1

Introduction: learner corpus research – past, present and future

Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier

Written and spoken data produced by learners has always been a key resource for the study of second language acquisition (SLA). However, for a long time the data used was rather artificial, i.e. resulting from highly controlled language tasks, and therefore not necessarily a reflection of what learners do in more natural communication contexts. In addition, the data samples were usually quite small, often involving no more than a handful of learners, and therefore raised concerns in terms of representativeness. The combined wish to address these two issues and produce more learner-aware/learner-focused pedagogical tools prompted the emergence of learner corpora, which can be defined as electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria. Learner corpora gave rise to a flurry of studies, which have come to be grouped under the umbrella term of 'learner corpus research' (LCR). This new research strand emerged in the late 1980s as an offshoot of corpus linguistics, a field which had shown great potential in investigating a wide range of native language varieties (diachronic, stylistic, regional) but had neglected the non-native varieties. In the case of English, by far the most widely investigated language at the time, this was hardly justified in view of the fact that the number of non-native speakers far outnumbers that of native speakers.

Having access to electronic collections of L2 data presents two significant advantages. First, as these collections are usually quite large and are collected from a great number of learners, they are arguably more representative than smaller data samples involving a limited number of learners. Second being in electronic format, the data can be analysed with a whole battery of software tools that greatly speed up the analysis and enable a wide range of investigations that either cannot be performed manually at all or only at huge cost in terms of human resources. Part-of-speech taggers, for example, assign to each word in a learner

corpus a tag that indicates its grammatical category, thereby facilitating investigations into learners' use of specific grammatical categories such as prepositions or auxiliary verbs. Concordance programs, on the other hand, have contributed to bringing lexis and phraseology to the forefront of L2 studies: they generate frequency lists of both single words and phrases, present all the instances of a linguistic item in their immediate linguistic context and include functionalities such as automatic extraction of collocations and word clusters or identification of keywords. For the analysis of errors, researchers can rely on error editors, which allow the insertion of error annotations into text files and play a key role in the type of error analysis carried out within LCR, known as computer-aided error analysis. Recent developments in automatic error detection and correction offer hope for increased automation of this key aspect of learner corpus research.

Although still quite young, the field of learner corpus research has already undergone remarkable developments. At first, learner corpus studies were limited to learner English. This was understandable in view of the position of English as the major lingua franca internationally, but in an increasingly multilingual society it is good to see the LCR field embrace an ever larger number of different L2s. The 'Learner corpora around the world' website maintained by the University of Louvain[1] currently contains 137 learner corpora, 82 (60%) representing L2 English, the rest focusing on other languages (Arabic, French, German, Korean, Spanish, etc.). In terms of medium and text type, the dominant focus was – and to a large extent still is – on writing, in particular essay writing, but there is a general diversification of data types and, especially, a growing number of projects on learner speech. Another significant trend concerns the research design: while there is still a preponderance of cross-sectional studies, i.e. studies that sample data from learners at a single point in time, the design of longitudinal corpora made up of data sampled from the same learners across time is showing a slow but steady rise. There is also a growing awareness among the learner corpus community of the need to pay greater attention to individual variability. Learner corpora lend themselves particularly well to the study of whole learner populations, and this global perspective undoubtedly has many benefits, especially from a teaching perspective. However, in view of the high degree of variability between and within learners, the exclusive use of aggregate data may be misleading. The growing application of more sophisticated statistical techniques in LCR has begun to remedy this weakness. Learner corpus researchers have started to realise that by using the appropriate statistics, it is possible to combine the best of both worlds: keep the group perspective which constitutes one of the strengths of LCR, while at the same time taking individual variability into account.

[1] www.uclouvain.be/en-cecl-lcworld.html (last accessed on 13 April 2015).

The Cambridge Handbook of Learner Corpus Research, with its aim to provide a state-of-the-art introduction to all the facets of the fast-expanding field of learner corpus research, reflects these recent developments. For example, it describes studies in a variety of L2s, deals with both speech and writing, includes a chapter on longitudinal research design and points to the need to attach greater importance to individual learners. At the same time, however, it is representative of the field in its current state, which necessarily means that certain target languages, text types, techniques or objects of study are less frequently dealt with than others.

While LCR shares with mainstream SLA research the objective of gaining a better understanding of the mechanisms of foreign or second language acquisition, LCR stands out because of its strong applied orientation. Initially limited to the sphere of foreign language teaching, it now includes a wide range of applications – in particular in natural language processing (NLP), such as automated scoring and automatic error detection and correction – which accordingly take pride of place in this handbook. As a result, LCR has become a truly interdisciplinary field at the crossroads between corpus linguistics, second language acquisition, language teaching and natural language processing. This makes LCR particularly fertile ground, but also brings with it the need to pull together the different research strands that are still insufficiently integrated. Several recent initiatives have been launched to foster greater synergy. The *Learner Corpus Association*,[2] set up in 2013, acts as an interdisciplinary forum for discussion and exchange of information on learner corpus research and coordinates the planning of a biennial international conference, the *Learner Corpus Research Conference*.[3] A new journal launched in 2015, the *International Journal of Learner Corpus Research*,[4] provides a dedicated publication outlet for research covering methodological, theoretical and applied work in any area of learner corpus research.

In line with these initiatives, *The Cambridge Handbook of Learner Corpus Research* aims to provide the rapidly growing community of researchers, teachers and students who are interested in this field with an overview of all the key aspects of learner corpus research.

The handbook is subdivided into five main parts:

1. learner corpus design and methodology
2. analysis of learner language
3. learner corpus research and second language acquisition
4. learner corpus research and language teaching
5. learner corpus research and natural language processing.

[2] www.learnercorpusassociation.org/ (last accessed on 13 April 2015).
[3] LCR2011 in Louvain-la-Neuve (Belgium), LCR2013 in Bergen (Norway) and LCR2015 in Nijmegen (the Netherlands).
[4] https://benjamins.com/#catalog/journals/ijlcr/main (last accessed on 13 April 2015).

Each of the chapters, all written by experts in their fields, introduces a different facet of LCR. They present state-of-the-art reviews and place emphasis on theoretical, methodological and applied aspects of wider relevance. The large number of chapters in Part I reflects the importance of corpus design and methodology for a good corpus analysis. The following topics are tackled: learner corpus design and collection, learner corpus methodology, learner corpora and psycholinguistic research, learner corpus annotation, speech annotation, error annotation and statistics for learner corpus research. Part II deals with the main foci of linguistic analysis in learner corpus research: lexis, phraseology, grammar, discourse and pragmatics. Part III situates learner corpus research within the general field of SLA and highlights more particularly the issues of transfer, formulaic language, developmental patterns, variability and the impact of the learning context. Part IV considers the links between LCR and language teaching, both in general and specific settings, introduces the notion of 'pedagogic corpus' and gives an overview of learner-corpus-informed pedagogical materials and testing practices. The last part is devoted to NLP applications, in particular automatic grammar- and spell-checking, automated scoring and automatic identification of the learner's native language. As a result of the interdisciplinary nature of LCR, each of the different research domains represented in the handbook has its own paradigms, theories and methodologies and it is essential to respect these. While it would not have been possible (nor indeed desirable) to present a unified theoretical or methodological framework, we have taken great care to enhance the coherence of the volume by including a large number of cross-references to help readers navigate through the chapters and confront different perspectives.

All the chapters in the handbook follow the same general format. After an introduction to the topic, the authors expand on a number of issues which they consider to be of particular importance. The third section describes in some detail two to four representative studies. This is an important section, as a handbook is meant not only to provide key theoretical information, but also to contain precise guidelines about how to do research in the field; the representative studies supply models of how to conduct learner corpus analyses. The fourth section takes a critical look at past and current research and points to promising future directions. This last section is especially relevant as the field of LCR is still very new, and twenty-five years after its advent it is time to take stock of the progress made and identify priorities for the future. To further assist users of the handbook, recommended key readings are provided at the end of each chapter, together with a short summary and an indication of their relevance to the topic of the chapter. These key readings, as well as the extended general bibliography, will allow researchers in the field to delve more deeply into all the aspects addressed. The handbook also features four indexes to facilitate navigation. Besides the traditional author and

subject indexes, there is one for all the corpora referred to in the volume and another for the software tools.

One of our main priorities in editing this handbook has been to cater for both novice and seasoned learner corpus researchers. For budding or would-be researchers, it will serve as an accessible introduction to all aspects of the field, with no unnecessary jargon and with all technical terms defined and illustrated. The representative studies described in each chapter make up a sort of how-to guide on how to conduct learner corpus research in a wide range of areas. For more experienced learner corpus researchers, the handbook will act as a prompt to embrace new perspectives: revise some of their methodological practices, add a new theoretical dimension, adopt a higher level of computational or statistical sophistication, enrich the interpretative side of the analysis or imagine new applications. A greater awareness of the interdisciplinary nature of LCR could also be an incentive to collaborate more closely with researchers in other disciplines.

This handbook aims to provide a comprehensive survey of learner corpus research. The multifaceted picture of the field that emerges from the different chapters highlights some of the major strengths of the research conducted to date but also points to shortcomings that need to be addressed and gaps that need to be filled. It is our hope that the handbook will be useful to a wide range of researchers from different disciplinary backgrounds. We also hope that it will play a key role in turning LCR into a fully mature field with stronger theoretical substrates and increased methodological rigour and that it will contribute to the production of a wide range of exciting new applications.

Part I

Learner corpus design and methodology

2

From design to collection of learner corpora

Gaëtanelle Gilquin

1 Introduction

Since the development of the field of second language acquisition (SLA), which Gass et al. (1998: 409) situate in the 1960s or 1970s, use has been made of authentic data representing learners' interlanguage. However, what has characterised many of these SLA studies is the small number of subjects investigated and the limited size of the data collected. This can be illustrated by the case studies selected by Ellis (2008: 9–17) as an 'introduction to second language acquisition research': Wong Fillmore's (1976, 1979) study of five Mexican children, Schumann's (1978) study of Alberto, Schmidt's (1983) study of Wes, Ellis's (1984, 1992) study of three classroom learners and Lardiere's (2007) study of Patty. While such studies have allowed for a very thorough and detailed analysis of the data under scrutiny (including individual variation and developmental stages), their degree of generalisation can be questioned (Ellis 2008: 8). In this respect, the expansion of corpus linguistics to the study of interlanguage phenomena has opened up new possibilities, materialised in the form of learner corpora.

Like any corpus, the learner corpus is a 'collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety' (McEnery et al. 2006: 5). What makes the learner corpus special is that it represents language as produced by foreign or second language (L2) learners. What makes it different from the data used in earlier SLA studies is that it seeks to be representative of this language variety. This element is emphasised by some of the definitions of learner corpora found in the literature, e.g. Nesselhauf's (2004: 125) definition as '*systematic* computerized collections of texts produced by language learners' (emphasis added), where 'systematic' means that 'the texts included in the corpus were selected on the basis of a number of – mostly external – criteria (e.g. learner level(s), the

learners' L1(s) [mother tongue(s)]) and that the selection is representative and balanced' (Nesselhauf 2004: 127). Design criteria are essential when collecting a learner corpus and will therefore be dealt with as one of the core issues (Section 2.2).

Another issue when defining learner corpora is their degree of naturalness. Granger's (2008a: 338) definition of learner corpora as 'electronic collections of *(near-) natural* foreign or second language learner texts assembled according to explicit design criteria' (emphasis added) suggests that they may comprise texts that are not, strictly speaking, naturally occurring texts.[1] This is because, for learners (especially foreign language learners), the target language fulfils only a limited number of functions, most of which are restricted to the classroom context. When learners engage in activities like writing a mock letter to an imaginary friend or doing role-plays with their classmates, the main objective is for them to practise and improve their skills in using the target language rather than to convey a genuine message. Data collected in such situations therefore do not represent the linguistic output of 'people going about their normal business' (Sinclair 1996), as would be expected of fully natural data. However, as is the case with corpora in general (see Gilquin and Gries 2009: 6), learner corpora may display varying degrees of naturalness, even when collected within the context of the school/university, from the more natural (e.g. the computer-mediated interactions between German and American students gathered in *Telekorp*; see Belz 2006)[2] to the more constrained (e.g. the retellings of a silent Charlie Chaplin movie included in the *Giessen-Long Beach Chaplin Corpus*; Jucker et al. 2003), through the semi-natural case of essay writing (e.g. *ICLE*, the *International Corpus of Learner English*; Granger et al. 2009), a pedagogical task that is natural in the context of the language learning classroom. In accordance with this continuum, and following Nesselhauf (2004: 128), learner data collected with more control on the language produced (e.g. the translations contained in the *UPF Learner Translation Corpus*; Espunya 2014) may be considered 'peripheral learner corpora'. When so much control is exerted that the learner is no longer free to choose his/her own wording, for instance in the case of a reading-aloud task, the term 'learner corpus' will normally be avoided.[3] Note that 'database' is sometimes used to refer to collections of learner data that have been gathered from both natural and less natural contexts, for example

[1] The definition also underlines, like Nesselhauf's (2004), the importance of design criteria in the compilation of learner corpora (see Section 2.2).

[2] *Telekorp* is the *Telecollaborative Learner Corpus of English and German*. It contains data produced by the students in their L1 and L2.

[3] It must be pointed out, however, that, e.g., Atwell et al. (2003) refer to *ISLE* (*Interactive Spoken Language Education*) as a corpus, although it includes recordings of German and Italian learners reading English texts. According to Gut (2014: 287), such collections of 'decontextualized sentences or text passages that are read out or repeated' qualify as 'peripheral types of learner corpora'. See also Chapter 6 (this volume) for a very broad use of the term 'learner corpus', covering highly constrained types of spoken data.