Gilquin, G. 2020. 'Learner corpora'. In M. Paquot & S. Gries (eds) *A Practical Handbook of Corpus Linguistics* (pp. 283-303). Cham: Springer.

Learner Corpora

Gaëtanelle Gilquin
Université catholique de Louvain

Abstract This chapter deals with learner corpora, that is, collections of (spoken and/or written) texts produced by learners of a language. It describes their main characteristics, with particular emphasis on those that are distinctive of learner corpora. Special types of corpora are introduced, such as longitudinal learner corpora or local learner corpora. The issues of the metadata accompanying learner corpora and the annotation of learner corpora are also discussed, and the challenges they involve are highlighted. Several methods of analysis designed to deal with learner corpora are presented, including Contrastive Interlanguage Analysis, Computer-aided Error Analysis and the Integrated Contrastive Model. The development of the field of learner corpus research is sketched, and possible future directions are examined, in terms of the size of learner corpora, their diversity, or the techniques of compilation and analysis. The chapter also features representative corpus-based studies of learner language, representative learner corpora, tools and resources related to learner corpora, and annotated references for further reading.

Keywords: second language acquisition, foreign language learning, interlanguage, longitudinal corpus, local learner corpus, metadata, annotation, Contrastive Interlanguage Analysis, Computer-aided Error Analysis, Integrated Contrastive Model

1 Introduction

Learner corpora are corpora representing written and/or spoken 'interlanguage', that is, language produced by learners of that language. Typically, the term covers both foreign language and second language situations, that is, respectively, situations in which the target language has no official function in the country and is essentially confined to the classroom (and, possibly, international communication), and situations in which the target language is learned by immigrants in a country where it is the dominant native language. It is normally not used to refer to corpora of child language, which are made up of data produced by children acquiring their first language (see Chap. 14), nor corpora of institutionalized second-language varieties, which are collected in countries that have the target language as an official, though not native, language (cf. 'New Englishes' like those represented in the International Corpus of English), although their data may also reflect a process of learning or acquisition.

While the first corpora were compiled in the 1960s, it took some thirty years before the first learner corpora started to be collected, both in the academic world (International Corpus of Learner English (ICLE)) and in the publishing world (Longman Learners' Corpus). Initially, they were corpora of written learner English, keyboarded from handwritten texts. Gradually, however, learner corpora representing other languages as well as spoken learner corpora made their appearance, while written learner corpora were increasingly compiled directly from electronic sources, which facilitated the compilation process. The nature of learner language made it necessary to rethink and adapt some of the general principles of corpus data collection and analysis. This led, among other things, to the creation of new types of corpora, like longitudinal corpora representing different stages in the language learning process, to the collection of new types of metadata, such as information about the learner's

mother tongue and exposure to the target language, and to the use of new methods to annotate or query the corpus, for example to deal with the errors found in learner corpora. These specificities, and others, will be considered in Sect. 2.

2 Fundamentals

2.1 Types of learner corpora

Like other corpora, learner corpora can include written, spoken and/or multimodal data; they can be small or large; and they can represent any (combination of) languages. The 'Learner Corpora around the World' resource (see Sect. 4) reveals that the majority of learner corpora are made up of written data, and that these data often correspond to learner English. Other types of corpora, however, including spoken learner corpora and corpora representing other target languages, are becoming more widely available. As for size, many of the learner corpora listed in the 'Learner Corpora around the World' resource are under one million words, with some of them not even reaching 100,000 words and a couple just containing some 10,000 words. It is likely that among those learner corpora that are not listed but exist 'out there', most can be counted in tens of thousands rather than in millions of words. Yet, there are also learner corpora that are much larger, especially those that have continued to grow over the years (like the Longman Learners' Corpus, which now comprises 10 million words) and those that come out of the testing/assessment world, such as EFCAMDAT (Geertzen et al. 2014) or TOEFL 11 (Blanchard et al. 2013).

One of the defining features of corpora is that they should be made up of authentic texts. This concept of authenticity, however, tends to be problematic in the case of learner corpora. Learner language, most of the time, is not produced purely for communicative purposes, but as part of some pedagogical activity, to practise one's language skills. Writing an argumentative essay or role-playing with a classmate, for example, may be natural tasks in the classroom, but they are not authentic in the sense of being "gathered from the genuine communications of people going about their normal business" (Sinclair 1996). Our understanding of the concept of authenticity must therefore be adapted to the context of learner corpora and encompass tasks that would not be described as natural in other contexts. It must also be acknowledged that some learner corpora will be more "peripheral" (Nesselhauf 2004: 128), as is the case of spoken learner corpora like the Giessen-Long Beach Chaplin Corpus (Jucker et al. 2003) which are elicited on the basis of a picture or a movie and thus include data of a more constrained nature. Another, related feature of learner language is that it usually does not cover the whole spectrum of genres that is characteristic of native varieties. Because its use tends to be associated with educational settings, there are certain genres that are difficult to capture or simply do not exist in the target language. Having a spontaneous conversation with a friend, for example, is more likely to occur in the mother tongue (L1) than in the target language (L2). As a result, most learner corpora represent one of a limited number of genres, including argumentative essays, academic writing, narratives and interviews.

One type of learner corpus that is worth singling out, because it is specific to varieties that are in the process of being learned or acquired (including child language), is the longitudinal learner corpus. In such a corpus, data are collected from the same subjects at different time intervals, so as to reflect the development of their language skills over time. Belz & Vyatkina (2005), for example, use longitudinal data from the Telecollaborative Learner Corpus of English and German (Telekorp) to study the development of German modal particles over a period of nine weeks, with one data collection point every week. Most longitudinal learner corpora, however, are less 'dense', in that they include data collected at

longer intervals, sometimes only once or twice a year (cf. LONGDALE, the Longitudinal Database of Learner English; Meunier 2016). Note that non-longitudinal learner corpora can sometimes also be used to investigate the development of learner language. Thus, if a learner corpus contains data produced by distinct learners from different proficiency levels, like the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) Corpus (Izumi et al. 2004), it is possible to identify developmental patterns by comparing subcorpora representing different levels, even if all the data were collected at a single point in time. Such learner corpora are called 'quasi-longitudinal' corpora and, because they are easier to collect than longitudinal corpora, they have often been used to study interlanguage development.

2.2 Metadata

Given the "inherent heterogeneity of learner output" (Granger 1998: 177), it is crucial that information about the data included in a learner corpus should be available. Learner corpora tend to be characterized by a large amount of such metadata. These metadata can have to do with the text itself (genre, length, conditions in which the task took place, etc.), but they can also concern the learners: what is their mother tongue? how old are they? how long have they been learning the target language? what kind of exposure to the target language have they received? do they know any other languages? etc. Usually, some of these variables are controlled for in the very design of the corpus, in the sense that the corpus only includes data corresponding to a certain value, e.g. only written essays (in ICLE) or only native speakers of English learning Spanish (in the Spanish Learner Language Oral Corpora (SPLLOC)). For the variables that are not controlled for during the compilation of the corpus, it is often possible for users to find information that enables them either to use a subset of the data meeting specific criteria (e.g. only texts written in exam conditions) or to examine the distribution of the results according to these variables (e.g. percentage of a given linguistic phenomenon among male vs female learners). Using the Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context (MERLIN),[1] for example, one can select a number of criteria, like the task (essay, email, picture description, etc.), the learner's mother tongue, his/her age, gender or proficiency level according to the Common European Framework of Reference for Languages (CEFR), in order to define a subcorpus and then restrict the search to this subcorpus. Figure 1 is a screenshot from the MERLIN website that shows the selection of a subcorpus made up of data produced by French-speaking learners of Italian with an A2 CEFR level (test and overall rating) and aged between 30 and 59. The ICLE interface (Granger et al. 2009) also allows users to define a subcorpus according to certain criteria. In addition, it makes it possible to visualize, in the form of tables and graphs, the distribution of the results according to all the other variables encoded in the metadata. Figure 2 is a screenshot from the ICLE interface that represents the output of a search for the word *informations* in the ICLE data produced by learners with Chinese (or Chinese-Cantonese/Chinese-Mandarin) as their mother tongue (ICLE-CH). More particularly, the graph shows the distribution of the results according to the time available to write the essay and indicates that the incorrect pluralization of *information* is more frequent in timed than in untimed essays.

[1] http://merlin-platform.eu/

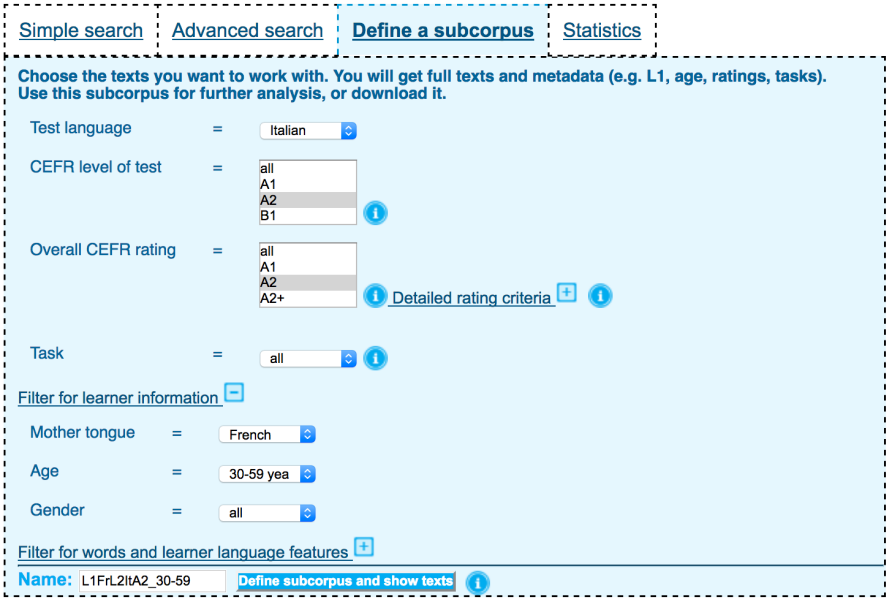

Fig. 1 Selection of a subcorpus on the MERLIN platform (criteria: target language = Italian; mother tongue = French; CEFR level of test = A2; overall CEFR rating = A2; age = 30-59) (source: http://merlin-platform.eu/)

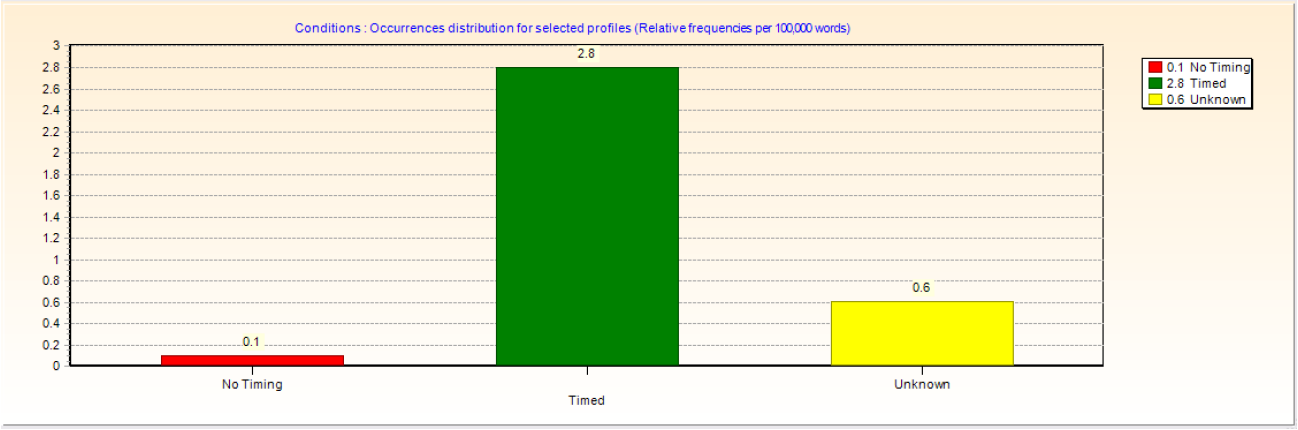

Fig. 2 Relative frequency of *informations* in ICLE-CH according to time available (source: Granger et al. 2009)

Despite the wealth of metadata that accompany most learner corpora and despite the facilities that some of these corpora provide to access them, it must be recognized that metadata are not used to their full potential in learner corpus research. One variable that is regularly taken into account is that of the learner's L1 background (e.g. Golden et al. 2017, based on ASK, the Norsk andrespråkskorpus), which makes it possible to identify probable cases of transfer from the L1. Sometimes it is another variable that is investigated, for example exposure to the target language through a stay abroad (Gilquin 2016) or presence of a native or non-native interlocutor (Crossley & McNamara 2012). Studies that examine the possible impact of several variables, on the other hand, are relatively rare, although such studies can offer important insights into the factors that are likely to affect learner language. The problem with this type of approach is that, because of the relatively limited size of most learner corpora, selecting many variables may result in a very small subset of data (see Callies 2015: 52), which, in effect, may make any kind of generalization impossible.

2.3 Annotation

Learner corpora can be enriched by means of the same types of annotation as all other corpora, including part-of-speech (POS) tagging, parsing, semantic annotation, pragmatic annotation and, for spoken learner corpora, phonetic and prosodic annotation (see Chap. 2 and Chap. 11). One issue to bear in mind, however, is that, with very few exceptions, the tools that one has to rely on to annotate learner corpora automatically are tools that have been designed to deal with native data. Applying them to non-native data may therefore cause certain difficulties. For POS tagging, for example, the many spelling errors found in written learner corpora have been shown to lower the accuracy of POS taggers (de Haan 2000, Van Rooy & Schäfer 2002). As for parsing, punctuation and spelling errors in written learner corpora have the highest impact according to Huang et al. (2018), and in spoken learner corpora Caines & Buttery (2014) have demonstrated that disfluencies and (formal and idiomatic) errors can lead to a 25% decrease in the success rate of the parser. However, while tools and formats of annotation specifically designed for learner data would of course be desirable (as suggested by Díaz-Negrillo et al. (2010) for POS tagging), it must be underlined that some attempts to automatically annotate learner corpora with off-the-shelf tools have been quite successful. Granger et al. (2009: 16), for example, report accuracy rates between 95% and 99.1% for the POS tagging of ICLE. A first attempt at POS tagging the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010) revealed an accuracy rate of 92% (Gilquin 2017). As for parsing, it seems to be more affected by the nature of learner language than POS tagging (see Huang et al. 2018). However, Geertzen et al. (2014: 247) note that the parser they used actually scored slightly better on EFCAMDAT, a written learner corpus, than on the Wall Street Journal corpus (89-92% for EFCAMDAT, to be compared with 84-87% for the Wall Street Journal). These reasonably good accuracy rates – given the non-native nature of the corpora – may be explained by the fact that the errors and disfluencies found in learner language are compensated by the relatively simple structure of the sentences which learners tend to produce (see Meunier (1998: 21) on POS tagging and Huang et al. (2018) on parsing). Another possible explanation is that most learner corpora represent university-level interlanguage (like ICLE and LINDSEI) and that such data are arguably easier to deal with for a POS tagger or parser than data produced at a lower proficiency level. Geertzen et al. (2014: 248) point out that the accuracy rate of the parser was higher on the more advanced EFCAMDAT data, although "the effect seem[ed] small". Next to these automated methods of annotation, learner corpora can also be annotated manually. While a full annotation of the corpus may not be feasible (nor, in fact, desirable), one type of annotation that may be particularly useful is problem-oriented tagging (de Haan 1984). This tagging is geared towards a specific research question and consists in annotating only those items that are of direct relevance to the research question. Spoelman's (2013) study of partitive case-marked noun phrases in learner Finnish, for instance, involved tagging instances of this phenomenon, depending on the category they represented. Such tagging then opens the way to automatic treatment of the annotated corpus.

Besides these types of annotation that are common to all corpora, there is one that is typical of learner corpora (and also child-language corpora, see Chap. 14), namely error tagging, which consists in the annotation of the errors found in a corpus (syntactic errors, unusual collocations, mispronunciations, etc.). The Fehlerannotiertes Lernerkorpus ('error annotated learner corpus', Falko), for instance, is an error-tagged corpus of learner German. The annotation of errors is usually accompanied by a correction (the 'target hypothesis') as well as a tag indicating the category of the error (e.g. spelling error, error in verb morphology, complementation error). Figure 3 shows an error-tagged sentence from Falko,

as retrieved from the ANNIS platform.[2] Falko uses a multi-layer standoff architecture, in which each layer represents an independent level of annotation (see also Chap. 3). The 'tok' (= token) layer shows the original sentence as produced by the learner. 'ZH1' provides a corrected version of the sentence (ZH = Zielhypothese 'target hypothesis'), with 'ZH1Diff' highlighting the differences between the original and corrected versions, and 'ZH1lemma' and 'ZH1pos' corresponding, respectively, to a lemmatized and POS-tagged version of the sentence. In this case, the learner has mistakenly used the article ('ART') *der* instead of the correct form *die*, an error which involves a changed token ('CHA') in the target hypothesis. Note that the multi-layer architecture of the corpus allows for enough flexibility to encode competing target hypotheses (Reznicek et al. 2013). In Falko, the step of attributing an 'edit tag' to the error (change, insertion, deletion, etc.) can be automated by comparing the learner text and the (manually encoded) target hypothesis/hypotheses. In learner corpus research, attempts have also been made to automate the process of error detection itself, although this is usually restricted to specific phenomena, for example preposition errors (De Felice & Pulman 2009), article errors (Rozovskaya & Roth 2010) or spelling errors (Rayson & Baron 2011). Most of the time, however, the whole error tagging procedure is done manually, a time-consuming task that can be facilitated by the use of an error editor like the Université Catholique de Louvain Error Editor (UCLEE; see Dagneaux et al. 1998 and Sect. 4). Once a learner corpus has been error tagged, it becomes possible to automatically extract instances of erroneous usage, which, as will be described in the next section, lies at the basis of one of the methods of analysis that have been developed to deal with learner corpora.

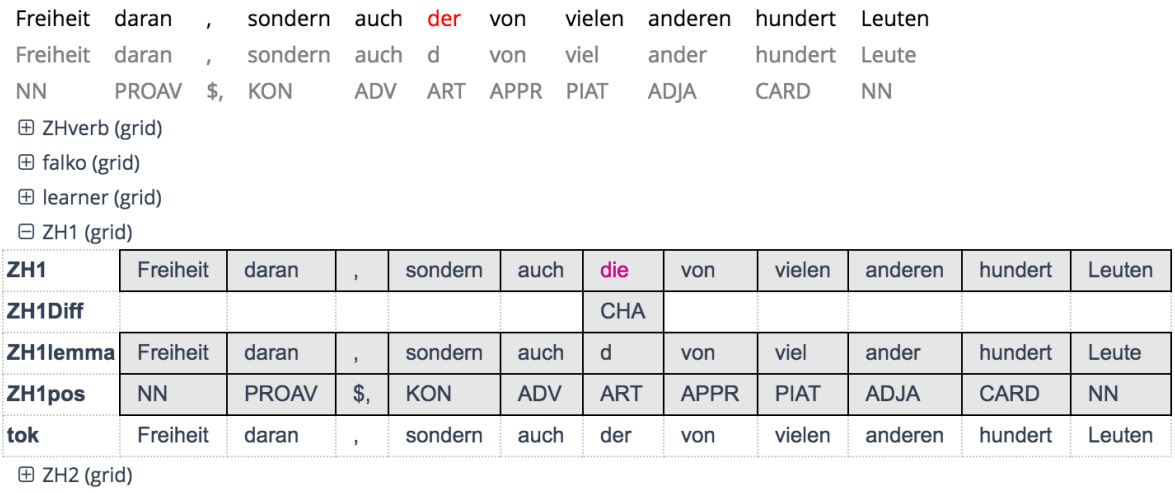

Fig. 3 Example of an error-tagged sentence in Falko (FalkoEssayL1v2.0: dhw015_2007_06) as displayed on the ANNIS platform

2.4 Methods of analysis

In addition to the application of well-established corpus linguistic methods, like the use of concordances (Chap. 8), frequency lists (Chap. 4) or collocations (Chap. 7), a number of techniques have been developed to deal specifically with learner corpora. Among these, we can mention Computer-aided Error Analysis (Dagneaux et al. 1998), Contrastive Interlanguage Analysis (Granger 1996) and the Integrated Contrastive Model (Granger 1996, Gilquin 2000/2001). Computer-aided Error Analysis (or CEA) relies on the use of an error-tagged learner corpus (cf. Sect. 2.3). Through error tagging, errors are identified and

[2] https://korpling.german.hu-berlin.de/falko-suche/

categorised according to a taxonomy, such as that developed by Dagneaux et al. (2008) to error tag ICLE. These error tagging systems are usually hierarchical, distinguishing for example between grammar, lexis, lexico-grammar and style at a high level of annotation, and then making further distinctions within each of these categories, for example grammatical errors related to nouns, pronouns or verbs, and within grammatical verb errors, those having to do with number, tense, voice, etc. This hierarchy is reflected in Dagneaux et al.'s (2008) tagset: grammatical errors are indicated by the letter 'G', grammatical verb errors by 'GV', and grammatical errors in verb tense by 'GVT'. Such tags make it very easy to automatically retrieve all the annotated errors in a certain category (e.g. all the complementation errors) or all the occurrences of a word representing a certain type of error (e.g. all the cases where the verb *enjoy* is used with an erroneous complement). These errors are the focus of analysis of CEA, as was the case in traditional error analysis (see James 1998). Unlike traditional error analysis, however, CEA allows the linguist to examine the errors in context, to consider correct uses along with incorrect uses, and to easily quantify the results (percentage of incorrect uses out of all uses or relative frequency of the error per 10,000 words, for instance).

Contrastive Interlanguage Analysis (CIA) consists of two types of comparison: a comparison of learner language with native language and a comparison between different learner varieties (Granger 2009: 18). These two types of comparison should preferably be combined with each other, but they can also be drawn separately. The comparison between native and learner language lies at the basis of a majority of the studies in learner corpus research (Flowerdew 2015: 469). Such a comparison helps identify non-standard forms (cf. CEA), but also, importantly, instances of 'overuse' and 'underuse' (or 'overrepresentation' and 'underrepresentation', see Granger 2015: 19). These terms, which are not meant as being evaluative but purely descriptive, refer to cases in which a given linguistic phenomenon (word, construction, function, etc.) is used significantly more or significantly less in the learner corpus than in a comparable native corpus, as indicated by a measure of statistical significance. The study of over- and underuse has been a real eye-opener in learner corpus research, because it has shown that the foreign-soundingness of learner language, especially at advanced levels of proficiency, is to be attributed as much (or perhaps even more) to differences in the frequency of use as to downright errors (Granger 2004: 132). The second type of comparison in CIA involves comparing different learner varieties, most notably varieties produced by learners from different L1 backgrounds. Such a comparison helps detect possible traces of transfer from the mother tongue: if a feature is only found among learners from a specific L1 population, say Italian learners of French, it might be a sign that it is the result of crosslinguistic influence, that is, interference from the L1 (Italian) on the L2 (French) (see Jarvis & Pavlenko 2008 on crosslinguistic influence, and Osborne 2015 on its link with learner corpus research). The learner varieties that are compared with each other could however differ along another dimension, which could be any of the variables encoded in the corpus metadata (comparison of foreign and second language learners, of male and female learners, of learners who have spent some or no time in a target language country, etc.). Recently, a revised version of CIA, called CIA^2, has been proposed by Granger (2015). Among its major developments, we can mention the fact that this revised model no longer advocates the exclusive use of native language as a reference point against which to compare learner varieties. Instead, it promotes the comparison of "interlanguage varieties" against one or several "reference language varieties" which, in the case of English, could include, in addition to native English, New Englishes (like Hong Kong English or Singapore English) and English as a Lingua Franca (i.e. English as used by competent L2 users). CIA^2 also includes an explicit reference to a number of variables (diatypic, dialectal, task and learner

variables), thus encouraging researchers to take these into account in the application of the model.

The Integrated Contrastive Model (ICM) is partly based on CIA, but it also integrates a contrastive analysis (CA), comparing the target language and the mother tongue thanks to comparable or parallel corpora (cf. Chap. 12). The model aims to predict possible cases of transfer (when the CA shows the target language and the mother tongue to differ in a certain respect) and seeks to explain problematic uses – misuse, overuse, underuse – in the learner corpus (by checking whether they could be due to discrepancies between the target language and the mother tongue). It thus has both predictive and diagnostic power. By combining careful analyses of learner, native and bilingual corpora, the model avoids the trap of misattributing certain phenomena to transfer simply because intuition seems to suggest that this is a plausible interpretation. Liu & Shaw (2001: 179), for example, claim that the frequent use of the causative constructions *make sb/sth feel* and *make sb/sth become* by Chinese learners of English "may be attributable to L1 interference" because such sequences "have word for word translational equivalents in Chinese". However, such a claim would require a thorough contrastive analysis of English and Chinese to confirm the equivalence between the English and the Chinese constructions. Moreover, a study of causative constructions in different varieties of learner English has demonstrated that the overuse of *make sb/sth feel* and *make sb/sth become* is in fact characteristic of several other L1 populations of learners (Gilquin 2012), which suggests that Liu & Shaw's (2001) results do not point to a case of transfer (or at least not only), but a more general tendency.

The last few years have witnessed a general refinement of the methods of analysis in learner corpus research. One major change is the increasingly prominent role of statistics in the field. While statistical significance testing has almost always been part of learner corpus studies, through the notions of over- and underuse, criticism has recently been voiced against this type of monofactorial statistics. Gries & Deshors (2014), for example, argue that, instead of comparing overall frequencies in learner and native corpora, researchers should look at the linguistic contexts in which an item is used – or not – by learners and native speakers, as determined by a multifactorial analysis involving a variety of morpho-syntactic and semantic features. Statistics also help researchers go beyond the typical global approach of corpus linguistics (studying corpora as wholes), by taking corpus/learner variation into account through statistical techniques such as Wilcoxon tests (e.g. Paquot 2014) or linear modelling (e.g. Meunier & Littré 2013). By adopting this more individual type of approach, learner corpus research is following the general quantitative trend in corpus linguistics as well as theories in second language acquisition (SLA) research like the Dynamic Systems Theory, which focuses on "individual developmental paths" (De Bot et al. 2007: 14). The link with theoretical frameworks, incidentally, is another way in which learner corpus research has evolved over the last few years. More and more learner corpus studies nowadays are grounded in SLA theories (see Myles 2015) or usage-based theories like cognitive linguistics (see De Knop & Meunier 2015), which gives such studies a more solid background and helps improve their explanatory power. Finally, methodological refinement in learner corpus research also comes from its rapprochement with the field of natural language processing, which has provided powerful tools and techniques for the automated analysis of large datasets (see Meurers 2015).

Representative study 1: Altenberg, Bengt, and Sylviane Granger (2001) The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics* 22(2): 173-194.

This study of the grammatical and lexical patterning of the high-frequency verb *make* among French- and Swedish-speaking learners of English seeks to test a number of hypotheses from the literature, e.g. the idea that a core verb like *make* is safe to use (hence not error-prone) or the contradictory claims that high-frequency verbs tend to be underused/overused by learners. It uses the French and Swedish components of ICLE, as well as a comparable native English corpus, the Louvain Corpus of Native English Essays (LOCNESS). The article provides a good overview of some of the techniques that can be applied to learner corpus data, including a comparison of the overall frequency of *make* in the three (sub)corpora, an examination of the distribution of its main semantic uses, a phraseological analysis of the collocates of the verb, and a syntactic and error analysis of its causative uses. In addition, the potential role of the mother tongue is examined, and some possible cases of transfer are highlighted, as well as strategies that appear to be common to the two groups of learners (e.g. a "decompositional" strategy which results in constructions like *make the family live* instead of *support the family*). Interestingly, the article also discusses the methodological issue of how accurate and useful an automatic extraction of collocates is. More generally, it demonstrates the benefits of combining an automatic and manual analysis, as well as a quantitative and qualitative approach.

Representative study 2: Lüdeling, Anke, Hagen Hirschmann, and Anna Shadrova (2017) Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning* 67(S1): 96-129.

This study focuses on German as a foreign language, and how advanced learners acquire morphological productivity for German complex verbs, that is, prefix verbs (like *verstehen* 'to understand') and particle verbs (like *aufstehen* 'to get up'). Looking at the treatment of morphological productivity in different acquisition models, including generative and usage-based models, the authors put forward a number of hypotheses, which are then tested against a learner corpus. The corpus is Falko (see Sect. 2.3) and its L1 equivalent. The study combines Contrastive Interlanguage Analysis and Computer-aided Error Analysis. First, it compares the frequency and uses of complex verbs in learner and native German. Second, it relies on the error tagging of Falko to identify grammatical and ungrammatical uses of complex verbs and to determine error types. The results show that learners tend to underuse prefix verbs and, especially, particle verbs, and that the variance between individual learners is greater than that between individual native speakers. Learners also appear to use complex verbs productively, although the new forms they produce sometimes result in errors. The paper illustrates some of the latest developments in learner corpus research, such as a solid grounding in theories and a combined aggregate and individual approach. It also makes the interesting methodological point that, through corpus annotation, categorization of the data can be made explicit and available to other researchers.

Representative study 3: Alexopoulou, Theodora, Jeroen Geertzen, Anna Korhonen, and Detmar Meurers (2015) Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1): 96-129.

This study is based on one of the large learner corpora coming out of the testing/assessment world (see Sect. 2.1), namely EFCAMDAT, the EF Cambridge Open Language Database. EFCAMDAT is made up of 33 million words, representing 85,000 learners and spanning sixteen proficiency levels. Although the corpus includes longitudinal data for certain individual learners, this study adopts an aggregate approach, considering each proficiency level as a 'section', but with the acknowledgment that "combining the cross-sectional perspective with an analysis of individual learner variation is a necessary next step" (p. 126). The paper investigates the development of learners' use of relative clauses. Like Lüdeling et al. (2017), it is grounded in theories of (second language) acquisition. In addition, it illustrates the rapprochement between learner corpus research and natural language processing (NLP), since it makes use of NLP tools and techniques to automatically extract relative clauses from a "big data" resource and to analyze their uses. The study reveals that very few relative clauses are found before Level 4, that their frequency increases until Level 6 and that it then remains more or less stable, with a peak at Level 11. The results show some limited effect of learners' nationalities (in terms of the types of relative clauses that are used) and a strong task effect. This focus on tasks echoes Granger's (2015) recommendation to take this kind of variable into account (see Sect. 2.4). However, other variables that are equally important in learner corpus research cannot be investigated in EFCAMDAT because of the relative lack of metadata about learners (information about their L1, for example, is so far not available but has to be approximated through nationality and country of residence). This shows that, at the moment, learner corpus size may still come at the expense of rich metadata.

Representative corpora

International Corpus of Learner English (ICLE; Granger et al. 2009)

One of the first learner corpora to have been compiled, ICLE is a mono-L2 and multi-L1 corpus, in that it contains data from a single target language, English, produced by (high-intermediate to advanced) learners from different L1 backgrounds. It is a written learner corpus made up of argumentative (and some literary) essays written by university students under different conditions (exam or not, timed or untimed, access to reference tools or not). It is accompanied by rich metadata which can be queried through the interface that comes with the released version of the corpus. In its current version, it contains 3.7 million words, representing 16 L1 backgrounds. The whole corpus has been POS tagged.

Corpus Escrito del Español L2 (CEDEL2; http://cedel2.learnercorpora.com)

CEDEL2, directed by Cristóbal Lozano, is a mono-L2 and multi-L1 learner corpus, made up of Spanish learner data produced by speakers of various L1s. It includes texts written by learners of all proficiency levels, from beginners to advanced learners. The texts were collected via a web application, together with detailed metadata. Unlike many learner corpora which fail to include precise information about learners' proficiency levels (see Sect. 3), CEDEL2 provides, for each learner, the result of an independent and standardized placement test which the participants also took online. The corpus currently includes over

one million words. It comes with native Spanish corpora built according to the same design criteria, which can be used for L1-L2 comparisons.

Parallèle Oral en Langue Étrangère (PAROLE; Hilton et al. 2008)
PAROLE is a multi-L2 and multi-L1 spoken learner corpus, which represents L2 Italian, French and English speech produced by learners from various L1 backgrounds and proficiency levels. It also contains some data produced by L1 speakers. The data were collected through five oral production tasks, which correspond to varying degrees of naturalness. Next to the usual type of information (learner's L1, knowledge of other languages, etc.), the metadata include, for each learner, measures of L2 proficiency, phonological memory, grammatical inferencing and motivation. PAROLE is a speech (or speaking) learner corpus, which means that, unlike so-called mute spoken learner corpora, it comes with sound files. The data have been transcribed according to the CHILDES system (see Sect. 3) and the transcriptions have been time-aligned with the sound files (see Chap. 11 on time-alignment).

3 Critical assessment and future directions

Over the last few years, learner corpora have grown in number, size and diversity. Written learner corpora are already quite numerous and large. In the near future, we should see the release of more and bigger spoken learner corpora, like the (still growing) Trinity Lancaster Corpus (Gablasova et al. 2017). In this respect, it is to be hoped that the developments in speech recognition will one day make it possible to automatically create reliable transcriptions based on recordings of learner language. In Zechner et al. (2009), the authors tested the reliability of a speech recognizer that they had trained on non-native spoken English produced by learners from a wide range of L1 backgrounds and proficiency levels. The result was that about one word in two was (wholly or partly) incorrectly transcribed. Although progress has been made in the meantime, Higgins et al. (2015: 593) still acknowledge that the performance of speech recognizers "can degrade substantially when they are presented with non-native speech".

Another possible development is that the learner corpora of the future will be mega databases (rather than corpora in the strict sense), bringing together data produced by the same learners in different contexts, with different degrees of monitoring (thus including some constrained data, even perhaps of an experimental nature, in addition to the more naturalistic data), at different stages in their learning process and in different languages, including their mother tongue. The last-mentioned type of data, L1 data to be compared with L2 data from the same subjects, can help distinguish between linguistic behaviours that are typical of a person, regardless of whether s/he is using his/her mother tongue or a non-native language (e.g. a slow speech rate), and those that the person only displays when using the L2. García Lecumberri et al. (2017), for instance, have compiled a bi-directional corpus made up of speech produced by English and Spanish native speakers in both their L1 and their L2, and they show how such a corpus can open up new possibilities for the study of learner language.

More and more learner corpora nowadays come with an equivalent L1 corpus representing the target language (cf. CEDEL2 and PAROLE). This is a welcome development, as it makes it possible to carry out contrastive interlanguage analyses on the basis of fully comparable data. Such target language data are likely to be included in the mega databases of the future. What would also be desirable is input data, which should strive to represent the language that learners get exposed to, so that correlations between input and output can be measured. While in the past learners' input has been approximated by means of

textbook corpora (cf. Römer 2004), it is clear that, especially in the case of an international language like English, learners' input is no longer limited to textbooks, even in foreign language situations, and that additional sources of exposure to the target language should therefore be taken into account.

At the same time as we should witness an exponential growth in the size of learner corpora/databases, we should also observe the creation of new types of learner corpora, some of which have already started to be collected. The PROCEED corpus (Process Corpus of English in Education),[3] for example, is a 'process learner corpus' which aims to reflect the whole of the writing process among language learners. It does so by combining screencast and keystroke logging and by examining at a micro-level the different steps leading to the final product (see Gilquin Forthcoming). Multimodal learner corpora (see Chap. 16) like the Multimedia Adult ESL Learner Corpus (MAELC; Reder et al. 2003) are likely to become more common, as well as translation learner corpora (corpora of texts translated by non-native students / translator trainees) like the MeLLANGE Learner Translator Corpus (Castagnoli et al. 2011) or the Multilingual Student Translation (MUST) corpus (see Chap. 12). More generally, it seems as if the new generation of learner corpora will be characterized by a higher degree of diversification than is currently the case: more (target and first) languages will be represented, more proficiency levels (including young learners, as in the International Corpus of Crosslinguistic Interlanguage; Tono 2012), more tasks, etc. The use of web applications to collect learner corpus data (cf. CEDEL2) will also make it possible to include the production of a wider range of non-native populations, and in particular learners outside universities, where, for reasons of convenience, many participants so far have been recruited.

In addition to an expansion and diversification of learner corpora, we can also expect these corpora to come with more additional information than ever before, in the form of metadata and annotation. Starting with metadata, although learner corpora have included a large variety of them from the very beginning, there is also a growing recognition that these may not be enough to reflect the complexity of the second language acquisition process. Limiting target language exposure to the 'time abroad' factor, for example, means neglecting other possible sources of exposure like the Internet, TV series or songs, all of which have become omnipresent in the lives of many young people. Proficiency is another case in point. While typically it has been evaluated on the basis of external criteria such as age or number of years of English instruction, scholars like Pendar & Chapelle (2008) have demonstrated that these may only give a very rough approximation of a learner's actual proficiency, which speaks in favour of having the participants take a placement test as part of the data collection procedure (cf. CEDEL2) and/or having the corpus data rated according to a scale like the CEFR. More cognitive measures are also likely to be added in the future, as is the case in PAROLE or in the Secondary-level COrpus Of Learner English (SCooLE), which relies on a whole battery of psychometric tests measuring verbal comprehension, reasoning, perseverance, anxiety and many others (see Möller 2017). In terms of annotation, we can expect learner corpora to more systematically be POS tagged, parsed and/or error tagged (to cite only the main types of annotation mentioned in Sect. 2.3), which should be easier once adequate tools have been designed or adapted to deal with learner language more accurately. As with other types of corpora (see, e.g., spoken corpora in Chap. 11), standardization will become even more important as metadata and annotation keep being added. A project like the Child Language Data Exchange System (CHILDES)[4] has contributed to the standardization of child-language corpora by proposing a common format for transcription, POS tagging, etc.

[3] https://uclouvain.be/en/research-institutes/ilc/cecl/proceed.html
[4] https://childes.talkbank.org/

Although some learner corpora have adopted this system too, like PAROLE or the French Learner Language Oral Corpora (FLLOC),[5] they are relatively rare, and there is currently no corresponding system for learner corpora which could ensure the same degree of standardization.

The availability of more, more diverse, bigger and more richly annotated learner corpora will have an impact on the way we conduct learner corpus research. Ellis et al. (2015), for example, call for "more longitudinal studies based on dense data". This will involve, first, the compilation of bigger and denser longitudinal learner corpora. Once these corpora have been collected, appropriate techniques will have to be developed to automate the analysis of individual developmental trajectories in large datasets (see Hokamura 2018 for a step in this direction, based on a set of twenty data collection points but limited to two learners). With such techniques, it will become possible to investigate much larger populations of learners than is currently the case and thus achieve a higher degree of reliability. It can also be hoped that new and better resources will attract more users. In particular, teachers should be encouraged not only to use learner corpora, but also to collect data produced by their own students, in the form of 'local learner corpora' (Seidlhofer 2002). With more and more teachers receiving some training in corpus linguistics, we can expect that an increasingly large number of them will want to apply the methods of learner corpus research in their classrooms, thus bringing learner corpora closer to those who, ultimately, should benefit from their exploitation, namely learners.

4 Tools and resources

Learner Corpus Bibliography: this bibliography is made up of references in the field of learner corpus research. The bibliography can be found on the CECL website (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html). A searchable version is accessible to members of the Learner Corpus Association in the form of a Zotero collection.

Learner Corpora around the World (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html): this website, which is regularly updated, contains a list of learner corpora, together with their main characteristics (target language, first language, medium, text/task type, proficiency level, size) as well as information about whether (and how) they can be accessed.

Université Catholique de Louvain Error Editor (UCLEE; Hutchinson 1996): this program facilitates error tagging thanks to a drop-down menu that makes it possible to select an error tag. It also facilitates the insertion of a corrected form. A new version of the software is currently in preparation.

Compleat Lexical Tutor (Lextutor; http://www.lextutor.ca): this website, created by Tom Cobb, is mainly aimed at teachers and learners (of English, but also some other languages like French). However, among the many tools it offers, some will be useful to researchers working with learner corpora. VocabProfile, in particular, can analyse (small) learner corpora according to vocabulary frequency bands, making it possible to check whether, say, learners of English tend to rely heavily on the 1000 most frequent words of the English language.

[5] www.flloc.soton.ac.uk/

5 Further reading

Granger, Sylviane. 2012. How to use foreign and second language learner corpora. In *Research methods in second language acquisition: A practical guide*, eds. Alison Mackey, and Susan M. Gass, 7-29. Chichester: Blackwell Publishing.
After briefly introducing learner corpora, this paper clearly presents the different stages that can be involved in a learner corpus study: choice of a methodological approach, selection and/or compilation of a learner corpus, data annotation, data extraction, data analysis, data interpretation and pedagogical implementation.

Díaz-Negrillo, Ana, Nicolas Ballier, and Paul Thompson, eds. 2013. *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins.
This edited volume covers many important methodological issues related to learner corpora, such as the question of interoperability, multi-layer error annotation, automatic error detection and correction, or the use of statistics in learner corpus research.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
This handbook provides a comprehensive overview of the different facets of learner corpus research, including the design of learner corpora, the methods that can be applied to study them, their use to investigate various aspects of language, and the link between learner corpus research and second language acquisition, language teaching and natural language processing.

References

Alexopoulou, Theodora, Jeroen Geertzen, Anna Korhonen, and Detmar Meurers. 2015. Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research* 1(1): 96-129.

Altenberg, Bengt, and Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics* 22(2): 173-194.

Belz, Julie, and Nina Vyatkina. 2005. Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: The case of German modal particles. *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 62(1): 17-48.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A corpus of non-native English*. Princeton, NJ: Educational Testing Service.

Caines, Andrew, and Paula Buttery. 2014. The effect of disfluencies and learner errors on the parsing of spoken learner language. *First joint workshop on statistical parsing of morphologically rich languages and syntactic analysis of non-canonical languages*, Dublin, Ireland, August 23-29 2014, 74-81.

Callies, Marcus. 2015. Learner corpus methodology. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 35-55. Cambridge: Cambridge University Press.

Castagnoli, Sara, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2011. Designing a learner translator corpus for training purposes. In *Corpora, language, teaching, and resources: From theory to practice*, ed. Natalie Kübler, 221-248. Bern: Peter Lang.

Crossley, Scott A., and Danielle S. McNamara. 2012. Interlanguage talk: A computational analysis of non-native speakers' lexical production and exposure. In *Applied natural language processing: Identification, investigation and resolution*, eds. Philip M. McCarthy, and Chutima Boonthum-Denecke, 425-437. Hershey: IGI Global.

Dagneaux, Estelle, Sharon Denness, and Sylviane Granger. 1998. Computer-aided error analysis. *System* 26(2): 163-174.

Dagneaux, Estelle, Sharon Denness, Sylviane Granger, Fanny Meunier, JoAnne Neff, and Jennifer Thewissen. 2008. *Error tagging manual version 1.3*. Louvain-la-Neuve, Centre for English Corpus Linguistics.

de Bot, Kees, Wander Lowie, and Marjolijn Verspoor. 2007. A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1): 7-21.

De Felice, Rachele, and Stephen Pulman. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal* 26(3): 512-528.

de Haan, Pieter. 1984. Problem-oriented tagging of English corpus data. In *Corpus linguistics: Recent developments in the use of computer corpora*, eds. Jan Aarts, and Willem Meijs, 123-139. Amsterdam: Rodopi.

de Haan, Pieter. 2000. Tagging non-native English with the TOSCA–ICLE tagger. In *Corpus linguistics and linguistic theory*, eds. Christian Mair, and Marianne Hundt, 69-79. Amsterdam: Rodopi.

De Knop, Sabine, and Fanny Meunier. 2015. The 'learner corpus research, cognitive linguistics and second language acquisition' nexus: A SWOT analysis. *Corpus Linguistics and Linguistic Theory* 11(1): 1-18.

Díaz-Negrillo, Ana, Nicolas Ballier, and Paul Thompson, eds. 2013. *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1-2): 139-154.

Ellis, Nick C., Rita Simpson-Vlach, Ute Römer, Matthew Brook O'Donnell, and Stefanie Wulff. 2015. Learner corpora and formulaic language in second language acquisition research. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 357-378. Cambridge: Cambridge University Press.

Flowerdew, Lynne. 2015. Learner corpora and language for academic and specific purposes. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 465-484. Cambridge: Cambridge University Press.

Gablasova, Dana, Vaclav Brezina, Tony McEnery, and Elaine Boyd. 2017. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics* 38(5): 613-637.

García Lecumberri, María Luisa, Martin Cooke, and Mirjam Wester. 2017. A bi-directional task-based corpus of learners' conversational speech. *International Journal of Learner Corpus Research* 3(2): 175-195.

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In *Selected proceedings of the 2012 second language research forum: Building bridges between disciplines*, eds. Ryan T. Miller, Katherine I. Martin, Chelsea M. Eddington, Ashlie Henery, Nausica Marcos Miguel, Alison M. Tseng, Alba Tuninetti, and Daniel Walter, 240-254. Somerville, MA: Cascadilla Proceedings Project.

Gilquin, Gaëtanelle. 2000/2001. The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3(1): 95-123.

Gilquin, Gaëtanelle. 2012. Lexical infelicity in English causative constructions. Comparing native and learner collostructions. In *Analytical causatives. From 'give' and 'come' to 'let' and 'make'*, eds. Jaakko Leino, and Ruprecht von Waldenfels, 41-63. München: Lincom Europa.

Gilquin, Gaëtanelle. 2016. Discourse markers in L2 English: From classroom to naturalistic input. In *New approaches to English linguistics: Building bridges*, eds. Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, and Sarah Chevalier, 213-249. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle. 2017. POS tagging a spoken learner corpus: Testing accuracy testing. Paper presented at the 4th Learner Corpus Research Conference, Bolzano/Bozen, Italy, 5-7 October 2017.

Gilquin, Gaëtanelle. Forthcoming. Hic sunt dracones: Exploring some *terra incognita* in learner corpus research. In *variation in time and space: Observing the world through corpora*, eds. Anna Čermáková, and Markéta Malá. Berlin: De Gruyter.

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

Golden, Anne, Scott Jarvis, and Kari Tenfjord. 2017. *Crosslinguistic influence and distinctive patterns of language learning: Findings and insights from a learner corpus*. Bristol: Multilingual Matters.

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in contrast. Text-based cross-linguistic studies*, eds. Karin Aijmer, Bengt Altenberg, and Mats Johansson, 37-51. Lund: Lund University Press.

Granger, Sylviane. 1998. The computer learner corpus: A testbed for electronic EFL tools. In *Linguistic databases*, ed. John Nerbonne, 175-188. Stanford: CSLI Publications.

Granger, Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In *Applied corpus linguistics: A multidimensional perspective*, eds. Ulla Connor, and Thomas Upton, 123-145. Amsterdam: Rodopi.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In *Corpora and language teaching*, ed. Karin Aijmer, 13-32. Amsterdam: John Benjamins.

Granger, Sylviane. 2012. How to use foreign and second language learner corpora. In *Research methods in second language acquisition: A practical guide*, eds. Alison Mackey, and Susan M. Gass, 7-29. Chichester: Blackwell Publishing.

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

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

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

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

Higgins, Derrick, Chaitanya Ramineni, and Klaus Zechner. 2015. Learner corpora and automated scoring. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 587-604. Cambridge: Cambridge University Press.

Hilton, Heather, John Osborne, Marie-Jo Derive, Nejma Succo, Jean O'Donnell, Sandra Billard, and Sandrine Rutigliano-Daspet. 2008. *Corpus PAROLE (Parallèle Oral en Langue Étrangère). Architecture du corpus & conventions de transcription*. Chambéry: Laboratoire LLS – Équipe Langages, Université de Savoie. http://archive.sfl.cnrs.fr/sites/sfl/IMG/pdf/PAROLE_manual.pdf. Accessed 4 August 2018.

Hokamura, Michiyo. 2018. The dynamics of complexity, accuracy, and fluency: A longitudinal case study of Japanese learners' English writing. *JALT Journal* 40(1): 23-46.

Huang, Yan, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics* 23(1): 28-54.

Hutchinson, John. 1996. Université Catholique de Louvain Error Editor. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Izumi, Emi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management* 12(2): 119-125.

James, Carl. 1998. *Errors in language learning and use: Exploring error analysis*. London and New York: Longman.

Jarvis, Scott, and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. New York and London: Routledge.

Jucker, Andreas H., Sara W. Smith, and Tanja Lüdge. 2003. Interactive aspects of vagueness in conversation. *Journal of Pragmatics* 35(12): 1737-1769.

Liu, Eric T.K., and Philip M. Shaw. 2001. Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics in Language Teaching* 39(3): 171-194.

Lüdeling, Anke, Hagen Hirschmann, and Anna Shadrova. 2017. Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning* 67(S1): 96-129.

Meunier, Fanny. 1998. Computer tools for interlanguage analysis: A critical approach. In *Learner English on computer*, ed. Sylviane Granger, 19-37. London and New York: Addison Wesley Longman.

Meunier, Fanny. 2016. Introduction to the LONGDALE Project. In *Studies in learner corpus linguistics. Research and applications for foreign language teaching and assessment*, eds. Erik Castello, Katherine Ackerley, and Francesca Coccetta, 123-126. Berlin: Peter Lang.

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

Meurers, Detmar. 2015. Learner corpora and natural language processing. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 537-566. Cambridge: Cambridge University Press.

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

Myles, Florence. 2015. Second language acquisition theory and learner corpus research. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 309-331. Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2004. Learner corpora and their potential in language teaching. In *How to use corpora in language teaching*, ed. John Sinclair, 125-152. Amsterdam: John Benjamins.

Osborne, John. 2015. Transfer and learner corpus research. In *The Cambridge handbook of learner corpus research*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 333-356. Cambridge: Cambridge University Press.

Paquot, Magali. 2014. Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. In *EUROSLA Yearbook 14*, eds. Leah Roberts, Ineke Vedder, and Jan H. Hulstijn, 240-261. Amsterdam: John Benjamins.

Pendar, Nick, and Carol A. Chapelle. 2008. Investigating the promise of learner corpora: Methodological issues. *CALICO Journal* 25(2): 189-206.

Rayson, Paul, and Alistair Baron. 2011. Automatic error tagging of spelling mistakes in learner corpora. In *A taste for corpora: In honour of Sylviane Granger*, eds. Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, 109-126. Amsterdam: John Benjamins.

Reder, Stephen, Kathryn Harris, and Kristen Setzler. 2003. The Multimedia Adult ESL Learner Corpus. *TESOL Quarterly* 37(3): 546-557.

Reznicek, Marc, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In *Automatic treatment and analysis of learner corpus data*, eds. Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, 101-124. Amsterdam: John Benjamins.

Römer, Ute. 2004. Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In *Corpora and language learners*, eds. Guy Aston, Silvia Bernardini, and Dominic Stewart, 152-168. Amsterdam: John Benjamins.

Rozovskaya, Alla, and Dan Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 154-162. Los Angeles: Association for Computational Linguistics.

Seidlhofer, Barbara. 2002. Pedagogy and local learner corpora: Working with learning-driven data. In *Computer learner corpora, second language acquisition and foreign language teaching*, eds. Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, 213-234. Amsterdam: John Benjamins.

Sinclair, John. 1996. Preliminary recommendations on corpus typology, Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards). www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. Accessed 4 August 2018.

Spoelman, Marianne. 2013. The (under)use of partitive objects in Estonian, German and Dutch learners of Finnish. In *Twenty years of learner corpus research: Looking back, moving ahead*, eds. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 423-433. Louvain-la-Neuve: Presses universitaires de Louvain.

Tono, Yukio. 2012. International Corpus of Crosslinguistic Interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In *Developmental and crosslinguistic perspectives in learner corpus research*, eds. Yukio Tono, Yuji Kawaguchi, and Makoto Minegishi, 27-46. Amsterdam: John Benjamins.

Van Rooy, Bertus, and Lande Schäfer. 2002. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20: 325-335.

Zechner, Klaus, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51(10): 883-895.