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.

LINDSEI, the *Louvain International Database of Spoken English Interlanguage* (Gilquin et al. 2010), which is made up of (in decreasing order of naturalness) free informal discussions, monologues on a set topic and picture descriptions.

Related to the concept of naturalness is what could be referred to as the degree of monitoring (in the sense of Krashen 1977) or editing of the data included in the corpus. Some data are produced with no prior planning and no subsequent editing (this is typically the case of speech, which by its very nature is more spontaneous than writing). When given sufficient time before and/or after language production, however, the learner can organise his/her discourse more carefully and (in the case of written discourse) revise and improve the text, possibly with the help of reference tools or feedback from an instructor. Some recent learner corpus projects which aim to make the writing process visible show various stages in the drafting of a text and thus reflect different degrees of editing/monitoring. The *Hanken Corpus of Academic Written English for Economics* (Mäkinen and Hiltunen 2014), for example, consists of the first drafts and final versions of end-of-term papers (before and after the teacher's feedback). The *Marburg Corpus of Intermediate Learner English* (*MILE*), on the other hand, seeks to represent the changes made during the writing process by marking deletions, additions or line breaks when digitising the learners' (handwritten) data (see Kreyer 2014). As noted by Kreyer (2014: 56), such alterations are interesting in that they can be 'regarded as an additional window onto the development of L2 competence'.

The above characterisation of the learner corpus normally excludes corpora like the *ELFA* (*English as a Lingua Franca in Academic Settings*) corpus, which contains data produced by L2 users (rather than L2 learners, see Mauranen 2011), and like *ICE* (*International Corpus of English*), which includes data produced by speakers of indigenised varieties of English – often, rather confusingly, referred to as English as a Second Language, but differing from the varieties included in learner corpora in that these indigenised varieties are used in countries where English is not a native language but has an official or semi-official status (see Chapter 19, this volume, on these 'new' varieties of English). However, these distinctions are not always clear-cut. The *NUS Corpus of Learner English* (Dahlmeier et al. 2013), for example, contains data produced by Singaporeans, who are speakers of an indigenised variety of English rather than learners of English in the strict sense; in this case, the use of the term 'learner corpus' might be justified by the fact that the data included in it were produced by undergraduate university students, not adult users. Nesselhauf (2004: 128), however, notes that the term 'learner' (and hence 'learner corpus') may also be applied to adult speakers 'in countries in which the status of the language in question is somewhere between foreign and second language (for example English in Hong Kong)'.

2 Core issues

2.1 Learner corpus typology

Several types of learner corpora can be distinguished, differing along one or more dimensions, some of which are common to all corpora while others are specific to learner corpora. The first dimension, which is crucial in determining how the data will be collected and turned into a corpus, is that of medium. Learner corpora can consist of written texts or transcriptions of spoken discourse. Unsurprisingly, the first learner corpora, which started to be collected in the late 1980s, were of the former type. Spoken corpora, which are more laborious to collect (see Section 2.3), only appeared later. Today, written learner corpora are still more numerous than spoken learner corpora – they are over twice as common according to the list of 'Learner Corpora around the World' (LCW) compiled by the University of Louvain[4] – but a number of spoken learner corpora have become available over the last few years and have started to form the basis of extensive research. Among spoken learner corpora, a distinction can be made between those that simply consist in written transcriptions of spoken discourse and those that are distributed with their corresponding sound files and thus give access to the speech signal; the terms 'mute spoken corpus' (Chapter 6, this volume) and 'speech corpus' (Wichmann 2008) can be used to describe this difference. Some learner corpora include both written and spoken data, like the *Santiago University Learner Corpus*.[5] As newcomers to the field, multimodal (or audio-visual) learner corpora (like *MAELC*, the *Multimedia Adult ESL Learner Corpus*; Reder et al. 2003) include video recordings, which give access to new domains of investigation like the analysis of learners' gazes or gestures, such as Hashimoto and Takeuchi's (2012) study of non-verbal elements in presentations, based on their *Multimedia Learner Corpus of Basic Presentation* (*MLCP*), in which each video-recorded presentation is accompanied by peer evaluations from the audience.

Genre is another aspect that may serve to categorise learner corpora. In principle, any genre (or combination of genres) may be represented in a learner corpus. However, in practice, the variety of genres tends to be limited as a result of (i) the restricted number of genres for which a second or (especially) foreign language variety is actually used (see Section 1) and (ii) learner corpus compilers' preference for certain genres, for example argumentative essays among written learner corpora, which correspond to over half of the written learner corpora included in the LCW list. Most learner corpora to date correspond to language as it is used for general purposes, but recently language for specific purposes (LSP) learner corpora have made their appearance. Unlike general learner corpora, which are

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

[5] www.sulec.es/ (last accessed on 13 April 2015).

mainly collected within the framework of general language courses, LSP learner corpora are made up of 'discipline and genre-specific texts written by learners within the framework of LSP or content courses' (Granger and Paquot 2013: 3142; see also Chapter 21, this volume). An example of such a corpus is the *Active Learning of English for Science Students (ALESS) Learner Corpus* (Allen 2009), which consists of research papers written by Japanese students majoring in science. Particularly interesting are learner corpora that contain a variety of genres, like the *MiLC Corpus* (Andreu Andrés et al. 2010) which includes, among others, essays, reports, formal and informal letters, summaries and business letters, as these kinds of corpora make it possible to compare interlanguage across genres.

Learner corpora can also be distinguished on the basis of the target language they represent. English, the language of the first learner corpora that were collected, is still the most predominant target language. However, over the last few years, new projects have been launched that seek to collect data representing other target languages, most notably French (e.g. *FLLOC*, *French Learner Language Oral Corpora*),[6] German (e.g. *Falko*, *Fehlerannotiertes Lernerkorpus*)[7] and Spanish (e.g. *CEDEL2*, *Corpus Escrito del Español L2*; see also Section 3.1),[8] which are the most widely represented target languages after English according to the LCW list. While most learner corpora are monolingual, containing data from only one target language, a small number of learner corpora are multilingual, like the *MiLC Corpus* mentioned above, which contains learner data in Catalan, English, French and Spanish, or the *USP Multilingual Learner Corpus* (Tagnin 2006), which has English, German, Italian and Spanish as target languages.

Besides the target language, one has to take the learner's mother tongue into account. Among the learner corpora that contain data produced by a single L1 population ('mono-L1 learner corpora'), it seems, on the basis of the LCW list, that Asian learners are the most widely represented, e.g. the *Taiwanese Learner Corpus of English* (Shih 2000) or the *Japanese English as a Foreign Language Learner (JEFLL) Corpus* (Tono 2007), but many other L1 populations are represented as well. Quite a few learner corpora (about a third of all the learner corpora included in the LCW list) are 'multi-L1'; in this case, learners from several L1 populations have contributed to the corpus (see Granger (2012a: 12) on the distinction between mono- and multi-L1 learner corpora). One such corpus is the *International Corpus of Learner Finnish* (Jantunen 2011), which contains data produced by learners of Finnish from several mother-tongue backgrounds, including Estonian, German, Polish, Russian and Swedish. Multi-L1 learner corpora are very useful for the study of L1 influence (see Chapter 15, this volume) as they are generally made up of subsets of data that are comparable across the

[6] www.flloc.soton.ac.uk/ (last accessed on 13 April 2015).

[7] www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko (last accessed on 13 April 2015).

[8] www.uam.es/proyectosinv/woslac/collaborating.htm (last accessed on 13 April 2015).

different L1 populations, thus making it possible to isolate interlanguage features that are typical of certain populations. It should be noted that some multi-L1 learner corpora do not allow for such comparisons as the learners' L1 is not identified (or at least not precisely enough), e.g. the *EF-Cambridge Open Language Database* (*EFCAMDAT*),[9] which currently includes information about the learners' nationalities but not about their L1.

In the same way as a general corpus may include data from one period in time (synchronic corpus) or from several periods (diachronic corpus), a learner corpus may be a snapshot of learners' knowledge of the target language at a particular moment or a representation of the evolution of their knowledge through time. Most learner corpora are of the former type, being made up of cross-sectional data. Corpora that seek to gather learner output produced at different stages in their development are called longitudinal corpora. They may vary in density depending on how often the data are gathered: the more regular the data collection, the denser the corpus. The *Longitudinal Database of Learner English* (*LONGDALE*)[10] is a project that aims to follow the same learners over a minimum period of three years, with at least one data collection per year. Increasing the number of collections per year would make the corpus denser. Belz and Vyatkina (2008: 33) use the term 'developmental learner corpus' to refer to dense corpora 'in which learner performance is documented at close intervals or at all points of production' – in their case, data from *Telekorp* that were collected from the same students over a two-month period (see Chapter 13, this volume). Longitudinal (and developmental) corpora make it possible to investigate learners' progress (or lack thereof) over time and are therefore a precious resource (see Chapter 17, this volume). However, because such corpora are difficult to compile (among other things because some learners drop out during the course of the data collection), there are very few currently available. For want of longitudinal learner corpora, researchers may instead resort to corpora of pseudo-longitudinal data (Gass and Selinker 2008: 56–7), also referred to as quasi-longitudinal data (Granger 2002: 11). Such corpora are gathered at a specific point in time but from (different) learners representing different proficiency levels. The *NICT JLE Corpus* (*National Institute of Information and Communications Technology Japanese Learner English Corpus*; Izumi et al. 2004) is quasi-longitudinal as it contains data produced by different learners and divided into nine proficiency levels. Learner corpora may also include both truly and quasi-longitudinal data, as illustrated by the *Corpus of Learner German* (*CLEG13*).[11]

[9] http://corpus.mml.cam.ac.uk/efcamdat/index.php (last accessed on 13 April 2015).

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

[11] http://korpling.german.hu-berlin.de/public/CLEG13/CLEG13_documentation.pdf (last accessed on 13 April 2015).

A distinction can also be drawn between global and local learner corpora. Most learner corpora are global, being part of large-scale projects and being collected among learners who are subjects providing data for inclusion in the corpus. Local learner corpora, on the other hand, are typically collected by teachers among their students, who are both contributors to and users of the corpus. The objective of this approach is to identify one's own learners' specific needs through a corpus analysis of their output and thus provide tailor-made solutions to their problems. Mukherjee and Rohrbach (2006) illustrate the compilation and use of a local learner corpus, the *Giessen-Göttingen Local Learner Corpus of English* (see also Millar and Lehtinen (2008) for an overview of the compilation and analysis of local learner corpora).[12]

Finally, learner corpora can be distinguished on the basis of their origin and the main purpose for which they were created. Commercial learner corpora are collected by publishing houses with a view to developing pedagogical materials (dictionaries, coursebooks, etc.) based on authentic learner output (see Chapter 22, this volume). Most of the time, these corpora are not publicly available. The two most notable examples of commercial learner corpora are the *Longman Learners' Corpus*[13] and the *Cambridge Learner Corpus*.[14] Unlike commercial learner corpora, academic learner corpora are initiated by researchers and/or teachers working in educational settings and interested in learning more about interlanguage (possibly with pedagogical aims in mind).

In what precedes, a number of learner corpora have already been mentioned, many of which are publicly available for research purposes. However, there are some cases in which it may be necessary, or desirable, to collect one's own data. This might be because, as with local learner corpora, the researcher wants to have access to data collected in a specific environment (his/her classroom, school, area, etc.), or because the ready-made learner corpora that are available do not suit his/her research purposes (e.g. they are too small, do not contain enough metadata or, in the case of (mute) spoken learner corpora, have not been transcribed in sufficient detail). The next sections will provide an overview of how to go about collecting a learner corpus, starting with the issue of design criteria. The focus will be on those features that are specific to learner corpora; for a good overview of the issues to be considered when compiling a (general) corpus, see the chapters collected in O'Keeffe and McCarthy (2010, Section II).

[12] A further type of corpus is sometimes recognised in between global and local learner corpora, namely in-house learner corpora, i.e. 'local reference learner corpora which reflect the production of a given learner population' (Rankin and Schiftner 2011: 430). In this case, the contributors and the users are not the same students, but they come from the same population (typically, the same school/university), which enhances the relevance of the analyses of these data for the users.

[13] www.pearsonlongman.com/dictionaries/corpus/learners.html (last accessed on 13 April 2015).

[14] www.cambridge.org/gb/cambridgeenglish/about-cambridge-english/cambridge-english-corpus (last accessed on 13 April 2015).

2.2 Design: environment, task and learner variables

The importance of adopting strict criteria when designing a corpus has been regularly emphasised in the literature (e.g. Atkins et al. 1992). In the case of learner corpora, design criteria are even more crucial given the highly heterogeneous nature of interlanguage, which can be affected by many variables related to the environment, the task and the learner him-/herself. Before embarking on the collection of a learner corpus, the researcher therefore has to think carefully about what exactly will be included in the corpus for, as pointed out by Granger (2013a: 3235), '"mixed bag" collections of L2 data present little interest'. In this respect, the dimensions presented in the previous section will obviously have a role to play: whether one wishes to collect, say, spoken or written learner data will have an influence on the way the corpus will be compiled and how the data will be analysed and interpreted. But there are many other variables that could be taken into account. These variables can pertain to the environment in which the data are collected, the tasks which the subjects are carrying out during the data collection, and the learners whose performances are being recorded.

In terms of the environment, a major distinction can be made between cases where the target language is a native language that is used in everyday interactions in the learner's environment (second language) and cases where the target language has no such functions and is normally confined to the classroom (foreign language). In addition, one can distinguish between collection of the data in an educational setting (at school/university) and in a natural setting (outside school/university). This distinction is especially relevant for second language learner corpora since second languages can be used in a wider variety of contexts, but foreign languages can sometimes also be used outside the educational setting, for example when a learner writes a letter or an email to a pen friend from home.

Task variables are closely related to the notions of medium and genre (see Section 2.1). Producing an argumentative essay or orally describing a picture, for instance, will activate very different mechanisms and will offer different possibilities for controlling the way the task is performed. Written tasks can involve variables like time constraints (did the learner have a limited amount of time available to write the text?), availability of reference tools (dictionaries or grammar books), intertextuality (did the learner have access to secondary sources such as articles or other students' essays?) and computerisation (did the learner write by hand or using a computer?). Task variables for spoken learner corpora include preparation time (did the learner have time to think about what s/he was going to say?), written support (did the learner have access to some written support, either notes of his/her own or text to which s/he is supposed to react?) and technique of recording (e.g. was the technique invasive or not?). In addition, one should consider whether the task was part of an

exam, as exam conditions may place students under increased pressure. Topic also has an important role to play as it may influence certain aspects of learner production (especially more lexical ones).

Unsurprisingly, many of the variables that affect the nature of interlanguage concern the learners themselves. Some of these variables are general, being applicable to any speaker/writer, native or not, e.g. age, gender, country/area, mother tongue. Other variables are more specifically relevant to learners, like the parents' native languages, the language(s) spoken at home, the learner's proficiency level, exposure to the target language inside the classroom (e.g. number of years spent learning the target language, pedagogical materials used) and outside the classroom (e.g. contact with the target language in everyday life, stays in target-language countries), or knowledge of other foreign languages. Different measures of the learner's proficiency and/or motivation may also be provided. The *PAROLE Corpus* is an example of a learner corpus that offers a particularly wide variety of measures, including motivation, listening comprehension skills, grammatical and lexical competence, aptitude for grammatical analysis and phonological memory (see Hilton 2008).

In the case of spoken learner corpora, other participants may be involved in addition to the learner, and it may be useful to include variables about these participants too. In *LINDSEI*, for example, information was also gathered about the interviewer (gender, mother tongue, knowledge of other foreign languages and familiarity with the learner) as it was thought to have a possible influence on the learner's production (for example, a learner may be more likely to resort to words from his/her mother tongue if the interviewer has knowledge of this language; see also Chapter 13, this volume, on the potential impact of the relationship between the interlocutors).

Not all these variables should necessarily be controlled for, but (at least) some of them should be recorded. In other words, the learner corpus compiler does not have to take all these variables into account when deciding what to include or not in the learner corpus, but s/he should keep a record of as many of them as possible so that their impact on the learner's linguistic behaviour can be assessed (see Chapters 15, 18 and 19, this volume, on the attested influence of some of these variables). When designing a learner corpus, the researcher should therefore identify the features that will be shared by all the data (i.e. variables that are kept constant) and those that can vary across the data. S/he may, for instance, want to restrict the learner corpus to data produced by Italian-speaking learners with at least five years of learning Spanish, but leave it unspecified whether the learners should have spent time in a Spanish-speaking country or not (although this variable may still be recorded). Note that some variables are less likely to be kept constant in a learner corpus, for example gender: whenever possible, corpus designers will seek to strike a balance between male and female learners rather than targeting only

males or females. Usually, information about the variables is gathered through a form that is (partly) completed by the learner. This 'learner profile questionnaire' is often combined with the consent form that learners are required to sign if they allow their data to be used for research purposes. For certain variables, it might be necessary to have the learners take a test, for example to determine their proficiency level or motivation.

Finally, recording all these metadata is of little use if they are not made available to the corpus user, together with the actual data produced by the learners. As Burnard (2005: 31) puts it, '[w]ithout metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity'. Minimally, the metadata could be given in the header of each text making up the corpus (ideally in an XML-like format). In *ICLE*, the metadata are not integrated into the text files directly, but are included in a database which is linked to the text files, so that the user can select any number of variables and then extract the part of the corpus that corresponds to these criteria.

2.3 Collection of learner corpora

Once the learner corpus has been carefully designed, the first concrete step in collecting it is to select the subjects who will contribute to it. In practice, the learners tend to be recruited among the students with whom the compiler is in (direct or indirect) contact. When the task performed is integrated into the students' pedagogical activities (e.g. essay written within the frame of an exam), all the students may be expected to participate, and the selection will then be based on which students gave permission for their data to be used and which fulfil the criteria established during the design of the corpus (see Section 2.2). When the task performed is not part of the learners' normal curriculum, on the other hand, the compiler will often be dependent on their willingness to participate voluntarily. In this case, self-selection may introduce a bias in that certain types of learners may be more likely to volunteer than others (e.g. female rather than male learners, learners who are self-confident, motivated and/or consider their proficiency level to be relatively high). This, some may argue, can compromise the balance and representativeness of the learner corpus. However, it should be emphasised that the compiler is still free to remove some data from the corpus if they do not match the predefined criteria. Furthermore, as McEnery et al. (2006: 73) point out, the notions of balance and representativeness should be 'interpreted in relative terms, i.e. a corpus should only be as representative as possible of the language variety under consideration', as '[c]orpus-building is of necessity a marriage of perfection and pragmatism'. *CEDEL2* is an example of a corpus where the contributors, while volunteers and hence self-selected, come from a large, diverse and thus presumably representative pool of learners, since calls for participation were distributed via a

wide range of mailing lists and the learners could contribute data to the corpus via an online application from anywhere in the world (see Lozano and Mendikoetxea 2013 and Section 3.1).

The next steps involved in the collection of a learner corpus differ widely depending on the type of corpus that is collected. In what follows, a major distinction will be made between written and spoken learner corpora. Some other types of learner corpora will also be mentioned in passing. Written learner corpora start with either handwritten or typed texts. Handwriting was the norm when the first learner corpora were compiled, which involved keyboarding by the researcher. This part can be quite tricky, as the texts have to be reproduced exactly as they are, including the learners' errors but without introducing additional ones. Illegible handwriting can further complicate the task. Having typewritten texts scanned and converted through optical character recognition is another method of collection (here again the researcher should check that the result is an exact reproduction of the learner's output), but today most written learner corpora start straight from computerised versions of the learners' texts, either transferred electronically to the corpus compiler or directly uploaded (and even typed) via an online interface, which can also serve to collect the metadata related to the learner and to the text produced.

Once the raw texts have been collected, some mark-up may be added, such as a header containing a reference and details about the text, or metatextual information within the text itself, indicating, for example, formatting and layout properties. It may also be necessary to identify (by means of special tags) and/or remove some chunks of text, especially quotations (which do not represent the learner's own use of language and may therefore have to be excluded from the analysis of the corpus) and elements that may reveal the learner's identity.

If the learner corpus design is cross-sectional or quasi-longitudinal (see Section 2.1), the corpus compilation is complete once the data of all the selected learners have been collected. For longitudinal written learner corpora, the above procedure has to be repeated among the same learners at different points in time, as many times as required, depending on the desired density of the corpus.

Spoken corpora start from a sound, not a text. Collecting spoken learner data therefore requires as an initial step that the spoken output be recorded. This should be done with high-quality material so that the sound files are fully exploitable, also for phonetic purposes. The first spoken learner corpora were recorded on cassette tapes, which had to be digitised when more modern technologies became available. Nowadays, most recording equipment produces sound files which can be imported straight onto a computer. The recordings form the basis of transcription, that is, the transformation of an oral format into a written one. The transcription process can be performed via a simple text-editing program or

using more sophisticated tools, e.g. transcript editors like *CLAN*, *Praat* or *EXMARaLDA* (see Chapter 6, this volume), which, by showing the waveform or spectrogram of the audio files, can facilitate the transcription process. Varying degrees of precision can be aimed at when transcribing the data, from very basic orthographic transcription, which just seeks to reproduce the words uttered by the learners, to very detailed phonological and phonetic transcription, which shows how the words were actually pronounced by the learners; both orthographic and phonological/phonetic transcription can be more or less broad or narrow. For obvious reasons of economy (see below), most spoken learner corpora are made up of orthographic transcripts. The *LeaP* (*Learning Prosody in a Foreign Language*) corpus is one of the few exceptions: next to a word tier which contains an orthographic transcription of the data, it includes tiers for syllables, segments, tone and pitch (Gut 2012; see Section 3.3). The degree of delicacy of the transcription will mainly depend on the resources available (time and money) and the research purposes. If the corpus is primarily compiled to carry out lexical analyses of spoken interlanguage, then an orthographic transcription is probably sufficient; if, on the other hand, the main goal is to investigate learners' pronunciation and prosody, it might be worth investing in a narrower type of phonetic transcription. More often than not, however, a (spoken) learner corpus is compiled with no one particular research question in mind, or at least, with a view to allowing the larger community of linguists to benefit from it as well. In such cases, pragmatism may prevail over perfection (see McEnery et al.'s (2006) quotation earlier in this section) and the compilers may decide to keep the transcription relatively broad, not only to reduce the costs and efforts involved, but also in acceptance of the fact that a spoken learner corpus, however delicate its transcription, will never answer all of the questions that a syntactician, semanticist, phonetician or SLA specialist may want to study, and that the user of the corpus may therefore have to add a level of transcription him-/herself before embarking on a specific research project. It should be underlined at this stage that even a 'simple' orthographic type of transcription can be quite costly. According to Ballier and Martin (2013: 33), it is estimated that one word of 'simple' orthographic transcription costs about one euro. In terms of time, it was calculated within the framework of the *LINDSEI* project that each minute of learner speech requires some twenty to thirty minutes for transcription (including post-transcription checks).

In addition, transcribing speech verbatim is a complex undertaking. If this is true of any type of speech, it is all the more so of learner speech, which tends to be difficult to decode because of the many dysfluencies and errors (including pronunciation errors) that it contains and that have to be transcribed (Gilquin and De Cock 2011). This has been shown to lead to 'a higher degree of perceptual reconstruction by the transcriber in L2 than in L1' (Detey 2012: 234), possibly influenced

by the transcriber's L1 (Bonaventura et al. 2000), and to a substantially lower rate of inter-transcriber agreement (Zechner 2009). In other words, because transcribers sometimes hear different things when listening to learner speech, they may come up with different transcriptions of the same stretch of discourse. There are also more specific problems when transcribing spoken interlanguage, such as the issue of how to deal with deviant forms (e.g. *choregraphy* instead of *choreography* or *womans* instead of *women*). While reproducing the deviant form may prevent it from being extracted automatically from the corpus (if the researcher uses the standard form as a search item), normalising the form results in a loss of information (in the case of *choregraphy*, for instance, the possible influence of the L1 if the learner is French-speaking, as the equivalent French word is *chorégraphie*). Matters get even worse when mispronunciation results in a different word (e.g. *law* pronounced as *low* or *dessert* pronounced as *desert*) or when the word simply does not exist in the target language. Admittedly, confused pairs of words or invented words occur in written learner corpora too and can present problems for the automatic extraction of words, but in a written corpus it is the learner who selects a particular spelling, whereas in a spoken corpus it is the transcriber who is responsible for choosing a certain transcription.

As a final note on transcription, it must be said that attempts have been made to transcribe learner speech (semi-)automatically, either as a first step before manual correction (see Bonaventura et al. 2000) or with the aim of developing automatic speech recognition software applicable to non-native speech (e.g. Wang and Schultz 2003). However, it is fair to say that there is still a long way to go before spontaneous learner speech can be transcribed accurately in a fully automatic manner, so that for the next years or decades to come, researchers will probably have to go through the arduous and time-consuming process of manual transcription (unless they can have 'Turkers' do the work for them, see Evanini et al. 2010).[15]

In the case of longitudinal spoken learner corpora, as with longitudinal written learner corpora, the procedure has to be repeated at several points in time. Multimodal learner corpora, being made up of video-recorded speech, rely on some of the steps described above for their compilation. In comparison with spoken learner corpora, they require video recording of the speech event and may involve some sort of 'transcription' of the video as well (e.g. indication of the gestures made by the learner).

For all types of learner corpora, different options are available for the encoding of the final product (SGML, XML, etc.); for spoken and multimodal learner corpora that are distributed with their sound/video files, it is possible (and often desirable) to align the text transcript with the sound/

[15] Turkers are users of the crowdsourcing Amazon Mechanical Turk platform (www.mturk.com, last accessed on 13 April 2015) who get paid to perform (usually simple) tasks online.

video so that the two can be examined and queried simultaneously. Since these features are common to all corpora, native and non-native, they will not be further discussed here. As for the post-processing of learner corpora (lemmatisation, part-of-speech tagging, error annotation, phonetic annotation, etc.), this will be dealt with in Chapters 5 to 7 (this volume).

3 Representative studies

This section presents three studies which describe the design and collection of different types of learner corpora. The first one, by Lozano and Mendikoetxea (2013), deals with the compilation of a written learner corpus, while the other two concern the compilation of a spoken learner corpus. In the case of Jendryczka-Wierszycka (2009), the corpus can be defined as a mute spoken learner corpus, whereas in the case of Gut (2012) it is a speech learner corpus with speech–text alignment (see Section 2.1 on the distinction between mute and speech corpora). The three studies also represent a range of target languages: Spanish in Lozano and Mendikoetxea (2013), English in Jendryczka-Wierszycka (2009), and English and German in Gut (2012).

3.1 Lozano, C. and Mendikoetxea, A. 2013. 'Learner corpora and second language acquisition: The design and collection of CEDEL2', in Díaz-Negrillo, A., Ballier, N. and Thompson, P. (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: Benjamins, pp. 65–100.

Lozano and Mendikoetxea (2013) describe the compilation of *CEDEL2* (see above and Lozano 2009a), a cross-sectional corpus of L2 Spanish compositions written by English-speaking learners. They demonstrate that *CEDEL2* is a well-designed and carefully constructed corpus by showing how it follows the ten key design criteria set out by Sinclair (2005). Their main arguments are summarised here:

1. Content selection: the texts included in *CEDEL2* have not been selected on the basis of the language they contain; they are supposed to represent the use of learner Spanish under natural conditions.
2. Representativeness: *CEDEL2* represents a large sample of learners, from all proficiency levels and writing on a wide range of topics.
3. Contrast: a corpus of native Spanish, designed according to the same criteria as the learner corpus, allows for legitimate comparisons between native and non-native writing.
4. Structural criteria: *CEDEL2* is simply structured according to the writers' L1 (English or, for the native comparable corpus, Spanish) and the learners' proficiency level (beginner, intermediate or advanced).
5. Annotation: tags (in XML format) are stored separately from the texts (in raw text format).

6. Sample size: *CEDEL2* is made up of complete, unedited texts which may vary in length.
7. Documentation: detailed information about the structure of *CEDEL2*, as well as about the learners who contributed to it and their compositions, is available.
8. Balance: while limited to written language, *CEDEL2*, as a specialised corpus, is claimed to provide a good basis for the study of interlanguage phenomena.
9. Topic: the twelve composition topics writers could select from are assumed to be varied enough to elicit a large range of linguistic phenomena.
10. Homogeneity: texts submitted online that do not satisfy the design criteria are not included in *CEDEL2*.

In accordance with the need for metadata underlined in Section 2.2, Lozano and Mendikoetxea (2013) also explain how they have collected information through two forms to be completed by each participant: (i) a learning background form, which asks for writers' personal details (age, gender, institution, etc.), linguistic details (L1, parents' L1, stay in Spanish-speaking countries, etc.) and self-rated proficiency in speaking, listening, reading and writing (in Spanish and in other languages they may have learned); (ii) a composition form, which includes the composition itself, but also information about background research (did the writer conduct any research before writing the composition, and if so, how long and by what means?), composition title (among the twelve possible topics), writing location (in class, at home or both) and writing tools if any (dictionaries, spell-checkers, native help, etc.).

In addition to considerations concerning the design of *CEDEL2*, Lozano and Mendikoetxea (2013) describe the current state of the corpus (which included about 750,000 words produced by some 2,500 participants in March 2011 but continues to be expanded, with an intended target of 1 million words), the distribution of the data it contains and the preliminary post-processing it has undergone. What is particularly interesting about this corpus is that, unlike most learner corpora which are collected in a small number of environments (often depending on the location of the researchers involved in their compilation, see Section 2.3), *CEDEL2* was collected via a web application, through which speakers of Spanish all over the world were invited to contribute. This results in a wide range of writer profiles, using different varieties of (learner and native) Spanish. Another advantage of the corpus is that it comes with an assessment of each learner's proficiency level. This is done via the learning background form, which requires learners to self-rate their proficiency in the four skills (see above). In addition, learners' actual proficiency is assessed by means of an independent and standardised placement test, the University of Wisconsin placement test, which the participants can take online. As

will be noted in Section 4, proficiency is a variable that is often lacking (or determined with insufficient precision) in learner corpora, and this double proficiency measure in *CEDEL2* is therefore a major asset (also because, as suggested by the authors, it allows the comparison between self-rated and real proficiency). The compilation of a native counterpart to the learner corpus should be underlined as well, as it makes it possible to compare native and non-native writing using data that are fully comparable since they were collected according to the same design criteria.

Finally, it is noteworthy that the paper is written from an SLA perspective and that *CEDEL2* is designed to answer questions that are (also) of interest to SLA researchers. The paper and the corpus therefore represent a laudable attempt to bring learner corpus research and SLA closer together (see Chapter 14, this volume, on the relation between the two fields), on the grounds that 'if corpus-based research is going to make a significant contribution to the field of SLA, new, well-designed corpora need to be made available to the research community' (p. 89).

3.2 Jendryczka-Wierszycka, J. 2009. 'Collecting spoken learner data: Challenges and benefits. A Polish L1 perspective', in Mahlberg, M., González-Díaz, V. and Smith, C. (eds.), *Proceedings of the Corpus Linguistics Conference, University of Liverpool, UK, 20–23 July 2009*.

Written by one of the partners in the *LINDSEI* project, Jendryczka-Wierszycka's (2009) paper has the interesting feature that it not only describes the compilation of a component of *LINDSEI* (the Polish component), but it also underlines the many challenges one can face when collecting spoken learner data. Jendryczka-Wierszycka starts by introducing the project as a whole, meant as a spoken counterpart to *ICLE*.[16] Being a multi-L1 learner corpus, *LINDSEI* is made up of several subcorpora that are each compiled according to the same principles, which ensures comparability across the different subcorpora. These principles include the fact that the data consist of informal interviews and that the participants are advanced foreign language learners of English. The structure of the interviews, in three parts, also needs to be adhered to: after choosing one among three set topics and talking about it for a few minutes, the learners answer questions about what they have just said and about more general topics like hobbies or life at university, and finally they are asked to describe a four-picture cartoon. In addition, a learner profile questionnaire has to be completed by every learner who contributes to *LINDSEI*. The questionnaire gathers information about the learner (age, gender, stay in an English-speaking country, other foreign languages known, etc.) and also includes the learner's consent for his/her data to be used for research purposes.

[16] See the *LINDSEI* handbook (Gilquin et al. 2010) for more details.

Among the challenges mentioned by Jendryczka-Wierszycka, the first one has to do with student recruitment. She notes that '[s]ince getting people's time, even for the sake of science, is well-known not to be the easiest task if there is no money involved, we had low expectations of the number of volunteers for our corpus'. In the end, Jendryczka-Wierszycka was able to recruit fifty-one students by appealing to them during lectures and by announcing a prize draw among the volunteers. In the description of these students' characteristics, which are also summarised in an appendix, another problem is alluded to, which is valid for all the *LINDSEI* subcorpora (and many other learner corpora too, see Section 4), namely the identification of the learners' proficiency level. In *LINDSEI*, all learners should be in their third or fourth year at university, and on this basis are expected to be (upper-intermediate to) advanced learners of English. However, Jendryczka-Wierszycka rightly points out that this 'may be a faulty assumption as the level naturally differs from one university to another even within one country, not to mention university level differences worldwide'. In the particular case of Polish learners, she adds that the quality of English classes that are taught in different schools across Poland is so uneven that the number of years of English at school cannot be a good indication of the learners' proficiency either.

Another major challenge that is described at length in the paper is the transcription of the data. Although *LINDSEI* comes with its own transcription guidelines, which are outlined on the project website,[17] Jendryczka-Wierszycka recognises that transcribing the interviews (which was done with the help of the *SoundScriber* software)[18] was not an easy task. Besides technological difficulties (one of the recorded interviews would not play back), some passages were unintelligible, due, among other things, to overlapping speech and external noises, and certain items (like fillers) proved particularly hard to transcribe consistently. Jendryczka-Wierszycka explains that in the transcription process she was helped by a group of M.A. English linguistics students. While this reduced the time of transcription considerably, this also involved training of the transcribers, good coordination of the group and correction of the transcripts by the coordinator. All in all, Jendryczka-Wierszycka notes that each interview (lasting about fifteen minutes) required an average of five hours to be transcribed, which included listening to the sound file at least twice and making a final check of the transcript. The transcription phase took seven months, as against two months for the recording of the interviews.

[17] See www.uclouvain.be/en-cecl-lindsei.html (last accessed on 13 April 2015). Note that the conventions described by Jendryczka-Wierszycka correspond to an earlier version of the transcription guidelines. Thus, the use of square brackets and vertical alignment to signal overlapping speech has now been replaced by the tag <overlap /> which is inserted in each of the two overlapping utterances, while foreign words are now marked by means of <foreign> … </foreign>, in lieu of italics.

[18] www-personal.umich.edu/~ebreck/code/sscriber/ (last accessed on 13 April 2015).

The paper also provides some useful statistics about the composition of the Polish component of *LINDSEI*, as well as a summary of two case studies published by the same author: one on vague language and the other on discourse markers. Finally, it briefly describes the author's attempt to apply a part-of-speech (POS) tagger designed for the annotation of native data, *CLAWS*, on learner data, underlining the problems that appeared and proposing some possible solutions to them. The paper ends with the hope that the corpus can become a useful resource not only for linguists but also for language teachers and translators.

3.3 Gut, U. 2012. 'The LeaP corpus. A multilingual corpus of spoken learner German and learner English', in Schmidt, Th. and Wörner, K. (eds.), *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: Benjamins, pp. 3–23.

This paper by Gut (2012) describes the compilation and annotation of the *LeaP* corpus, a corpus of spoken learner German and learner English totalling over twelve hours of recording which, as its name indicates (*LeaP* stands for ***Lea****rning* ***P****rosody in a Foreign Language*), is primarily aimed at studying the second language acquisition of prosody.[19] Unlike a learner corpus like *LINDSEI*, which is made up of data from a rather specific learner population (young, relatively advanced learners in their third or fourth year at university), the *LeaP* corpus was designed to be as representative as possible of the German and English interlanguages, including data from a wide range of learners (seventeen mother-tongue backgrounds, ages from 18 to 60, first contact with the target language from 1 to 33 years of age, etc.). Certain groups of learners were selected so as to answer predefined research questions, e.g. a group of very advanced, native-like learners 'to test what type of ultimate phonological attainment is possible' (p. 5) and a group of learners who were recorded before and after a course in pronunciation to measure the impact of formal training. In addition, a few native speakers of German and English were recorded as a baseline. The corpus also includes different types of speech: free speech collected in a semi-structured interview setting, prepared reading of a story, semi-spontaneous retelling of this story and reading of a list of nonsense words. To ensure the high quality of the audio files and their possible phonological exploitation, the recordings took place in a sound-treated chamber.

The paper also describes the detailed transcription and annotation of the corpus. Being a speech (rather than mute) learner corpus (see Section 2.1), the *LeaP* corpus is distributed with its audio files in the form of time-aligned phonological and phonetic transcriptions, where the transcription is linked to the corresponding part of the recording by means of time-stamps set at the beginning and end of each relevant unit (word,

[19] The *LeaP* corpus is freely available for research purposes at http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI671070%23 (last accessed on 13 April 2015).

syllable, phoneme, etc.). This text-to-sound alignment makes it possible, by means of appropriate software like *Praat* (which was used to annotate the *LeaP* corpus), to have simultaneous access to transcription and sound – the latter represented in *Praat* by a waveform, a spectrogram and a pitch track. Annotations are similarly aligned, with each type of annotation constituting an individual tier. The *LeaP* corpus contains up to eight tiers, six of them carried out manually and the other two added automatically:

1. Phrase tier: division into intonation phrases, with indication of interrupted intonation phrases, unfilled pauses, hesitation phenomena, elongated phonemes and some non-speech events (noise, breath, laughter).
2. Word tier: orthographic transcription and annotation of the beginning and end of each word.
3. Syllable tier: broad phonetic transcription and annotation of the beginning and end of each syllable.
4. Segment tier: vocalic intervals, consonantal intervals and intervening pauses.
5. Tone tier: pitch accents and boundary tones.
6. Pitch tier: phonetic properties of pitch range (initial high pitch, final low pitch, intervening pitch peaks and valleys).
7. POS tier: automatic annotation of parts of speech.
8. Lemma tier: automatic annotation of lemmas.

For each minute of recording, an average of 1,000 events were annotated. This was done by six annotators who received training in annotation (criteria for the division into intonation phrases, annotation schemes, etc.). The reliability of the manual annotation was measured by means of inter-annotator agreement (to what extent do all annotators agree on the annotation of the same recording, i.e. how stable are the annotations?) and intra-annotator agreement (to what extent does an annotator agree with him-/herself when annotating the same recording twice, i.e. how reproducible are the annotations?). Both measurements yielded differing results, depending on the complexity of the task, so that '[t]he higher the number of different categories in an annotation scheme, the lower the agreement' (p. 11). Experience with annotation was also shown to have a positive influence on the reproducibility of the annotations.

The corpus includes metadata (which Gut refers to as 'non-linguistic annotation') comprising a wide range of information such as date and place of the recording, gender, age and native language of the learner, duration and type of stays abroad, prosodic knowledge, motivation and attitude towards the target language, importance self-attributed to competence in pronunciation, and even experience and ability in music and in acting. These metadata are integrated into the corpus data, which are in an XML-based format specially developed for the

corpus, the Time Aligned Signal data eXchange (TASX) format. Analysis of the corpus is possible on the basis of this format or through conversion to other file formats compatible with various search tools. The paper ends with an illustration of how the *LeaP* corpus can be used to explore fluency in learner German and learner English, thus convincingly demonstrating that 'a corpus with rich annotations and a standardised data format despite having a relatively small size offers numerous possibilities of testing previous concepts and claims in L2 acquisition research' (p. 20).

4 Critical assessment and future directions

In this section, we will reconsider the three core issues outlined in Section 2 from a more critical perspective, pointing to certain limitations of the typology, design and collection of learner corpora, and adding some suggestions for possible future developments in these three areas.

In terms of learner corpus typology, it must be recognised that there is a striking imbalance between the types of learner corpora that are currently available. There are more written than spoken corpora, more general than specific corpora, more corpora of English than any other language, and more corpora containing cross-sectional than longitudinal data. This imbalance is a partial reflection of the ease with which data can be collected: the transcription of spoken texts is more time consuming than the keyboarding, scanning or electronic collating of written texts, learners of English for general purposes are more numerous than learners of English for specific purposes and, especially, learners of other languages, and multiple data collection from the same learners requires heavier logistics than one-off data collection. Because of researchers' tendency to collect data from learners who are easy to reach (see Section 2.3), we also notice a predominance of learner corpora representing relatively advanced university students, often majoring in the target language, whereas beginners and young learners are less often represented.[20] While all sorts of practical constraints may make it difficult to collect certain types of learner data, a collective effort should nonetheless be made not only to enlarge our repertoire of learner corpora but also to diversify it, as the types of learner corpora that are lacking are bound to provide invaluable information about interlanguage (see, e.g., Chapter 17, this volume, on longitudinal corpora). In particular, two types of learner corpora that are currently extremely rare seem to hold special promise: multimodal and local learner corpora. While the

[20] An exception is the *International Corpus of Crosslinguistic Interlanguage* (*ICCI*), which consists of data produced by beginner to lower-intermediate learners of English; see http://cblle.tufs.ac.jp/llc/icci/ (last accessed on 13 April 2015).

former open up a whole range of new possibilities in the study of interlanguage by adding the picture to the text and sound, the latter invite teachers and students alike into the field of learner corpus research by making them both providers and beneficiaries, thus resulting in learner corpora being directly useful to those for whom, ultimately, they have been compiled. It might also be interesting to collect more multilingual mono-L1 learner corpora like the *USP Multilingual Learner Corpus* (Tagnin 2006), which contains data produced by Brazilian learners in different foreign languages. Unlike monolingual multi-L1 learner corpora (like *ICLE*), which make it possible to identify the universal vs L1-specific problems learners encounter when learning a specific target language, such corpora offer the opposite perspective and show what difficulties learners of a given mother-tongue background experience when learning a foreign language.

As for the types of learner corpora that are likely to appear in the near future, it seems as if we might be moving towards multidimensional learner corpora (or perhaps databases) that contain several subsets of data designed according to similar criteria but representing different language varieties (including a native counterpart), different media, different genres, different tasks, different acquisition settings, etc., so that legitimate comparisons can be drawn and reliable statements can be made about the possible influence of these factors. In view of the growing awareness of the internal variation of learner corpora and the individuality of learners (see Gilquin and Granger 2015), it may be expected that, whenever possible, these data will be collected from the same learners, who will be required, say, to produce spoken and written interlanguage, write an essay in timed and untimed conditions, describe a picture in the target language and in their mother tongue, or participate in linguistic experiments whose results will be recorded and then confronted with their more natural production. The *ASU Corpus* (Hammarberg 2010) is a first step in this direction, since it contains spoken and written data produced at regular intervals and through various tasks by the same learners of Swedish, and also includes a native Swedish counterpart built in a similar way (though, obviously, with different informants).

Moving on to design criteria, we can applaud the fact that most learner corpora are built according to some (more or less strict) criteria and, above all, that they are often accompanied by information about the profile of the learners who contributed to the corpus and about the circumstances in which the data were contributed. As regards the choice of variables that are recorded, we can only agree with Granger (2004: 126) that 'there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control'. One variable that is of crucial importance but whose identification has often been less than optimal, however, is that of proficiency. In *ICLE*, for example, proficiency is determined by means of external criteria like age and number of years of English at

university (Granger 1998a: 9). Yet, several linguists have objected that such external criteria do not necessarily offer an accurate representation of a learner's proficiency level (e.g. Thomas 1994; Pendar and Chapelle 2008; see also Section 3.2)[21] and that a more objective measure of proficiency should therefore be provided. As explained by Carlsen (2012: 165), there are two ways of doing this: through learner-centred or text-centred methods. The former (which also include external criteria like age) determine proficiency by examining the learner's characteristics, for example by having him/her take an independent, standardised proficiency test like the University of Wisconsin placement test for *CEDEL2* (see Section 3.1) or the Oxford Quick Study Placement Test for *MiLC* and *WriCLE*[22] (cf. Mediero Durán and Robles Baena 2012). Text-centred methods, on the other hand, examine the text itself to establish proficiency, as was done with the *ASK* corpus of Norwegian L2, whose individual texts were rated according to the Common European Framework of Reference for Languages (Council of Europe 2001) – see Carlsen (2012). In effect, such measures provide 'a description of the quality of one single essay produced by each individual learner, rather than a more independent assessment of the learners' overall proficiency' (Thewissen 2013: 79). Each method has its disadvantages: assessment on the basis of an independent placement test may not reflect the level of the text since 'one and the same learner may perform slightly differently from one day to the next or from one test to another' (Carlsen 2012: 168), whilst assessment on the basis of the corpus texts runs the risk of circularity as the texts are rated according to their linguistic features and then analysed linguistically to say something about the learner's proficiency (see Hulstijn 2010). However, both types of measure constitute an improvement over impressionistic evaluation of the proficiency level and should thus be encouraged in the design of learner corpora. In this respect, we can certainly welcome the fact that some of the most recent learner corpus projects have integrated an objective proficiency score in their design (see above examples).

Another variable that could usefully be improved is that of exposure to the target language. While traditionally this has been limited to a description of the acquisition setting (foreign or second language environment), the number of years of instruction in the target language and the time spent in a target-language country, there are many other elements that could have an influence on learners' degree of exposure, especially in today's high-tech world, where resources and contacts in other languages

[21] This has been confirmed by a CEFR (Common European Framework of Reference for Languages; Council of Europe 2001) evaluation of a sample of essays from *ICLE* which, on the basis of the external criterion of number of years of English at university, were supposed to represent the same level, but whose actual scores ranged from B2 and lower (40%) to C2 (less than 20%) (see Granger et al. 2009: 11–12). The same sort of contrast emerged from the CEFR rating of a sample of *LINDSEI* (see Gilquin et al. 2010: 10–11).

[22] *WriCLE* stands for *Written Corpus of Learner English*; see www.uam.es/proyectosinv/woslac/Wricle/ (last accessed on 13 April 2015).

are at a learner's fingertips. If a learner spends all his/her free time watching TV series or playing multiplayer online games in English, for instance, this is likely to have more impact on his/her knowledge of English than a two-week holiday in the UK with his/her family.[23] In an attempt to approximate the learner's full experience with the target language, information could be gathered about incidental learning in everyday life through reading, entertainment, social networking, etc., along the lines of the questionnaire found in Schmitt and Redwood (2011: 206–7) or, for a much more detailed version, Freed et al.'s (2004a) 'language contact profile'.

Probably as important as the refinement of certain variables, however, is the urgent need to standardise the metadata that come with learner corpora. This would not necessarily mean that all learner corpora should include exactly the same metadata, but if they do include a certain type of information, it should follow a specific format (e.g. precise age in years and months rather than ranges of years), so that results related to these variables can be compared across different learner corpora. General initiatives have been undertaken to make recommendations about the selection and presentation of metadata (e.g. Dublin Core Metadata Initiative)[24] but, to date, similar initiatives specifically concerned with learner corpus metadata are still lacking.

Standardisation is key to the successful compilation and encoding of learner corpora too. In addition to adopting good practices such as those recommended by Sinclair (2005) – see Section 3.1 – it would be desirable, in order to increase compatibility between different learner corpora, to follow the same guidelines to represent text in electronic format (e.g. form of corpus headers, transposition of typographical features, indication of quotations, conventions of transcription). Again, such initiatives exist, like the Text Encoding Initiative (TEI),[25] but few learner corpora so far have applied these standards (for an exception, see *BACKBONE*, a corpus whose annotation relies on TEI-compliant XML and which includes a number of interviews with non-native speakers of English; see Kohn 2012).[26] Equally important in order to allow the community to benefit from a learner corpus are the availability of detailed documentation describing the compilation of the corpus and, of course, the accessibility of the corpus data (including sound/video files if appropriate) in the first place, so that studies based on these can be replicated and more studies can be undertaken. This is not necessarily obvious: in Schiftner's (2008) survey, documentation of the learner corpus projects turned out to be 'scattered and often scarce' (p. 48), and only half of the learner corpora

[23] See the *ReCALL* special issue edited by Cornillie et al. (2012) on the role of digital games for language learning.

[24] http://dublincore.org/ (last accessed on 13 April 2015).

[25] www.tei-c.org/ (last accessed on 13 April 2015).

[26] Of special interest in this respect is the *SACODEYL Annotator*, which makes it possible to create XML TEI-compliant annotations (see Pérez-Paredes and Alcaraz-Calero 2009).

were publicly available (though not always free of charge). It might be that one of the reasons (partly) accounting for certain researchers' tendency not to disseminate their learner corpora has to do with the subjects involved in the collection of such corpora, viz. mainly young people, whose participation may require a special set-up.[27] Thus, while every effort should be made, whatever the type of corpus, to follow ethical procedures and maintain anonymity, this is all the more crucial with young subjects, whose safety and welfare should be preserved at all costs, during but also after the data collection. This can be particularly problematic with multimodal learner corpora, which capture the participant's voice and face, and which should therefore be anonymised appropriately to conceal his/her identity – while bearing in mind that their complete anonymisation may limit the types of analyses that they allow (e.g. facial feature analysis; see Adolphs and Knight (2010: 43–4) on the anonymisation of multimodal corpora). Obtaining consent for the data to be collected and used for research purposes may also be a complex task, as consent may have to be granted by a parent or guardian if the subject is under-age and/or by a teacher, a headmaster or even a higher-level authority if the data collection takes place in a school; if the learner is old enough to give consent him-/herself, the researcher should make sure that s/he fully understands the nature of the research before signing the consent form.

Finally, looking at what might be the learner corpus of the future, it is likely that new technologies will have a major role to play in how it is collected (see also Chapter 18, this volume). Learner corpora of computer-mediated communication (like *Telekorp*) are an early illustration of this. Multiplayer online games, mentioned above for their possible impact on learners' knowledge of foreign languages, as well as the recent trend of massive open online courses (MOOCs), can also be used as a way of collecting learner data. Besides computers, data could be collected via smartphones and tablets, whose popularity among young people would contribute to the non-intrusive character of the process. These technologies, because they are part and parcel of the everyday life of the new generation of learners, make it possible to move corpus collection away from the academic setting and into a more natural environment, thus coming closer to the ideal of genuine communications that should be included in a corpus (Sinclair 1996; see Section 1). In a way, this is a natural development, since the knowledge of an L2 is an inherent feature of individuals, which is with them all the time and not just during the (limited) periods when they are actually learning it. It is therefore only normal that learner corpora, if they are to serve as repositories of interlanguages, should strive to reflect learners' full experience with the L2 as accurately as possible.

[27] See, e.g., the *Guidelines for Research with Children and Young People* published by the National Children's Bureau (Shaw et al. 2011).

Key readings

O'Keeffe, A. and McCarthy, M. (eds.) 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

The second section of this handbook (pp. 29–103), entitled 'Building and designing a corpus: What are the key considerations?', covers the basics of corpus compilation. It provides a step-by-step guide for how to build a spoken, written, small specialised and audio-visual corpus, and a corpus that represents a certain language variety (like American or academic English). Although the chapters do not deal specifically with learner corpora, they provide information that is relevant to learner corpus compilation, and also include a few references to learner corpora (especially in the chapter on small specialised corpora).

Granger, S. 1998a. 'The computer learner corpus: A versatile new source of data for SLA research', in Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 3–18.

This is one of the founding texts that introduced the learner corpus, situating it within the broader fields of corpus linguistics, second language acquisition and foreign language teaching, describing the main language- and learner-related design criteria relevant to learner corpus building (with an illustration by means of *ICLE*) and pointing to some of the difficulties involved in compiling a learner corpus.

Pravec, N. A. 2002. 'Survey of learner corpora', *ICAME Journal* 26: 81–114.

Though slightly dated, this survey of learner corpora of (written) English provides detailed information about the attributes (size, availability of learner background information, format, etc.) of the ten corpora that were available at the time, with a view to helping researchers select the corpus that is the most suitable for their purposes.

Tono, Y. 2003. 'Learner corpora: Design, development and applications', in Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, *UCREL Technical Paper 16*. Lancaster University, pp. 800–9.

This paper provides a good overview of the considerations that should be kept in mind when compiling (and analysing) a learner corpus. It also provides a survey of some twenty learner corpora and their features (including size, types of subjects and texts), as well as some directions for the future.

Schiftner, B. 2008. 'Learner corpora of English and German: What is their status quo and where are they headed?', *Vienna English Working Papers* 17(2): 47–78.

In addition to a detailed description of twenty-six English and five German learner corpora, the paper considers developments in English and German learner corpus compilation, with special emphasis on problems related to design and accessibility, and it offers useful suggestions for the compilation of English and German learner corpora, some of which might be relevant to other target languages as well.