Gilquin, G. 2021. '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). Berlin: De Gruyter.

Gaëtanelle Gilquin

Hic sunt dracones

Exploring some *terra incognita* in learner corpus research

Abstract: While the field of learner corpus research has made a great deal of progress since its advent some thirty years ago, there are still areas that remain to be explored. In this chapter, two such areas, related to temporal evolution, are demarcated. The first one is diachronic learner corpus research, which considers the possible evolution of learner language through time, as a result of changes in the typical acquisitional context or in the norms that learners depend on, for example. The second one is process learner corpus research, which uses screencasting and keyboard logging to examine the evolution of learner texts during the writing process and to gain insights into learners' competence on that basis. These areas require new types of learner corpora, some of which are introduced and tested through small case studies. Preliminary results confirm the relevance of diachronic learner corpus research and process learner corpus research and thus underline the dynamic nature of learner language, which changes and evolves in many different ways and at many different levels.

Keywords: Diachronic learner corpus research, process learner corpus research, Americanization, keystroke logging, screencasting

1 Terra cognita

Since its advent in the late 1980s, learner corpus research (LCR) can be said to have reached a relatively mature state, as testified by the fact that it now has a handbook devoted to it (Granger, Gilquin & Meunier 2015), a journal (*International Journal of Learner Corpus Research*) and an international association (*Learner Corpus Association*). Over the last thirty years or so, it has led to many advances in the study of learner language. By applying the tools and techniques of corpus linguistics to the study of learner language, it has made it possible to adopt a new perspective and investigate

Gaëtanelle Gilquin, Université catholique de Louvain, Belgium, gaetanelle.gilquin@uclouvain.be

https://doi.org/10.1515/9783110604719-004

aspects of interlanguage that had been less emphasized, and sometimes even neglected, in traditional second language acquisition (SLA) research (see also Gilquin & Granger 2015). From a qualitative point of view, unlike SLA research, whose main goal is to characterize learners' competence, LCR focuses on performance. It uses naturalistic data, not constrained data, and it has strong links with teaching. Quantitatively speaking, LCR has been able to investigate larger numbers of learners and words than is usually the case in SLA research, and to identify and quantify linguistic preferences, with concepts such as overuse and underuse. Informal surveys (e.g. Gilquin 2016, based on the review of the 81 representative studies included in Granger, Gilquin & Meunier 2015) and meta-analyses of the field (cf. Paquot & Plonsky 2017) have highlighted areas that have been relatively well explored and constitute the *terra cognita* ('known land') of LCR. Thus, in terms of variety, LCR has mainly studied learner writing, rather than learner speech, and learner English, more than any other languages (Gilquin 2016). As far as learners are concerned, it has devoted most of its attention to learners of English as a foreign language (EFL) (rather than learners of English as a second language (ESL)), advanced learners (rather than intermediate learners and beginners) and young adults studying at university (Paquot & Plonsky 2017). The preferred subject has been lexis, followed by grammar, then discourse and pragmatics (ibid.). Some of the specific linguistic phenomena that have often been investigated are the learner phrasicon, with collocations and lexical bundles, and discourse features, with connectors, discourse markers, stance markers and involvement features (Gilquin & Granger 2015). Most of the LCR studies have been of the synchronic type, examining data from learners at a single point in time, rather than looking at their development (Paquot & Plonsky 2017). And finally, methodologically speaking, LCR has mainly relied on the comparison of learner language with native language (Gilquin 2016).

The above summary of the main contributions of LCR, by listing aspects that have been given special emphasis in the field so far, also points to areas that have been underresearched, the *terra incognita* ('unknown land'), which used to be designated on old maps by means of the Latin phrase *Hic sunt dracones*, 'Here are dragons'.[1] In this chapter, I will explore two such areas and will illustrate them through the case of English. While doing so, I will put a spotlight on some studies that have gone off the beaten track to start to investigate these areas, showing what these studies have contributed and how they could be taken further. I will also present new types of resources that could be used in learner corpus research to approach the *terra incognita*. Each of the

1 In reality, the phrase itself only appears on two globes, one dating from the early 1500s and made of ostrich egg, and the Lenox globe, just a few years older, now found in the New York Public Library. Many old maps, however, include illustrations of dragons and other monsters, which are supposed to represent unexplored and presumably dangerous territories.

two main sections of this chapter corresponds to one unknown territory. Section 2 considers the possible diachronic evolution of learner language, and Section 3 examines the process of writing in a non-native language. Both areas are related to variation in time: how interlanguage, as a language variety, may change through time, from one generation of learners to the next, and how a text written by a learner gradually takes shape if we take into account the temporal dimension of the moment of writing. This exploration into some unknown areas of LCR seeks to enrich our understanding of learner language. In particular, it will underline its dynamic nature, both at a global and at an individual level, as will be explained in Section 4, the conclusion of this chapter.

2 Tempus fugit

The first territory to be explored is that of diachrony, and more precisely 'diachronic learner corpus research'. As was mentioned in Section 1, most LCR studies have been synchronic, focusing on learners at a single point in time. Laitinen (2016: 176) rightly points out, talking about English (but this is true of other languages as well), that "research on learner English has focused on interlanguage phenomena and proficiency in acquisition and not on time and diachronic processes". Although we know from diachronic linguistics that language is not static but evolves through time, the study of interlanguage has not really benefited from the findings of this field of research. When learner language change has been studied, it has been in the form of developmental patterns in pseudo-longitudinal and longitudinal studies. These two types of design are represented in Figures 1 and 2, respectively. *Pseudo-longitudinal studies* use data collected at a single point (theoretically, at least – see below) from different learners representing different proficiency levels. For example, at T1, we could collect data from a group of beginners (stage 1), a group of intermediate learners (stage 2) and one of advanced learners (stage 3), and compare these data to see how learner language develops across proficiency levels (see Thewissen 2015 for an illustration). In *longitudinal studies*, on the other hand, the data are collected from the same learners at different points in time, which correspond to different stages in their development: beginner (stage 1) at T1, intermediate (stage 2) at T2 and advanced (stage 3) at T3, for example (see Meunier & Littré 2013 for an illustration). This description of the longitudinal design gives the impression that the approach is truly diachronic, looking at the evolution of learner language through time. That this is not the case, however, appears from a comparison of Figure 2 with Figure 3, which shows a real diachronic approach, focusing on one stage (say, stage 1, that of the beginners) at different points in time. The longitudinal approach admittedly involves a time gap between the data collection points, but what the researcher is really interested in is not the temporal evolution per se (i.e. how

learner language changes from T1 to T3), but the developmental trends (i.e. how learners progress from stage 1 to stage 3). The tacit assumption, in such designs, is that if we were to collect similar data at a later point in time among a new but similar group of learners, starting the data collection with stage 1 at T2 or T3, for instance, the developmental trends would be identical.[2]

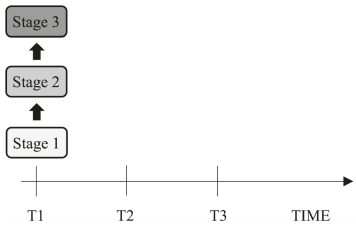

Fig. 1: Pseudo-longitudinal approach

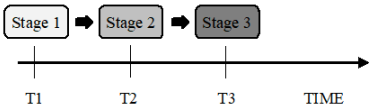

Fig. 2: Longitudinal approach

2 Thus, the natural order hypothesis, which suggests that grammatical morphemes are acquired in a certain order, has been tested for its applicability to second (as opposed to first) language acquisition, to learner corpus data (as opposed to SLA observational data) and to populations of learners from different mother tongue backgrounds, but as far as I know, it has not been tested for its validity through time. Moreover, learner-corpus-based studies that have tested the hypothesis of a natural/universal order of acquisition have tended to rely on learner corpora that were already a few years old, cf. Tono (2000), a pseudo-longitudinal study based on the Japanese English as a Foreign Language Learner Corpus (JEFLL), compiled in the mid-1990s, or Housen (2002), a longitudinal study based on the Corpus of Young Learner Interlanguage (CYLIL), whose compilation started in 1990.

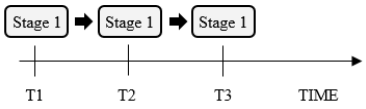

Fig. 3: Diachronic approach

This 'atemporal' assumption, valid for both longitudinal and pseudo-longitudinal studies, can be illustrated through Thewissen's (otherwise excellent) pseudo-longitudinal study of errors among B1 to C2 learners of English (Thewissen 2015). The data used in this study come from the International Corpus of Learner English (ICLE; Granger et al. 2009), and more precisely from the French (ICLE-FR), Spanish (ICLE-SP) and German (ICLE-GE) components, representing argumentative essays written by French-, Spanish- and German-speaking learners of English, respectively. Despite the pseudo-longitudinal design, which normally involves using data collected at a single point in time, the three components were compiled at different points, covering a period of potentially up to 10 years: the ICLE-FR data were collected between March 1990 and October 1993; the ICLE-GE data were collected between August 1992 and June 2000; and the ICLE-SP data were collected between April 1994 and October 1999. While ten years may not seem like a long time, it should be underlined that this period corresponded to crucial changes in the global position of the English language and in particular in the amount of exposure to the English language that learners were likely to receive, since these ten years witnessed the rise of the Internet, with the birth of software like Netscape (1994), Internet Explorer (1995), Hotmail (1996) or Google (1998). A learner of English growing up with the Internet and its large proportion of content in English can be expected to use English differently from a learner who never had access to the Internet. The developmental trends that Thewissen investigates in her study might therefore have been affected by the time differences, with some potential interaction between the developmental and temporal dimensions. Note that such interactions are also possible in truly longitudinal studies. If an event such as the advent of the Internet occurs in-between the data collection points of a longitudinal corpus, the linguistic consequences of this event might interfere with the developmental patterns, yielding results that would probably differ from those obtained with an earlier or later cohort of learners.

Crucially, the temporal dimension tends to be disregarded in other, non-developmental studies too. Thus, *intra-learner corpus comparisons* usually do not take into account the possible time differences between sections of the corpus. ICLE and its spoken counterpart LINDSEI (Louvain International Database of Spoken English Interlanguage; Gilquin, De Cock & Granger 2010) have often been used for intra-corpus comparisons, especially to compare data produced by learners from different mother tongue

backgrounds. However, the whole of LINDSEI spans 13 years, the second version of ICLE (ICLEv2; Granger et al. 2009) spans 17 years, and the next releases of these corpora, which are in preparation, will span even longer periods. In Paquot's (2013) study of lexical bundles in ICLE, for example, the comparison of ICLE-FR (collected in 1990-1993) with nine other subcorpora includes data that cover a period of up to eleven years. *Inter-learner corpus comparisons* often present the same problem, regardless of whether we compare two learner corpora or a learner corpus and a native corpus. In Gilquin (2015), the analysis of phrasal verbs by French-speaking learners in writing and speech involves a time gap between the French components of ICLE (1990-1993) and of LINDSEI (1995-1997) which were used as a basis for the comparison. And when Granger & Tyson (1996) use the Lancaster-Oslo-Bergen (LOB) corpus as a reference corpus against which to compare the use of connectors in ICLE-FR, in effect they are comparing learner language from the early 1990s and native language from 1961. Admittedly, the article makes the point that LOB is not an appropriate reference corpus. However, the reason given is not related to time, but has to do with register differences, ICLE being a corpus of essays written by university students and LOB a corpus comprising different registers of writing.

I can think of at least three reasons why time has not been considered a relevant variable in LCR so far. They can be summarized by means of the three following objections:

(i) The time spans considered in LCR are not long enough to allow for a diachronic approach.
(ii) Diachronic evolution is only found in native languages.
(iii) EFL is not a language/variety of its own and therefore cannot display any diachronic evolution.

Starting with the first objection, it cannot be denied that, in comparison with the periods usually considered in historical linguistics, the time spans that we are working with in LCR are very short, covering 10 to 15 years on average. However, next to so-called 'long-term diachrony', which investigates very long periods of time, potentially covering several centuries, we should also recognize 'short-term diachrony' (referred to as 'brachychrony' in Mair 1997), which focuses on shorter periods of just a few decades. Short-term diachronic studies such as Leech et al. (2009) have convincingly demonstrated that linguistic changes can take place over just a few years. There is therefore no reason why the time spans covered by learner corpora could not be approached from a short-term diachronic perspective. Moreover, one should not forget Virgil's adage, *Tempus fugit*: time flies in LCR, and although the field is often described as being relatively young, the idea of learner corpora has been around for close to 30 years now.

As time goes by, the gap between the first learner corpora and the most recent learner corpora will gradually increase, and soon, we will be talking about 30, 40, or even 50 years' differences. And while the notion of 'centuries old learner corpora' may almost sound like an oxymoron today, the day will come, eventually, when it corresponds to the reality of LCR.

Some people may also claim that diachronic evolution is only found in native languages and that diachrony therefore does not have any role to play in the study of non-native languages. However, diachronic linguistics has recently been applied to institutionalized second-language varieties of English (New Englishes). These varieties now have their own historical corpora, like the Historical Corpus of Singapore English or the Historical Corpus of Ghanaian English, and studies based on these, such as those found in Collins (2015), have shown that diachronic changes do occur in these varieties. In addition, a study on core and emergent modals by Laitinen (2016) suggests that English as a Lingua Franca (ELF) may also be characterized by diachronic evolution. It should however be pointed out that Laitinen's (2016) study is actually based on synchronic corpora representing spoken and written ELF, and that the diachronic conclusions are drawn on the assumption that changes first appear in speech before spreading to writing. This methodology makes it a kind of pseudo-diachronic study, whose results should be confirmed by a truly diachronic approach to ELF.

What is common to New Englishes and ELF is that they are thought to be varieties in their own right, with their own developing norms, and hence, possibly, their own diachrony. By contrast, the learner equivalent, English as a foreign language (EFL), is usually not considered a variety of its own and may therefore be said not to be affected by diachrony. Learner English is claimed to be "norm-dependent" (Kachru 1985: 17), and hence may be thought to only exist in relation to a 'norm-providing' variety. Some linguists, like Mukherjee & Götz (2015), refer to the EFL 'variants' (instead of 'varieties'), to make it clear that they are not proper varieties. Even if we do not recognize EFL as a variety in its own right, however, there are two facts that must be acknowledged. The first one is that, as a norm-dependent variety (or variant), EFL depends on models of native English (typically British or American English) which, as demonstrated by historical linguists, are subject to diachronic evolution. If learners' models vary through time, we may expect interlanguage to vary too, in accordance with the models. The second fact is that the context in which EFL learners acquire English can evolve – and has evolved over the last few years. Some twenty years ago, for example, it was difficult to get hold of English-language newspapers in Belgium. They were only to be found in specialized bookshops in large cities, and they were quite expensive. Today, English-language newspapers are easier to find in Belgium and are less expensive, but most importantly, they have become available at the click of a mouse on the Internet, most of the time for free. As suggested earlier with respect to the advent of the

Internet, a change in the acquisitional context may have linguistic consequences on EFL and therefore lead to a diachronic evolution in learner language (regardless of proficiency development).

What precedes seems to confirm the relevance of a diachronic approach to interlanguage. In order to check this, I conducted a case study comparing the influence of British English (BrE) vs American English (AmE) on EFL at two different points in time. My hypothesis was that EFL would be increasingly influenced by AmE. This hypothesis relies on the existence of two possible sources of influence. The first source is BrE, which many EFL learners use as a target norm, at least in Europe, and which has been said to show traces of Americanization (Leech et al. 2009: 252ff.). EFL learners reproducing the BrE model should thus also display a certain degree of Americanization. The second possible source of influence is more direct, since AmE, which has arguably become increasingly dominant worldwide, could influence learner English directly. Mair (2013) takes AmE to be the 'hyper-central variety' of his World System of Englishes, that is, the variety that is "a potential factor in the development of all [the] other [varieties]" (Mair 2013: 261). Among these varieties, Mair only explicitly mentions native varieties, New Englishes and ELF, but we could assume that this applies to EFL too. My analysis was based on a list of twenty pairs of items (including lexical items, syntactic constructions and phraseological patterns) taken from Algeo (2006) and displaying a significant difference between AmE and BrE as well as a distinctive preference for one variety over the other in the recent GloWbE, the Global Web-Based English Corpus (see Gilquin 2018 for details on the methodology). The frequency of these items was checked in two learner corpora: ICLEv2 (Granger et al. 2009), which is made up of essays written between 1988 and 2005, and the EFL data from the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Theodora & Korhonen 2014), which contains essays dating from a later period, namely 2011-2013. Americanness and Britishness rates were computed on the basis of the frequencies and made it possible to compare the respective influence of AmE and BrE. Table 1 shows the Americanness rate for each of the twenty pairs of items selected. The figures in bold correspond to cases where the AmE item is more common than the BrE item. There are nine such cases in ICLE and thirteen in EFCAMDAT. Among the six pairs of items that show a reversed tendency in ICLE and EFCAMDAT, five show a preference for the BrE item in ICLE and for the AmE item in EFCAMDAT (*a half/half a(n)*, *math/maths*, *MOVIE/FILM*, *APARTMENT/FLAT*, *right now/at the moment*), and in only one case, for *HAVE/TAKE a shower*, is it the opposite. In some cases, the difference between ICLE and EFCAMDAT is considerable. The pair *APARTMENT/FLAT*, for example, has an Americanness rate of 29% in ICLE and of 93% in EFCAMDAT. The pair *MOVIE/FILM* presents similar results (32% vs 92%). On average, ICLE has an Americanness rate of 49% and EFCAMDAT of 63%. These results thus seem to

confirm the growing influence of AmE on EFL, since the more recent data of EF-CAMDAT display a higher rate of American features than ICLE.

AmE/BrE pair	ICLEv2	EFCAMDAT-EFL
HAVE gotten/HAVE got	4.57%	6.19%
a half/half a(n)	47.71%	**65.54%**
math/maths	45.28%	**79.44%**
MOVIE/FILM	32.27%	**92.39%**
APARTMENT/FLAT	28.80%	**92.75%**
volunteer work/voluntary work	**55.56%**	**78.03%**
all year long/all year round	0.00%	21.74%
free time/spare time	**58.17%**	**84.28%**
right away/straight away	**100.00%**	**95.36%**
in college/at (the) college	**55.81%**	**69.60%**
right now/at the moment	34.36%	**65.38%**
in school/at school	31.71%	48.09%
on (the) WEEKEND/at (the) WEEKEND	**60.00%**	**72.83%**
toward/towards	10.89%	22.95%
CHAT with/CHAT to	**97.22%**	**96.91%**
different than/different to	42.11%	19.17%
MAKE a deal/DO a deal	**100.00%**	**96.55%**
GIVE it a try/GIVE it a go	**100.00%**	**98.44%**
TAKE a vacation/HAVE a holiday	0.00%	35.87%
TAKE a shower/HAVE a shower	**75.00%**	22.13%
AVERAGE	48.97%	63.18%

Tab. 1: Americanness rates in ICLE and EFCAMDAT (rates above 50% are shown in bold)

ICLEv2	EFCAMDAT-EFL
1988-2005	2011-2013
16 learner populations (of which 12 from Europe)	107 learner populations (of which 33 from Europe)
Higher-intermediate to advanced learners	16 levels, from complete beginners to expert users
University students	Adults of different ages

Tab. 2: Comparison of ICLEv2 and EFCAMDAT-EFL

However, it appears from Table 2 that, in addition to a difference in the period when the data were collected, ICLE and EFCAMDAT display a number of other differences. First, EFCAMDAT includes a larger number of learner populations than ICLE, corresponding to a wider geographical spread. In terms of proficiency level, there is also more variety in EFCAMDAT, which includes sixteen levels from complete beginners to expert users, than in ICLE, which includes higher-intermediate to advanced levels. Finally, while the learners in ICLE are all university students, in EFCAMDAT they are adults of different ages. Because of these differences in the design of the two corpora, we cannot exclude the possibility that the differing frequencies shown in Table 1 are not due to the time gap, but to one of these other differences. In order to avoid a possible influence of these factors, I have started collecting data of the ICLE type among French-speaking learners, a new resource that I call ICLE-FR+25, because the data are about 25 years older than the data included in ICLE-FR. It is made up of argumentative essays, on topics chosen among those listed in the ICLE manual. The learners are students at the University of Louvain, where most of the ICLE-FR data were collected. They have English as their major and are in their third year, thus presenting a profile similar to those of the learners who contributed data to ICLE-FR. In its current state, the corpus contains data from 62 students, who each contributed one or two essays, for a total of about 52,000 words. Because of this relatively small size, many of the twenty pairs of items investigated in ICLE and EFCAMDAT were not frequent enough, or simply not found at all, to make any reliable claims. The few pairs of items that were frequent enough, however, show interesting results. In the case of the *MOVIE/FILM* pair, it can be seen from Figure 4 that, while the BrE term *FILM* is the preferred option in ICLE-FR, 25 years later it is the AmE term *MOVIE* that predominates. Frequencies for two other pairs of items, *APARTMENT/FLAT* and *free time/spare time*, are low, but again they seem to point to a change over time, with a stronger influence of AmE. In ICLE-FR, only the BrE noun *FLAT* is found, whereas in ICLE-FR+25 it is only the AmE noun *APARTMENT* that is found. And while both *spare time* and *free time* are found in ICLE-FR, with no significant difference between the two, in ICLE-FR+25 only the AmE option *free time* is found. Although these results would have to be confirmed by a larger amount of data, it seems as if Americanization might be a phenomenon that affects EFL too.

It would be worth investigating whether other phenomena that have been shown to characterize the evolution of native English could also affect learner English. Figure 5 reveals for example that over the last 25 years, the colloquial adverb *maybe* has become more common in French learner English, whereas the formal adverb *perhaps* has become much less common. This suggests that the phenomenon of colloquialization, that is, "the shift to a more speech-like style" (Leech et al. 2009: 239), might have a role to play in the evolution of learner English, as is the case in British and American English

(see Leech et al. 2009). Larger diachronic learner corpora and more diachronic learner corpus research will be necessary to back up any claims about Americanization or colloquialization, but also to settle the issue of whether these phenomena directly affect learner English, whether they affect the native norm that EFL learners then reproduce, or whether, perhaps, these phenomena affect both the native norm and the EFL variety simultaneously.

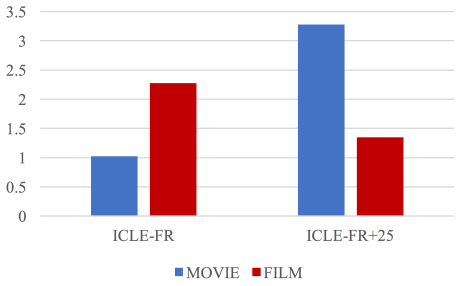

Fig. 4: Relative frequency per 10,000 words of *MOVIE* and *FILM* in ICLE-FR and ICLE-FR+25

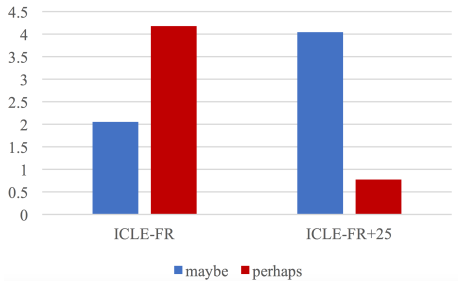

Fig. 5: Relative frequency per 10,000 words of *maybe* and *perhaps* in ICLE-FR and ICLE-FR+25

I would like to finish this section with a couple of caveats. The first one is that it would not be realistic to expect all learner corpora we are working with for non-diachronic purposes to date from exactly the same period. Collecting (learner) corpora takes time, and we have to accept that there will be time differences between some of the corpora we rely on, especially when the data are part of long-term international projects like ICLE. However, it is important, on the one hand, to be aware of such differences and to mention them as potential factors affecting the results, as would be the case for, say, proficiency level or time spent in an English-speaking country. On the other hand, one should avoid comparing corpora that present a very large time gap, for

example the LOB from the 1960s and ICLE-FR from the early 1990s, or ICLE-FR with the most recent subcorpora of the forthcoming version of ICLE (ICLEv3), the Serbian and Macedonian subcorpora, which include data from 2016. The second caveat is that not all linguistic phenomena may be equally affected by the temporal dimension. The order of acquisition of grammatical morphemes, for example (see footnote 2), might be less subject to diachronic changes than more lexical phenomena like the use of *film* vs *movie*. However, until we actually probe into the possible effect of time on learner language, we will not know for sure. The first step is therefore to investigate this issue further.

3 Mutatis mutandis

It was mentioned in Section 1 that LCR has devoted more attention to writing than to speech, and also that, unlike SLA research, it focuses on performance, rather than competence. Written performance is thus arguably one of the well-known areas of LCR. However, instead of performance *stricto sensu*, it would probably be more appropriate to say that LCR has focused on the *performed*, because it usually does not consider the writing process as such, but rather the finished product, in the form of a text. Written performance in the strict sense, on the other hand, is still largely *terra incognita* in LCR. This can be explained by the fact that we do not really have adequate resources to study the writing process. As pointed out by Wärnsby et al. (2016: 198), "the influence of corpora in the study of writing processes has been limited, partly because corpora typically contain only one version of a text". This means that, if one wants to study the process of writing in LCR and do 'process learner corpus research', new types of learner corpora are needed.

Very recently, learner corpora have started to emerge that give access to written performance (in the sense of writing process). We can distinguish between two types of corpora: corpora that include several drafts of the same text and corpora that represent handwritten texts with some traces of alteration. *Multiple-draft corpora* include the Hanken Corpus of Academic Written English for Economics (Mäkinen & Hiltunen 2016), the Malmö University-Chalmers Corpus of Academic Writing as a Process (MUCH; Wärnsby et al. 2016) and the CityU Corpus of Essay Drafts of English Language Learners (Lee at al. 2015). The Hanken Corpus is made up of academic writing by Finnish EFL learners studying economics and includes the first draft and the final version of each text. MUCH contains academic papers by Swedish EFL undergraduate to PhD students, with an average of three to five drafts per paper. The CityU Corpus includes essays by Chinese undergraduate ESL learners, with two to three drafts per essay. In addition, MUCH and the CityU Corpus include teacher comments, while

MUCH also includes peer comments. Thanks to this material, it is possible to investigate the writing process and how it can be influenced by the feedback provided. The statistics found in Lee et al. (2015) reveal for example that the total number of words and sentences increases between the first and the second draft, but decreases between the second draft and the following drafts, which suggests that improving a text does not necessarily imply writing more. It also appears that the more revisions there are, the longer the sentences become, from an average of 12 words per sentence in the first draft to an average of almost 17 words per sentence in the final drafts, which points to an increasingly large degree of complexity of the sentences. Lee et al. (2015) also present a case study on verb tenses showing, among other results, the feedback take-up rate for several combinations of tenses. When asked to change a tense to the present simple, for instance, learners responded at a rate of almost 77%, whereas they responded at a rate of less than 43% when asked to change a tense to the past perfect, which can presumably be explained by the fact that learners are more familiar with the present simple than with the past perfect. As for the improvement rate, it also varies depending on the tense. Changes to the present simple or the past simple were most of the time correct, whereas changes to the present perfect or, especially, the past perfect, mostly led to further errors.

The second type of corpus that can inform research on the writing process is a *corpus showing traces of revision in handwritten texts*. The Marburg corpus of Intermediate Learner English (MILE; Kreyer 2015) is one such corpus. It is a longitudinal corpus representing the written production of learners of English from grades 9 to 12 of German secondary schools. The texts included in the corpus were originally handwritten, and the electronic version seeks to reproduce the handwritten form as faithfully as possible, which involves text-internal mark-up to indicate line breaks, deletions or additions, for example. Surprisingly enough, the main assets of the corpus that are emphasized by Kreyer (2015) are its longitudinal nature, the fact that it includes data produced by intermediate learners of English, and its error-tagging. While these are indeed important features of the corpus, and features that are not so common in LCR, as appears from Section 1, one of the most innovative features of this corpus is the mark-up that makes visible some elements that were ultimately removed from the text. Kreyer (2015: 24) makes a very brief comment about this, noting that "the writing process itself, in particular those aspects that do not make it into the final version, provides us with valuable information about various interlanguage phenomena that are at play in the writing process". Despite the many possibilities that MILE opens up, its limitations must also be acknowledged. First, it only works with handwritten texts. Second, the data collection is very time-consuming, not only because the handwritten texts have to be keyboarded, but because all the elements of the manuscript have to be reproduced, which involves deciphering crossed out elements that are sometimes difficult or even

impossible to read. In addition to unreadable changes, there could be some invisible changes, for example through the use of an ink eraser or correction fluid/tape, which cannot be included in the corpus. Finally, the changes are not dynamic, in the sense that it is impossible to know, for example, when a specific alteration was made.

Unlike most written learner corpora, which include only the final product, corpora such as MILE or the multiple-draft corpora mentioned above offer a glimpse into the actual writing process by giving access to some of the steps involved in producing a written text. The difference between these new types of written learner corpora and the traditional corpora is illustrated in Figure 6, where the arrow symbolizes the process of writing, taking place along a temporal dimension. In the first arrow, which corresponds to traditional written corpora, the process is represented only by its very final step, namely the text as a finished product. The second arrow shows the different components of a multiple-draft corpus: one or several drafts and the final text. In the case of MILE, the draft and the text could be said to correspond to different layers of annotation in the corpus (draft = mark-up; text = transcript). With such corpora, we come closer to the writing process, since different steps are represented, but the gaps between the components indicate that the corpora fail to capture the whole process. Both traditional and multiple-draft corpora, therefore, only present a partial view of the writing process. What we need is a way of capturing the process as a whole, as illustrated by the last arrow in Figure 6.

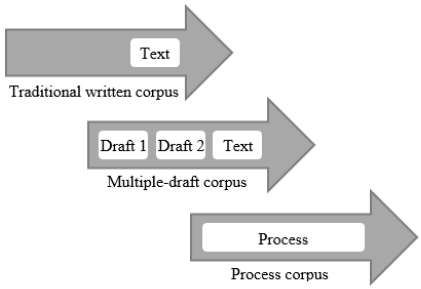

Fig. 6: From written text to writing process

Capturing the whole writing process is precisely the aim of the Process Corpus of English in Education (PROCEED),[3] a new type of resource that should make it possible to

3 See [https://uclouvain.be/en/research-institutes/ilc/cecl/proceed.html] (last accessed on 30 December 2018).

take the corpus-based study of learner writing one step further. PROCEED is what I would like to call a 'process learner corpus', i.e. a learner corpus reproducing the whole of the writing process, not just the final product or some of the steps involved in the process. This is achieved by means of keylogging and screencasting while learners are typing their texts on a computer. *Keylogging* (or keystroke logging) is the recording of any action on a computer keyboard. The recording was made by means of Inputlog (Leijten & Van Waes 2013), which shows the keys that have been struck on the keyboard and also provides very useful statistics about the writing process (e.g. number of revisions or pausing behaviour). *Screencasting* is the recording of anything that happens on a computer screen. OBS (Open Broadcaster Software) was used to record the screen activity. The output is a video that shows learners' behaviour during the writing task, including what they type, what changes they make, but also, say, what online tools they use when writing a text. The compilation of PROCEED started in February 2017, among (mostly) French-speaking learners of English, all of them university students majoring in English, who had to type a short argumentative essay on one of three set topics. The data collection took the form of an obligatory in-class assignment, but the students could choose whether they wanted to have their computer activity recorded, and whether they were willing to share their data for research purposes. There are three types of output: the final product, that is the learners' texts in Word format, which currently amount to some 60,000 words; the screencasts, which correspond to about 130 hours of video; and the keystroke logging data, which represent 160 log files. All of these data complement each other to reveal the process through which texts are written. Keystroke logging and screencasting have already been used to study the L2 writing process (e.g. Miller, Lindgren & Sullivan 2008 or Nie 2014 for keylogging and Luoma & Tarnanen 2003 or Séror 2013 for screencasting) and they have provided very interesting insights into the processes underlying L2 writing. However, such studies tend to be characterized by a small number of participants (from only one or two learners to some twenty at most). They are usually of one of two types: either they look at global revision statistics, for example pausing behaviour during writing, or they adopt a kind of discourse analysis approach, focusing on one individual writer and manually dissecting his or her writing behaviour. In these studies, keystroke logging or screencasting is often used in combination with other, more intrusive techniques like think-aloud protocols or eye tracking. Finally, these studies are embedded within frameworks like SLA research or psycholinguistics, rather than corpus linguistics (it is symptomatic, in this respect, that the term 'corpus' almost never occurs in these publications). However, if enough data are collected and if the format of the data allows for (semi-)automatic searches, then it becomes possible to apply the tools and techniques of corpus linguistics to the data and to make generalizations about the L2 writing process.

In comparison with traditional or multiple-draft learner corpora, one of the main benefits of PROCEED is that it provides information about the textual changes made during the writing process (deletions or additions, for example) as well as when and how these changes are made. It also gives information about the structuring of the text. We can notice, for example, that writing is not the linear process that one may imagine. Learners tend to move sentences from one paragraph to another, go back to an earlier part of the text in order to make alterations, etc. The videos make it possible to see whether and how learners use online resources such as Google, dictionaries, corpus interfaces or language forums to help them in the writing process. Preliminary observations are quite disappointing in this respect, since students appear to use mainly bilingual dictionaries and synonym dictionaries. They do not use corpus interfaces at all, although they were shown how to use some of them, and information about collocations is usually not looked for either. The screencasts also highlight cases where an error is triggered by the online resource itself – either because the information it contains is inappropriate, or because the learner has used the resource inappropriately. Interestingly, by giving access to performance in the strict sense (see beginning of this section), process data allow insights into learners' competence. They can for example help understand why certain errors were made, point to possible cases of avoidance, and more generally reveal some of the cognitive processes that are at work during the production of a written text. Thus, the sentence *Dreaming and imagination are essential cognitive skills which need**s** to be developed* in one of the PROCEED texts might at first sight give the impression that the learner does not master the rules of verb agreement, since she used the *-s* ending with a plural subject. However, the screencast tells a different story. In the video, we see that the original sentence was in the singular, with *imagination* as a subject. The learner then decided to add *dreaming* as a subject. She adapted the form of the main verb (*is* was changed to *are*) and she pluralized the noun *skill*, but she failed to adapt the verb of the subclause. Given the circumstances, an oversight is a more likely explanation than a problem of competence. In other words, to use the well-known distinction between errors and mistakes (see Corder 1967), this is more likely a mistake (an error of performance) than an error in the strict sense (an error of competence). As another illustration, consider the following sentence: *feminism is an ideology that can't be generalized*. The use of the contraction *can't* could be described as a spoken-like feature and taken as evidence for the learner's lack of register-sensitivity (cf. Gilquin & Paquot 2008). However, it appears from the screencast that the learner initially spelled the negative modal verb as two words: *can not*. She then looked for help in an online dictionary and found two forms listed (with no explanation): *cannot* and *can not*. The fact that in the video we see the learner highlight *can not* twice with the cursor suggests that she hesitated as to which form she should use. In the end, she opted for *can't*, probably because it allowed her not to choose between the two

spellings of *cannot*. This change thus points to a kind of avoidance strategy, rather than a preference for spoken-like features. In fact, the initial sentence included a personal subject (*feminism is an ideology that we can not generalize*), and at the same time as the learner substituted *can't* for *cannot*, she also replaced the personal pronoun *we* and the active voice with an impersonal passive voice, thus making the clause more written-like. Other videos illustrate the replacement of spoken-like features by features that are more typical of academic writing, which confirms a point made by Ädel (2008) on the basis of a comparison of timed and untimed essays, namely that learners need time to make their texts less involved, and hence more formal and written-like. These few examples show that the PROCEED corpus, and the screencasts in particular, can bring to light certain aspects that would otherwise have gone unnoticed in a traditional, product-oriented corpus.

Before turning to the conclusion of this chapter, a few general remarks concerning the process-oriented approach in LCR are in order. The first one has to do with the nature of process learner data. While some people may consider screencasting and keyboard logging as a constrained type of data and associate it with a more experimental kind of approach, as also suggested by the fact that these techniques have been used in SLA research and psycholinguistics, I would like to argue that process data can be recognized, *mutatis mutandis* (i.e. once the necessary changes have been made), as corpus data. Writing a text on a computer is an ecologically valid activity for students – almost the default option, as it turns out, since a British survey conducted in 2014 revealed that "one in three respondents had not written anything by hand in the previous six months".[4] In addition, neither screencasting nor keylogging is really intrusive. Once the two programs have been launched, the learner does not notice them and can just type his or her text as usual. These techniques can even be said to be much less intrusive than the audio recorder and microphone used to collect spoken corpus data. The second remark is that, although a process learner corpus offers a better view of the writing process than a more traditional written learner corpus, it only offers a partial view of the process, because there will always be things that the learner does or thinks that do not lead to any activity on the screen or on the keyboard. In order to access these aspects, more experimental (and intrusive) techniques would be necessary, such as think-aloud protocols, eye tracking or electroencephalography (EEG).[5] These, however, would affect the naturalness of the task and thus question the very notion of corpus.

4 See [https://www.theguardian.com/science/2014/dec/16/cognitive-benefits-handwriting-decline-typing] (last accessed on 13 May 2017).

5 Interestingly, the cursor sometimes gives an indication of the point of gaze, much like eye tracking would do.

4 Audeamus

In this chapter, we have dealt with two areas in LCR that are still largely unexplored: the possible diachronic evolution of learner language (to be distinguished from its evolution through proficiency levels) and the process underlying L2 writing (i.e. written performance in the strict sense). The preliminary exploration of these two areas suggests that learner language is characterized by variation in time. Learner English, as a variety (or variant), is not static, but changes as a result of the typical acquisitional context (cf., for instance, the impact of the Internet), the norms it depends on (which are themselves subject to diachronic evolution), and more generally the dominant position of English worldwide. The English produced by learners a few decades ago is thus likely to differ from present-day learner English, which in comparison might for example show more traces of Americanization or colloquialization. Such changes are expected to be global, affecting populations of learners across the board (although differences between learners are bound to exist too, with some learners being more affected by global changes than others). By contrast, the second type of change highlighted in this chapter is more individual and its timeline is shorter. If we consider the process through which a learner writes a text, we notice that it is not a linear process in which words are fixed once and for all. On the contrary, many changes take place at different levels while the text is taking shape: paragraphs are moved around, sentences are reorganized, words are deleted or replaced by other words, etc. Of course, this is by no means a novel finding, but thanks to techniques such as screencasting and keylogging, we can, for the first time, follow the process step by step and see how it affects the final product.

Since its advent some thirty years ago, the field of LCR has opened many doors and explored many areas. Yet, several doors remain to be opened and several areas remain to be explored. What this chapter has shown is that learner language is dynamic – it changes, it evolves through time, it is not a static entity set in stone. The field of LCR therefore needs to be dynamic too, and be open to new ideas, new experiences, new adventures. More generally – and this is true of all areas of linguistics, and probably of research at large – we should venture off the beaten path, take the roads less travelled, rather than go with the crowd, because only by exploring unknown territories can we truly expand our knowledge. So let us explore, let us not fear the dragons, because new discoveries await! *Audeamus* (‘let us dare’)!

References

Ädel, Annelie. 2008. Involvement features in writing: Do time and interaction trump register awareness? In Gaëtanelle Gilquin, Szilvia Papp & María Belén Díez-Bedmar (eds.), *Linking up Contrastive and Learner Corpus Research*, 35-53. Amsterdam/Atlanta: Rodopi.

Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: Cambridge University Press.

Collins, Peter (ed.) 2015. *Grammatical Change in English World-Wide*. Amsterdam/Philadelphia: John Benjamins.

Corder, S. P. 1967. The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching (IRAL)* 5 (4): 161-170.

Geertzen, Jeroen, Alexopoulou, Theodora & Korhonen, Anna. 2014. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In Ryan T. Miller, Katherine I. Martin, Chelsea M. Eddington, Ashlie Henery, Nausica Marcos Miguel, Alison M. Tseng, Alba Tuninetti & Daniel Walter (eds.), *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*, 240-254. Somerville, MA: Cascadilla Proceedings Project.

Gilquin, Gaëtanelle. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11 (1): 51-88.

Gilquin, Gaëtanelle. 2016. One norm to rule them all? Describing and evaluating learners' usages in learner corpus research. Plenary talk at the 12th international Teaching and Language Corpora conference (TaLC 12), Justus-Liebig-Universität, Giessen.

Gilquin, Gaëtanelle. 2018. American and/or British influence on L2 Englishes – Does context tip the scale(s)? In Sandra C. Deshors (ed.), *Modeling World Englishes: Assessing the Interplay of Emancipation and Globalization of ESL Varieties*, 187-216. Amsterdam/Philadelphia: John Benjamins.

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

Gilquin, Gaëtanelle & Granger, Sylviane. 2015. Learner language. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 418-435. Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle & Paquot, Magali. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1 (1): 41-61.

Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny & Paquot, Magali. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds.) 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Granger, Sylviane & Tyson, Stephanie. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15 (1): 17-27.

Housen, Alex. 2002. A corpus-based study of the L2-acquisition of the English verb system. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 77-116. Amsterdam/Philadelphia: John Benjamins.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & H. G. Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*, 11-30. Cambridge: Cambridge University Press.

Kreyer, Rolf. 2015. The Marburg Corpus of Intermediate Learner English (MILE). In Marcus Callies & Sandra Götz (eds.), *Learner Corpora in Language Testing and Assessment*. 13-34. Amsterdam/Philadelphia: John Benjamins.

Laitinen, Mikko. 2016. Ongoing changes in English modals: On the developments in ELF. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *New Approaches to English Linguistics: Building Bridges*, 175-196. Amsterdam/Philadelphia: John Benjamins.

Lee, John, Yan Yeung, Chak, Zeldes, Amir, Reznicek, Marc, Lüdeling, Anke & Webster, Jonathan. 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.

Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Smith, Nicholas. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Leijten, Mariëlle & Van Waes, Luuk. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30 (3): 358-392.

Luoma, Sari & Tarnanen, Mirja. 2003. Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing* 20: 440-465.

Mair, Christian. 1997. Parallel corpora: A real-time approach to the study of language change in progress. In Magnus Ljung (ed.), *Corpus-based Studies in English*, 195-209. Amsterdam/Atlanta, GA: Rodopi.

Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34 (3): 253-278.

Mäkinen, Martti & Hiltunen, Turo. 2016. Creating a corpus of student writing in economics: Structure and representativeness. In María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo & Ignacio M. Palacios-Martínez (eds.), *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora*, 41-58. Leiden/Boston: Brill.

Meunier, Fanny and Littré, Damien. 2013. Tracking learners' progress: Adopting a dual "corpus cum experimental data" approach. *The Modern Language Journal* 97 (S1): 61-76.

Miller, Kristyan S., Lindgren, Eva & Sullivan, Kirk P. H. 2008. The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly* 42 (3): 433-454.

Mukherjee, Joybrato & Götz, Sandra. 2015. Learner corpora and learning context. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds.), *The Cambridge Handbook of Learner Corpus Research*, 423-442. Cambridge: Cambridge University Press.

Nie, Yujing. 2014. Analysis of EFL learners' writing process in China: Comparison between English major and non-English major learners. *Studies in Literature and Language* 9 (1): 72-76.

Paquot, Magali. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18 (3): 391-417.

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

Séror, Jérémie. 2013. Screen capture technology: A digital window into students' writing processes. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie* 39 (3): 1-16.

Thewissen, Jennifer. 2015. *Accuracy across Proficiency Levels. A Learner Corpus Approach*. Louvain: Presses universitaires de Louvain.

Tono, Yukio. 2000. A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In Lou Burnard & Tony McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective. Papers from the Third International Conference on Teaching and Language Corpora*, 123-132. Frankfurt am Main: Peter Lang.

Wärnsby, Anna, Kauppinen, Asko, Eriksson, Andreas, Wiktorsson, Maria, Bick, Eckhard & Olsson, Leif-Jöran. 2016. Building interdisciplinary bridges. MUCH: The Malmö University-Chalmers Corpus of Academic Writing as a Process. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *New Approaches to English Linguistics: Building Bridges*, 197-211. Amsterdam/Philadelphia: John Benjamins.