you mind* ... resulting from the task of writing a letter to ask for a different appointment). Adopting such criteria should enable all learners to benefit equally from corpus-based resources, regardless of their profiles or the context of the writing task.

4.3 Transparency

The last recommendation is for materials writers to be more transparent about the source of the information (corpus or otherwise) that they include in their pedagogical resources. Very often, the resources provide a general mention of the learner corpus that they rely on. For printed grammars and textbooks, this is typically done on the cover of the book, with a few words about the learner corpus and its general contribution to the resource. However, inside the resource, it may not always be clear what parts of the book precisely result from the analysis of the learner corpus or how exactly the corpus data have been used. In the *Cambridge Grammar of English* (Carter & McCarthy, 2006), for example, some error warnings are signalled by a special symbol. This is how these warnings are described in the introduction to the grammar:

> We also had access during the writing of this book to a large learner corpus consisting of texts produced by learners of English from a wide range of lingua-cultures, coded for error and inappropriate use. This, along with our own language-teaching experience and that of our reference panel, has enabled us to give warnings of common areas of potential error where appropriate. (Carter & McCarthy, 2006, p. 3)

While this confirms the use of a learner corpus, it does not necessarily seem to guarantee that this corpus was used as the (main) source of information for all the warnings. In addition, the grammar includes many crossed out sentences, as in Example (8), without it being made clear whether these come from the learner corpus.

(8) I may be free. I'll have to check my diary. (~~I'll have got to check my diary.~~)
(Carter & McCarthy, 2006, p. 403)

In the *Objective IELTS Intermediate Student's Book* (Black & Sharp, 2006b), some exercises are said to be *based on* learner corpus data, which leaves it unclear how authentic they actually are, as illustrated by the following instructions: "Read the question below and the letter in response to it, which is based on an answer produced by an IELTS candidate" (Black & Sharp, 2006b, p. 57; emphasis added). It is also quite common to notice a mixture of authentic and (presumably) invented sentences, a distinction not always drawn very markedly. This mixture can be found within one and the same exercise, e.g. "Look at these sentences from job applications, some of them written by IELTS candidates" (Black & Sharp, 2006b, p. 57; emphasis added). It can also be found across similar exercises, for example error detection exercises in the *Objective IELTS Intermediate Workbook* (Black & Sharp, 2006a), some of which are explicitly described as taken from authentic data (e.g. "Correct the mistakes below made by IELTS candidates", Black & Sharp, 2006a, p. 49) whereas others are not (e.g. "Correct the 12 spelling mistakes in this paragraph", Black & Sharp, 2006a, p. 45). While all the pedagogical resources that rely on some learner corpus data are obviously a step in the right direction and while we cannot necessarily expect all the materials in these resources to be entirely derived from (learner) corpora, clearly signposting the contents that come from a corpus (by means of a symbol or a title, like the corpus spots in Hashemi & Thomas, 2010) would seem like a good practice to follow for materials writers.

5. Future directions of research

This final section considers some ways in which written learner corpora could further inform teaching in the future. The first direction is that of computer-assisted language learning (Section 5.1) and the second one is the study of the writing process (Section 5.2).

5.1 Computer-assisted language learning

Computer-assisted language learning (CALL) applications based on (written) learner corpora are by no means new. As early as 1992, for example, Liou et al. reported on a CALL project aimed at improving EFL learners' grammar, which relied on the analysis of an error-tagged corpus of essays written by Chinese students. CALL applications based on (written) learner corpora are not rare either. In fact, Granger (2015, p. 496) notes that “there is a much greater – and more diversified – use of learner corpus data in CALL materials than in other pedagogical resources”. Yet, she also stresses that “[m]ost tools are still at an experimental stage and presented as prototypes” (Granger, 2015, p. 499).

Like other pedagogical resources, CALL projects based on learner corpora mostly use them as repertoires of errors typically committed by (certain populations of) learners. Lee et al. (2016), for example, use error-tagged learner corpora to create a CALL system that automatically generates fill-in-the-blank exercises for preposition usage. The learner corpus data also serve to create plausible distractors, which are shown through an experiment to “rival … the quality of human-authored items” (Lee et al., 2016, p. 991). Lo et al. (2018) describe GEC Cool Edit, a CALL system for grammatical error correction (GEC) which relies on parallel sentences taken from an error-tagged learner corpus, with the original sentence and its corrected version. The web-based system allows students to write a text and automatically receive corrective feedback. The evaluation of the system points to its “competitive performance on a number of publicly available testsets” (Lo et al., 2018, p. 82) and the authors mention some avenues for further improvement of the system. These two examples should suffice to illustrate the great potential of CALL applications that rely on learner corpora, and we can only hope, as Granger (2015, p. 499) does, that “the buzzing activity in the field will lead to the production of fully fledged applications in the near future”.

5.2 Writing process

The resources and applications mentioned so far exploit written corpora that are made up of learners’ texts, that is, the finished products of the writing act. Some learner corpora, however, seek to represent the writing process, that is, how learners go about composing a text (Gilquin, 2021b). This is the case of learner corpora which include annotations showing revisions in handwritten texts, e.g. crossed out items or elements added in superscript, as in the Marburg corpus of Intermediate Learner English (Kreyer, 2015). While not all revisions are visible in such corpora (a crossed out item may be illegible or the learner may have used correction fluid), they provide valuable information about items that the learner may have got right in the end but that caused difficulties during the process, or about elements that the learner tried to write but that eventually disappeared from the text as a result of an avoidance strategy, all of which could usefully inform teaching.

Learner corpora that include several drafts of one and the same text, like the CityU Corpus of Essay Drafts of English Language Learners (Lee et al., 2015), also give snapshots of the writing process. They make it possible, for example, to examine how a text was revised and how learners addressed the feedback that they received. In terms of teaching applications, they could for instance be presented to learners to give them concrete examples of how a text can be gradually improved to reach a satisfactory result. In DDL activities, students could be asked to compare the different drafts of a text and be made to notice the improvements. They could also compare a sub-corpus made up of the first drafts with a sub-corpus made up of the final drafts, and be asked to identify the differences, for example items whose frequency has significantly increased or decreased. The final versions of the texts would, thus, serve as a reference corpus, which would make it unnecessary to resort to native corpora. Multiple-draft learner corpora could also be useful for CALL. Actually, the CityU Corpus is one of the learner corpora used in Lee et al. (2016) to create distractors in fill-in-the-blank exercises (see Section 5.1). This was done by looking for prepositions that were often edited between drafts, the assumption being that “frequent editing implies a degree of uncertainty on the part of the learner as to which of these prepositions is in fact correct” (Lee et al., 2016, p. 987).

Going even one step further in the representation of the writing process, the Process Corpus of English in Education (PROCEED) is a process-focused learner corpus that reproduces the whole writing process through the inclusion of keystroke logs, which contain a record of every key struck on the keyboard, and screencast videos, which show what happens on the computer screen while the learner is typing (Gilquin, 2022). Such data give an accurate picture of every revision that the text undergoes (deletion, insertion, movement, etc.) and in what order these revisions are carried out. The screencast videos make it possible to see the state of the text at each stage in the writing process and to observe its evolution second by second. The keystroke logs provide detailed statistics about the number and types of revisions, or the frequency and duration of pauses, for instance. Process data like these offer new possibilities in terms of teaching. Gilquin (2019) describes a pedagogical intervention which used PROCEED as a local learner corpus. Students were first asked to watch at least ten minutes of their own screencast video, which they then discussed individually with the teacher. The learner and the teacher also examined the learner's keystroke report, which was compared to the report of a proficient learner and that of a native writer, thus highlighting differences in the writing process (e.g. time devoted to reviewing the whole text at the end). Finally, the teacher showed some anonymised extracts from other screencast videos illustrating successful writing strategies and asked the learner to reflect on how they could adopt some of these strategies in their own writing. Focusing on another aspect of the writing process, Gilquin and Laporte (2021) explain how videos from PROCEED were annotated as to learners' use of online writing tools (dictionaries, corpora, etc.), thus revealing, among other findings, that learners predominantly rely on bilingual tools, that they tend to look for a single word using a single tool, and that they often lack critical thinking about the information which they retrieve from online tools. These findings were then exploited to create a self-learning online platform to help learners make better use of online writing tools. The platform consists of a number of modules, each one focusing on a particular type of tool or information to search for (e.g. monolingual dictionaries, thesauri, collocations, frequency). In each module, a video introduces the topic and demonstrates some tools, with particular attention to aspects that proved problematic according to the screencast videos. Each module also includes exercises that are based on sentences or text samples from PROCEED and that involve the use of specific (combinations of) tools to improve the sentence or text.

Following this and other avenues for research, and continuing to explore paths already trodden, we can hope that written learner corpora will further enhance teaching, enabling researchers, practitioners, and ultimately learners to fully benefit from their many advantages.

Further reading

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*(4), 319-335.

This is a plea for the inclusion of learner corpora in English for academic purposes pedagogy. It also includes a description of the MEDAL2 lexicographical project.

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

This handbook provides a state of the art in the field of learner corpus research. A whole section is devoted to “Learner corpus research and language teaching”, including Granger’s (2015) chapter on “The contribution of learner corpora to reference and instructional materials design”.

McCarthy, M. (2016). Putting the CEFR to good use: Designing grammars based on learner-corpus evidence. *Language Teaching*, *49*(1), 99-115.

This is a convincing demonstration of the powerful evidence offered by learner corpora to inform grammar teaching. The framework for the article is that of the English Profile project.

Notes

[1] Most of the resources mentioned in this section are published by Cambridge University Press, which has been a pioneer in the production of pedagogical materials based on learner corpora. The Cambridge Learner Corpus, on which these resources rely, is primarily made up of Cambridge examination scripts (i.e. learner writing) produced by learners from a range of mother tongue backgrounds. While McCarthy (2016) points out that spoken data are gradually being added to the corpus, it is unlikely that the resources described here have already benefited from the use of such spoken data.

References

Black, M., & Sharp, W. (2006a). *Objective IELTS intermediate workbook with answers*. Cambridge University Press.

Black, M., & Sharp, W. (2006b). *Objective IELTS intermediate student's book*. Cambridge University Press.

Capel, A. (2014). *Objective First*, *100 writing tips for Cambridge English: First*. Cambridge University Press.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge University Press.

Centre for English Corpus Linguistics. (2020). Learner corpora around the world. Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, *23*(3), 335-369. https://doi.org/10.1075/ijcl.16130.che

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*(2), 202-224. https://doi.org/10.1017/S0958344014000019

De Cock, S., & Granger, S. (2005). Computer learner corpora and monolingual learners' dictionaries: The perfect match. *Lexicographica*, *20*, 72-86. https://doi.org/10.1515/9783484604674.72

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). Benjamins. https://doi.org/10.1075/scl.5.21flo

Flowerdew, L. (2012). *Corpora and language education*. Palgrave Macmillan. https://doi.org/10.1057/9780230355569

Fuster-Márquez, M., & Gregori-Signes, C. (2018). Learning from learners: A non-standard direct approach to the teaching of writing skills in EFL in a university context. *Innovation in Language Learning and Teaching*, *12*(2), 164-176. https://doi.org/10.1080/17501229.2016.1142549

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

Gilquin, G. (2019, August 28-31). *Screencasting and keylogging as pedagogical tools to enhance writing skill development* [Paper presentation]. 27th EUROCALL conference, Louvain-la-Neuve, Belgium.

Gilquin, G. (2021a). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, *55*(1), 87-99. https://doi.org/10.1017/S0261444821000094

Gilquin, G. (2021b). Hic sunt dracones: Exploring some *terra incognita* in learner corpus research. In A. Čermáková & M. Malá (Eds.), *Variation in time and space: Observing the world through corpora* (pp. 65-86). De Gruyter.

Gilquin, G. (2022). The *Process Corpus of English in Education*: Going beyond the written text. *Research in Corpus Linguistics*, *10*(1), 31-44. https://doi.org/10.32714/ricl.10.01.02

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). Routledge.

Gilquin, G., & Granger, S. (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* (pp. 72-98). Multilingual Matters.

Gilquin, G., Granger, S., & Paquot, M. (2007a). Writing sections. In M. Rundell (Ed.), *Macmillan English dictionary for advanced learners. Second edition* (pp. IW1-IW29). Macmillan.

Gilquin, G., Granger, S., & Paquot, M. (2007b). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*(4), 319-335. https://doi.org/10.1016/j.jeap.2007.09.007

Gilquin, G., & Laporte, S. (2021). The use of online writing tools by learners of English: Evidence from a process corpus. *International Journal of Lexicography*, *34*(4), 472-492. https://doi.org/10.1093/ijl/ecab012

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 57-69). Rodopi.

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). Benjamins. https://doi.org/10.1075/scl.33.04gra

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485-510). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022

Granger, S. (2018). Has lexicography reaped the full benefit of the (learner) corpus revolution? In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX international congress: Lexicography in global contexts* (pp. 17-24). Ljubljana University Press. Available at https://euralex.org/category/publications/euralex-2018/.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199-209). Longman.

Hadkins, H., Lewis, S., & Budden, J. (2011). *Interactive. Student's book 2*. Cambridge University Press.

Harwood, N. (2005). What do we want EAP teaching materials for? *Journal of English for Academic Purposes*, *4*(2), 149-161. https://doi.org/10.1016/j.jeap.2004.07.008

Hashemi, L., & Thomas, B. (2010). *Objective PET. Student's book with answers*. Cambridge University Press.

Kreyer, R. (2015). The Marburg Corpus of Intermediate Learner English (MILE). In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 13-34). Benjamins. https://doi.org/10.1075/scl.70.01kre

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, *25*(1), 56-75. https://doi.org/10.1016/j.esp.2005.02.010

Lee, J., Sturgeon, D., & Luo, M. (2016). A CALL system for learning preposition usage. *Proceedings of the 54th annual meeting of the association for computational linguistics, Berlin, August 7-12, 2016* (pp. 984-993). Association for Computational Linguistics.

Lee, J., Yan Yeung, C., Zeldes, A., Reznicek, M., Lüdeling, A., & Webster, J. (2015). CityU corpus of essay drafts of English language learners: A corpus of textual revision in second language writing. *Language Resources and Evaluation*, *49*(3), 659-683. https://doi.org/10.1007/s10579-015-9301-z

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). Longman.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Longman.

Liou, H.-C., Wang, S. H., & Hung-Yeh, Y. (1992). Can grammatical CALL help EFL writing instruction? *CALICO Journal*, *10*(1), 23-44.

Lo, Y.-C., Chen, J.-J., Yang, C.-Y., & Chang, J. S. (2018). *Cool English*: A grammatical error correction system based on large learner corpora. *Proceedings of the 27th international conference on computational linguistics: System demonstrations* (pp. 82-85). Santa Fe, New Mexico, August 20-26, 2018.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135-157). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007

May, P. (2010). *First Certificate Trainer. Six Practice Tests with Answers*. Cambridge University Press.

McCarthy, M. (2016). Putting the CEFR to good use: Designing grammars based on learner-corpus evidence. *Language Teaching*, *49*(1), 99-115. https://doi.org/10.1017/S0261444813000189

McCarthy, M., McCarten, J., & Sandiford, H. (2012). *Viewpoint 1. Student's Book*. Cambridge University Press.

Meunier, F. (2012). Learner corpora in the classroom: A useful and sustainable didactic resource. In L. Pedrazzini & A. Nava (Eds.), *Learning and teaching English: Insights from research* (pp. 211-228). Polimetrica.

Moon, S., & Oh, S.-Y. (2017). Unlearning overgenerated *be* through data-driven learning in the secondary EFL classroom. *ReCALL*, *30*(1), 48-67. https://doi.org/10.1017/S0958344017000246

Moore, J. (2005). *Common mistakes at Proficiency ... and how to avoid them*. Cambridge University Press.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Benjamins. https://doi.org/10.1075/scl.12.11nes

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. https://doi.org/10.1075/ijcl.14086.oke

Puchta, H., & Stranks, J. (2007). *English in mind, Italian edition. Student's book 1. Second edition*. Cambridge University Press.

Reppen, R. (2012). *Grammar and beyond, level 1. Student's book*. Cambridge University Press.

Rundell, M. (Ed.). (2007). *Macmillan English dictionary for advanced learners. Second Edition*. Macmillan Education.

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). Benjamins. https://doi.org/10.1075/lllt.6.14sei

Swan, M., & Smith, B. (1987). *Learner English. A teacher's guide to interference and other problems*. Cambridge University Press.

Thewissen, J. (2015). *Accuracy across proficiency levels: A learner corpus approach*. Presses universitaires de Louvain.

Turton, N. D., & Heaton, J. B. (1996). *Longman dictionary of common errors*. Longman.

Urzúa, A. (2015). Corpora, context, and language teachers: Teacher involvement in a local learner corpus project. In V. Cortes & E. Csomay (Eds.), *Corpus-based research in applied linguistics. Studies in honor of Doug Biber* (pp. 99-122). Benjamins. https://doi.org/10.1075/scl.66.05urz

Walter, E. (2010). Using corpora to write dictionaries. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 428-443). Routledge. https://doi.org/10.4324/9780203856949.ch31