17

Developmental patterns in learner corpora

Fanny Meunier

1 Introduction

The understanding and description of learners' developmental patterns have been at the core of second language acquisition (SLA) research for about forty years now.[1] Kramsch (2000: 315) defines SLA as being concerned 'with the process by which children and adults acquire (learn) second (third or fourth) languages in addition to their native language' and adds that SLA is interested 'in the nature of these learners' language and their development throughout life'. Similarly, Ortega (2012: 133) writes that one of the key issues in SLA is to 'shed light on how interlanguage development proceeds over time, from initial emerging representations to a full-blown, mature system of the new language'. As evidenced by the two preceding quotes, the correlated notions of progress and time are central in SLA.

Learner corpora are one possible data type that can be used to analyse interlanguage development. Granger (2008a: 338) defines learner corpora as 'electronic collections of (near-) natural foreign or second language learner texts assembled according to explicit design criteria'. Although learner corpus research (LCR) has – from its onset – paid specific attention to the design criteria of learner corpora and to the collection of metadata (see Granger 1998a; Chapter 2, this volume), an even better control of some of the variables at play in SLA has only recently become a central concern in LCR. This focus on variables (be it in learner corpus data collection or analysis) combined with the use of ad hoc inferential statistics (see Chapter 8, this volume) now makes it possible for learner corpus specialists to use and analyse variables as dependent variables, potential

[1] Ortega (2013: 2) suggests taking Selinker's (1972) field-defining publication as a convenient official marker of the disciplinary beginnings of SLA.

predictors for the linguistic features of texts, or as dynamic factors that should be taken into account in the learning process.

Another key feature of LCR is that the data can be stored electronically and (pre-)processed with the help of (semi-)automatic corpus tools. This computational treatment makes it possible to analyse the production data of numerous learners (as opposed to more traditional SLA studies which typically involve few participants). As a result, LCR studies are able to replicate earlier SLA studies but on much larger populations. Murakami's (2013a) replication study on the order of acquisition of morphemes is one example of replication studies carried out within an LCR paradigm.

The aim of the present chapter is to illustrate how current learner corpora and LCR methods can be used to track development in the acquisition/learning of a language other than the mother tongue, both at group and individual level.

2 Core issues

2.1 Longitudinal vs pseudo-longitudinal

2.1.1 Study design

Unlike cross-sectional studies which examine the language behaviour of a group or groups of language learners at a single point in their development, longitudinal studies are defined by Johnson and Johnson (1999) as studies which examine the language behaviour of one or more subjects as that behaviour develops over time. Longitudinal study designs thus follow the same individual(s) over time and collect language-related data from this/these individual(s) at different points in time. Longitudinal research is defined as 'emphasizing the study of change and containing at minimum three repeated observations on at least one of the substantive constructs of interest' (Ployhart and Vandenberg 2010: 97). This minimum of three data-collection points makes it possible to fit a developmental line and visualise potential effects on that line (linear progression or regression, U- or reversed U-shaped behaviour).[2] Obviously, the more collection points, the more refined the interpretation of the development can be. When longitudinal data are collected at numerous intervals, the notion of 'dense data collection' is often used, especially when the corpus is accompanied by rich metadata. One example of a dense corpus (although devoted to the acquisition of the mother tongue) is the *Human Speechome Corpus*, which contains about 10 million words of transcribed recordings of child–caregiver interactions in natural contexts corresponding to about 120,000 hours of audio and 90,000 hours of video, capturing an

[2] One of the limitations of two-wave studies (i.e. with only two data collection points) is that any and all change from Time 1 to Time 2 will by default be linear (i.e. a straight line), which makes it impossible to determine a more precise form of change (Singer and Willett 2003: 9–10).

estimated 70 per cent of the child's first three years of waking hours[3] (see Roy et al. 2012 for further details on the corpus). The trade-off between the density of the data and the number – or representativeness – of the subjects whose language data is collected is still an unresolved issue in LCR (see Section 4 below for further comments).

Collecting longitudinal learner corpus data is a real challenge as it is both time consuming and requires much planning ahead. Another non-negligible problem, in today's stuck-in-fast-forward, competitive research agenda and publication timing, is that the analysis can only start when the entire data collection is over. Myles (Chapter 14, this volume) also mentions the prohibitive costs of collecting longitudinal data and the fact that research funders do not like to commit resources for very long periods of time. Attrition (i.e. the sometimes significant number of participants dropping out before each data-collection point) is another major challenge in dealing with longitudinal data, especially when it comes to learners of a second/foreign language whose learning histories cannot be predicted for certain (such as in the case of students who do not turn up for tests or data collections, decide to give up their studies, or change school or option). Such difficulties in collecting data mean that the high demand for longitudinal learner corpora is – quite unsurprisingly – met with few research teams collecting such data types.

When it is not possible to follow the same individuals over time, researchers can carry out a comparison of cross-sectional studies of different groups of learners at different developmental stages. Using such an approach yields what Johnson and Johnson (1999) call a pseudo-longitudinal effect as the learners' productions compared do not come from the same learners, hence the use of the 'pseudo' prefix. The 'time' variable (which can be measured directly in longitudinal studies) is thus measured in pseudo-longitudinal designs by a proxy such as age or proficiency level. In other words, instead of following one group of students through every step of their progress in acquiring a target language, researchers compare several groups of learners displaying different levels of proficiency. Those groups, whilst containing different learners, nonetheless often share a number of characteristics in order to warrant some homogeneity (e.g. same mother-tongue background or same learning context). Such study designs are thus called pseudo-longitudinal (Johnson and Johnson 1999; Huat 2012: 197) or quasi-longitudinal (Granger 2002; Thewissen 2013).

The longitudinal and pseudo-/quasi-longitudinal designs are graphically summarised in Figures 17.1a and 17.1b. They will be further illustrated with concrete examples in Section 3.

It is important to stress that individual trajectories can only be accessed indirectly in quasi- or pseudo-longitudinal studies as the data-collection

[3] To my knowledge, there is no such equivalent dense learner corpus available.

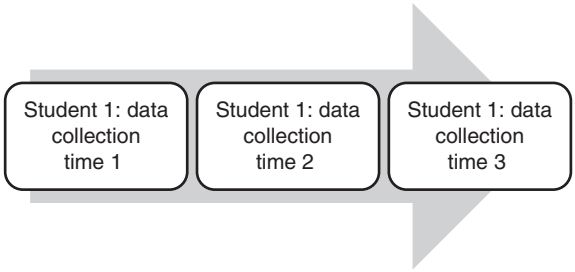

Figure 17.1a Graphical representation of a longitudinal study design

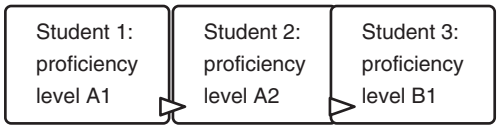

Figure 17.1b Graphical representation of a pseudo- or quasi-longitudinal study design (proxy for time used here: proficiency level)

procedure is inherently cross-sectional. In such designs, only group development can be measured. Individual variation within each group or sub-group can, however, be analysed.

With longitudinal study designs, in contrast, group progress, individual variation within groups and individual trajectories can be analysed. This requires the use of, for instance, multi-level modelling – also referred to as hierarchical linear modelling or mixed-effects models (see Raudenbush and Bryk 2002; Baayen et al. 2008; Cunnings 2012; Chapter 8, this volume). Multi-level modelling allows a variety of predictors to be analysed, with 'time' being a key predictor in longitudinal studies: do participants become more proficient as time goes by and, if so, how strong is the effect of time? Such statistical modelling can be applied to individuals within groups as well as to individuals as individuals, by analysing both endpoints and trajectories.

2.1.2 Learner corpora for developmental studies

Longitudinal learner corpora, which follow the same set of participants over multiple data-gathering sessions, are not very numerous.[4] They include, among others:

- five subcorpora of the *FLLOC* (*French Learner Language Oral Corpora*) project,[5] viz. the *LANGSNAP Corpus*, the *Newcastle Corpus*, the *Progression Corpus*, the *Brussels Corpus* and the *Salford Corpus*

[4] For a survey of existing longitudinal learner corpora, see www.uclouvain.be/en-cecl-lcworld.html (last accessed on 13 April 2015).

[5] See the project webpage at www.flloc.soton.ac.uk/index.html (last accessed on 13 April 2015) for a description of the *FLLOC* subcorpora. Each subcorpus is described in detail (types of learners, tasks, transcription conventions, headers used, database content).

- some of the subcorpora of the *InterFra* (*Interlangue française*) corpus[6]
- the longitudinal subcorpus of the *Corpus Ecrit de Français Langue Etrangère* (*CEFLE*)[7]
- the *Barcelona English Language Corpus* (*BELC*)[8]
- some subcorpora of the *Corpus of Learner German* (*CLEG13*)[9]
- the *Telecollaborative Learner Corpus of English and German* (*Telekorp*), a special kind of longitudinal corpus which contains bilingual contrastive learner data of computer-mediated communication between native German speakers and American non-native speakers of German (Belz and Vyatkina 2008)
- the *LONGDALE* (*Longitudinal Database of Learner English*) project (for further details, see Section 2.2).[10]

When learner corpora are collected according to a pseudo-longitudinal design, a different but comparable set of participants is used for each data-gathering session. The different samples of participants recruited for each separate data collection are comparable for certain attributes relating to the study carried out (language studied, type of instructional setting, etc.) and typically differ in terms of age or proficiency level. The number of learner corpora collected using a pseudo-longitudinal design are more numerous and include, among others:

- four subcorpora of the *FLLOC* project mentioned earlier, viz. the *Young Learners Corpus*, the *Linguistic Development Corpus* and the *UEA Corpus*
- some of the *InterFra* subcorpora
- the *Cambridge Learner Corpus* (*CLC*)[11]
- the *Cambridge English Profile Corpus* (*CEPC*)[12]
- the *National Institute of Information and Communications Technology Japanese Learner English Corpus* (*NICT JLE*)[13]
- the *Japanese EFL Learner Corpus* (*JEFLL Corpus*) (see Tono 2000b)
- the *Spanish Learner Language Oral Corpora* (*SPLLOC*)[14]
- some *CLEG13* subcorpora.

Table 17.1 provides a brief summary of the corpora listed above, their design, the targeted language studied and the data type covered.

The corpora listed in Table 17.1 have been collected in the framework of projects explicitly aiming at developmental studies. It should be added

[6] For a detailed description, visit www.su.se/romklass/interfra (last accessed on 13 April 2015).
[7] http://projekt.ht.lu.se/cefle/information/le-sous-corpus-longitudinal/ (last accessed on 13 April 2015).
[8] www.ubgral.com/corpus.html (last accessed on 13 April 2015).
[9] http://korpling.german.hu-berlin.de/public/CLEG13/CLEG13_documentation.pdf (last accessed on 13 April 2015).
[10] www.uclouvain.be/en-cecl-longdale.html (last accessed on 13 April 2015).
[11] www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB (last accessed on 13 April 2015). See also Chapters 22 and 23 (this volume) for more information on the corpus.
[12] www.englishprofile.org/index.php/corpus (last accessed on 13 April 2015).
[13] http://alaginrc.nict.go.jp/nict_jle/index_E.html (last accessed on 13 April 2015).
[14] www.splloc.soton.ac.uk (last accessed on 13 April 2015).

Table 17.1. Select list of learner corpora collected according to a longitudinal or pseudo-longitudinal design

Corpus name	Longitudinal design	Pseudo-longitudinal design	Target language	Data type
LONGDALE (several subcorpora)	✓		English	Oral and written data
FLLOC (several subcorpora)	✓	✓	French	Oral data
InterFra (several subcorpora)	✓	✓	French (some other L2s are also present in some of the subcopora, such as Swedish, Spanish, English and Italian)	Oral and written data
CLC		✓	English	Written responses to tests of English for Speakers of Other Languages
CEPC		✓	English	Oral and written data
NICT JLE		✓	English	Oral data
JEFLL		✓	English	Written data
CEFLE (longitudinal subcorpus)	✓		French	Written data
BELC	✓		English	Oral and written data
SPLLOC		✓	Spanish	Mainly oral data
CLEG13	✓	✓	German	Written data
Telekorp	✓		German and English	Computer-mediated communication

that some pure cross-sectional learner corpora have also been used to track development. Despite the fact that pure cross-sectional research cannot tell us anything about intra-individual or inter-individual change processes, the cross-sectional design initially adopted in some learner corpora has been supplemented by post hoc proficiency level assessment of the subjects, thereby allowing some researchers to perform pseudo-longitudinal research. For instance, some of the essays collected in the *International Corpus of Learner English* (Granger et al. 2009) have been assessed for proficiency after collection. This post hoc assessment variable made the adoption of a pseudo-longitudinal design possible, as for instance in Thewissen's (2013) study on accuracy developmental patterns. This approach was also adopted in one of the *FLLOC* subcopora, viz. the *Reading Corpus*, where students from an initially cross-sectional corpus (34 secondary school students, aged 16, who had all been learning French for five years, receiving four 35-minute lessons a week) were shown to display huge variation in terms of oral proficiency (from very little contribution to conversation to a level comparable with native speakers of French).

2.2 Analysing group and/or individual trajectories in LCR: some key considerations

Learners typically progress through various stages of proficiency, and in a vast majority of cases there is an overall positive correlation between the time and effort spent learning an additional language and the evolution of proficiency in the target language. This is clearly apparent in – and also intrinsically characteristic of – most educational settings, where a majority of learners gradually step from one level to a slightly higher one and tend to become increasingly proficient throughout their instructional path. The descriptors for linguistic competence included in the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) or in the ACTFL[15] Assessment of Performance toward Proficiency in Languages (AAPPL) clearly illustrate this approach to proficiency development. The CEFR and AAPPL descriptors are regularly used to inform curriculum design, language syllabuses, learning materials, tests (be they automated or not), language policies and teacher training programmes. Working with learner corpora can help inform pedagogical decisions, materials and practices at larger group- and proficiency levels (see Section 3 in this chapter for some illustrations, and also Chapters 20 to 23, this volume, for discussions of the various types of pedagogical applications of LCR).

An exclusive focus on larger groups would, however, be too restrictive as numerous types of individual differences are at play in SLA (see, for instance, Bigelow and Watson 2012; Duff 2012; Skehan 2012; Ushioda and Dörnyei 2012; Williams 2012 for recent research on individual differences). Individual differences typically include aptitude, motivation, identity issues, personality traits, type of working memory, socio-educational background, language proficiency in the mother tongue (L1) and other languages learnt, but also numerous aspects related to cognitive restructuring. As Bylund and Athanasopoulos (2014) recently put it: 'the extent and nature of cognitive restructuring in L2 [foreign/second language] speakers is essentially a function of variation in individual learners' trajectories'. Describing and understanding individual learning trajectories is thus – besides being valuable in itself – also essential to gaining a better understanding of group trajectories. The improved design, updated storage facilities and increasingly powerful descriptive and inferential methods of analysis used in LCR now make it possible to focus on learners as a group and learners within a group but also on individual trajectories as such, a too often neglected aspect in the early days of LCR.

Not all the variables potentially related to individual differences in SLA can be recorded in the metadata of learner corpora. Metadata related to personality traits, identity issues and L1 aptitude are, to my knowledge, not included in current learner corpora. In contrast, variables such as age

[15] American Council on the Teaching of Foreign Languages, see www.actfl.org/ (last accessed on 13 April 2015).

or information on the socio-educational background usually are. More recent longitudinal learner corpora also include complementary data types (other than authentic production data) to facilitate access to cognitive features of SLA. In the *LONGDALE* project, for instance, the same students are followed over a period of at least three years and data collections are organised at least once per year. The database contains argumentative essays, narratives and informal interviews, but also more guided types of productions (such as picture descriptions). Experimental data is also included for some of the subcorpora. The metadata are stored in a comprehensive learner profile which is gathered during each data-collection session. The variables include, among others: age, gender, educational background, country, language background, variables pertaining to the task and, when available, information on the proficiency levels of the students as measured by internationally recognised tests. The subcorpora in the *FLLOC* project also contain rich data types. As explained in the previous section, some of these corpora are longitudinal and others are cross-sectional. The *LANGSNAP* (*Languages and Social Networks Abroad Project*) *Corpus* is longitudinal and documents the development of modern language students' knowledge and use of the target language over a 23-month period including a 9-month stay abroad. *LANGSNAP* was collected to investigate learners' evolving social networks while abroad, the factors influencing type and amount of language engagement abroad, the kinds of learning opportunities afforded by target language interaction in a year-abroad context and the relationship between social networking, affect, social interaction and language learning. The data collected include authentic oral interactions but also day-long participant observation ('shadowing'). The *Young Learners Corpus*, also part of the *FLLOC* project, is cross-sectional but its design makes a pseudo-longitudinal approach possible. The corpus aims to document the development of linguistic competence among young classroom learners of French at three different starting ages, in primary and early secondary school classrooms, and identify similarities and differences (comparison of the rates of development at different ages after the same amount of classroom exposure; comparison of the classroom-learning strategies used by children at different ages and their attitudes to language learning). The corpus contains about forty hours of French language teaching for each of those three groups of learners. All language classes were recorded and, in addition, testing of the learners' French language proficiency took place at four different stages.

2.3 The focus of developmental studies in LCR

Numerous longitudinal studies have been carried out in SLA. Ortega and Iberri-Shea (2005) explain that the focus of such studies is often strictly linguistic (concentrating mainly on L2 morphology), and that the studies

typically involve few participants (although more recent studies tend to include more participants). The authors note, however, that there has been a recent broadening of the linguistic focus and also of the epistemological approach to language development, as illustrated, for instance, by longitudinal studies situated within a Vygotskian sociocultural framework. Ortega and Iberri-Shea (2005) refer to Belz and Kinginger's (2002) study of address form use as an example of such a new epistemological stance. The study documents critical incidents that contributed to the learning of indexical politeness (the use of French *tu/vous* and German *Du/Sie*) by two fourth-semester foreign language students at an American university taking part in a telecollaboration project. All the learners' interactions were collected in a learner corpus.

Thanks to its powerful lexical analysis techniques (in particular the extraction of collocations and recurrent word sequences), corpus linguistics has shifted the focus of linguistic analysis from grammar to lexis, and more especially phraseology, and LCR developmental studies reflect this trend. For example, Horst and Collins's (2006) longitudinal study tracks vocabulary growth and draws on an 80,000-word longitudinal corpus consisting of narrative texts produced by 210 beginner-level francophone learners of English. The samples were collected at four 100-hour intervals of intensive language instruction and the authors used lexical frequency profiling techniques. The authors find that although learners continue to use large proportions of frequent words over time, their productive vocabulary features fewer French cognates, a greater variety of frequent words and more morphologically developed forms. Marsden and David (2008) describe vocabulary use during semi-spontaneous oral production amongst instructed learners of French and Spanish at two different stages in the British educational system: Year 9 (near beginners) and Year 13 (approximately low intermediates). This pseudo-longitudinal study compares lexical diversity (range or variety of vocabulary used) across languages and across years, including analyses of different word classes. A last illustration is Bestgen and Granger's (2014) study, which focuses on phraseology and aims to assess the role played by phraseological competence in the development of L2 writing proficiency and text quality assessment. The authors use CollGram, a technique that assigns to each pair of contiguous words (bigrams) in a learner text two association scores, viz. mutual information and *t*-score, which are computed on the basis of a large reference corpus. The results show a longitudinal decrease in the use of collocations made up of high-frequency words that are less typical of native writers. As the study is conducted both longitudinally and pseudo-longitudinally, it also helps identify the respective contribution of each research design to the study of L2 writing development. Other lexically oriented studies tracking learners' development include Chen (2013b) on phrasal verbs, Crossley and Salsbury (2011) on lexical bundles (for a detailed description, see Chapter 10, this volume),

Kobayashi (2013) on the comparison between spoken and written productions, and Verspoor et al. (2012) on various types of lexical chunks (among other features examined in the study).

Other studies focus on morphology, grammar and syntax, often taking advantage of linguistic annotation tools such as part-of-speech (POS) taggers or, more rarely, parsers (see Chapter 5, this volume). For example, Vyatkina's (2013a) investigation of the development of grammatical complexity features relies on a POS-tagged corpus (see Section 3.3 for more details). Van Vuuren (2013), on the other hand, uses a syntactically annotated longitudinal corpus of student writing and compares it to a native reference corpus. She focuses on information structural transfer and analyses clause-initial adverbials in English as a Foreign Language writing produced by Dutch learners. Cross-linguistic differences in the information status of clause-initial position in a verb-second language like Dutch (compared to a Subject-Verb-Object order language like English) are hypothesised to result in an overuse of clause-initial adverbials in the writing of advanced Dutch learners of English. She observes that although there is a clear development in the direction of native writing, transfer of information structural features of Dutch can still be observed even after three years of extended academic exposure.

Some other papers address the relationship between lexical development and morphosyntactic measures. One example is David et al.'s (2009) study on lexical development in instructed L2 learners of French. The authors of this cross-sectional study analyse the relationship between lexical development and morphosyntactic measures in sixty instructed learners of French in Years 8, 10 and 12. The better understanding of SLA among young learners is meant to inform current primary language initiatives and educational practices in the United Kingdom and internationally.

2.4 Learner corpora, production data: what's in a name?

Different terminological options may sometimes lead to the conclusion that some data types are underrepresented. However, some mainstream, non-corpus-based SLA studies using data not referred to as learner corpora can be very similar to learner corpus studies and hence highly relevant to LCR. For example, Serrano et al. (2012) carry out a longitudinal analysis of the effects of one year abroad. They analyse the progress of fourteen Spanish-speaking learners of English during a one-year stay at a British university. Both oral and written data have been collected (three data-collection points) and the samples analysed in terms of fluency, syntactic complexity, lexical richness and accuracy. Two main research questions are investigated: (i) does L2 proficiency in oral and written production develop at the same pace while abroad or is improvement in one modality faster than in the other? and (ii) can learners' individual

variables, such as attitudes or chances to interact abroad, explain certain aspects of language development in oral and written production? Students' background information (referring to language attitude and language use) and authentic production data were collected and transcribed. The *Computerized Language Analysis* program (MacWhinney 2000) and the *Statistical Package for the Social Sciences* (*SPSS*) were used for the coding and analyses of the writing samples. The descriptive statistics for oral and written productions included fluency (syllables/minute), syntactic complexity (clauses/T-unit), lexical richness (Guiraud's Index) and accuracy (errors/T-unit). The results of the statistical analyses indicate that, while a few months abroad might be sufficient for some gains in oral performance, improvement in written production is slower. The type of interaction experienced and some attitudinal features have been shown to partly explain language development in some areas. Whilst Serrano et al.'s (2012) study would have escaped bibliographical searches relying on a keyword such as 'learner corpus' as the word is not used once in the article, it clearly has all the features of a longitudinal learner corpus study and is of high relevance for LCR.

Another example is Ferris et al.'s (2013) longitudinal study on written corrective feedback in L2 writing. The authors adopt a longitudinal (15-week semester), multiple-case (ten university-level L2 writers) classroom research design to address the impact of written corrective feedback for individual L2 writers. Although the focus of the study is primarily on students' descriptions of their own self-monitoring processes as they revise marked papers and write new texts, Ferris and colleagues set up data files for each of the ten students; those files include, among other things, the marked and revised texts (annotation for errors) and progress charts for each of the learners. After each of the first three timed writings, the researchers marked the 3–4 most prominent error patterns in each text written by the ten participants. In the study, 'prominent' could mean either most frequent or most serious for overall text effectiveness or some combination of the two. The individual researchers marking the texts were asked to use their best judgement about which 3–4 error types to mark. The authors argue that this procedure is similar to what classroom teachers do if they choose to mark student errors selectively rather than comprehensively. The types of errors marked by the researchers include article usage, lexical choices, missing words, sentence structure, agreement and punctuation. The ten case-study participants were shown to make a wide variety of errors in their timed writing assignments and individual learners' error patterns changed over the course of the semester, with, for instance, some learners making fewer word choice errors (vocabulary) but more sentence-level errors (syntax).

Here too, whilst some differences appear between what could be called learner corpus research and Ferris et al.'s (2013) work (notably more reliance on the use of intuition in terms of the marking of the errors, no

clear definition of what a prominent error may be and no error-tagging scheme provided in the article), the similarities with learner corpus data are numerous even if the term 'corpus' is not used in the article. Those similarities include the collection of metadata, of production data, the annotation of data and the establishment of progress charts. An additional strength of the article is the combination of various data types to best pinpoint the impact of instruction.

As has been shown in this section, longitudinal and pseudo-longitudinal studies point to the need for richly documented (multi)data and can be carried out using corpus data only or corpus data complemented by other data types. Whether the term learner corpus is explicitly mentioned or not, the validity of analysing authentic production data as one of the key data types to access learners' developmental patterns can no longer be questioned. To quote Larsen-Freeman and Cameron (2008b: 210), using learner corpora 'give[s] us access to stabilized patterns and variability around them'.

3 Representative studies

This section presents four studies illustrating the various designs presented in Section 2, with the first two being pseudo-longitudinal and the last two longitudinal. The publications also focus on various linguistic features and L2s: demonstrative reference for Austrian learners of English (Schiftner and Rankin 2012), prefabricated sequences for Swedish learners of French (Bartning and Forsberg 2006), syntactic complexity features for American learners of German (Vyatkina 2013a), and tense and aspect acquisition for French learners of English (Meunier and Littré 2013). The number of learners/subjects included in the representative studies also varies from two focal learners up to thirty-eight learners, and the data-collection points range from three up to fourteen.

3.1 Schiftner, B. and Rankin, T. 2012. 'The use of demonstrative reference in English texts by Austrian school-age learners', in Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: Benjamins, pp. 63–82.

Schiftner and Rankin's (2012) study seeks to identify developmental patterns in the usage of demonstrative reference in the written production of beginner and intermediate Austrian learners of English. The data used come from the *International Corpus of Crosslinguistic Interlanguage* (*ICCI*).[16] Seven data sets coming from the Austrian subcorpus of *ICCI* have been used, focusing on learners from Grade 5 to Grade 11 (i.e. with the age of

[16] See http://cblle.tufs.ac.jp/llc/icci/ (last accessed on 13 April 2015).

learners ranging from approximately 10 to 17). Part of the *Louvain Corpus of Native English Essays* (*LOCNESS*)[17] has been used as the native-speaker reference corpus.

Concordances for the four forms of demonstratives (*that*, *those*, *this* and *these*) were extracted using *WordSmith Tools* (Scott 2012). The demonstrative forms were coded manually for a number of grammatical and referential properties: grammatical function, proximity (proximal – P – or distal – D), number, type of reference (exophoric, anaphoric, cataphoric) and referent (noun phrase of proposition). It should be noted that learners sometimes use distal pronouns (e.g. *that*) to express proximal reference, hence the existence of, for instance, *that* P annotations. The results show that the overall frequency of demonstratives in all the learner subcorpora is lower than the overall frequency in the native corpus. However, there are differences between the individual demonstratives. The pronoun *that* is consistently overused (also when used as a proximal demonstrative) across all levels, while the pronoun *this* is consistently underused.

Despite many differences between the native and non-native corpora, even some low-proficiency learners show similarities with native speakers in their use of demonstratives. For example, demonstratives are used most frequently as short-range anaphors (i.e. referring to rather close antecedents in the text).

The authors also briefly comment on the pedagogical implications of their work. They suggest expanding the scope of teaching demonstratives beyond the properties of reference and proximity to include explicit comments on the larger syntactic patterns in which demonstratives occur.

3.2 Bartning, I. and Forsberg, F. 2006. 'Les séquences préfabriquées à travers les stades de développement en français L2', in *Actes du 16e congrès des romanistes scandinaves.* Department of Language and Culture, Roskilde University.

Bartning and Forsberg (2006) analyse the stages of development in prefabricated sequences (PSs) produced by Swedish learners of French. They map the acquisition of PSs with stages of morphosyntactic acquisition for the same learners. The six stages, described in detail in Bartning and Schlyter (2004a, 2004b), are labelled as initial, post-initial, intermediate, lower advanced, mid advanced and upper advanced. Only the first five stages have been examined in the present study.

Bartning and Forsberg (2006) divide the PSs into five main categories: lexical PSs (e.g. *coup de foudre*, *faire la fête*); grammatical PSs (e.g. *pas du tout*, *être en train de*); discursive PSs (*parce que*, *je veux dire que*, *tout à fait*); interlanguage PSs, which are syntactically or semantically deviant from native-like PSs but nonetheless used by learners as holistic and repeated PSs (such as the repeated use of *c'est tout passé bien* instead of the target

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

native PS which is *tout s'est bien passé*); and autobiographical PSs (e.g. *je m'appelle*, *j'ai* x *ans*).

A total of thirty semi-guided interviews have been used: twenty-five interviews with Swedish learners of French – coming from the *InterFra* (Bartning 2002) and *Lund* corpora (Granfeldt 2005) – and five interviews with native speakers of French. The beginner learners in the corpus had none or almost no French prior to data collection (and were aged 19 to 30); the intermediate ones had an average of 3.5 years of French (and were aged 16 to 18) and the advanced learners had from 4.5 up to 6 years of instruction in French and were university students aged 19 to 26. As can be seen, the design of the study is pseudo-longitudinal and the proxy used is proficiency level correlated with numbers of years of instruction in the target language (but not correlated with age).

In this study, the authors provide the numbers of PSs per type – and per learner too – and also individual and average percentages of words included in PSs out of the total number of words produced. The results show that lexical PSs are those that most distinguish learners from native speakers. Learners display a lack of progress from the initial to intermediate stages. Those first three stages are followed by substantial progress in stages four and five (i.e. lower and mid advanced). However, at these more advanced stages, the types and frequencies of PSs displayed by learners are still quite different from those produced by native speakers of French.

Grammatical PSs are the only ones that do not display clear progress, with similar percentages found throughout the various stages. The use of discursive PSs, in contrast, increases significantly from the initial to post-initial stage. Autobiographical PSs are found mainly in initial stages and are almost non-existent at more advanced stages. This is due to the fact that autobiographical PSs are typical of discussion topics at initial stages of proficiency. The interlanguage PSs decrease significantly with proficiency: from 21% of interlanguage PSs at the initial stage to only 1% at the mid advanced stage.

In terms of overall frequencies, Bartning and Forsberg (2006) show that the proportion of PSs in learners' speech increases with proficiency. The authors nonetheless conclude that whilst verbal morphology displays what they call a strict development (p. 19), prefabricated language does not seem to follow such strict development and is more sensitive to input and to the communicative style of individual learners.

3.3 Vyatkina, N. 2013a. 'Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus', *The Modern Language Journal* 97(S1): 11–30.

Vyatkina (2013a) analyses specific syntactic complexity by studying the developmental profiling of individuals on the basis of an annotated learner corpus. The study aims to track the development of syntactic complexity – with a focus on individual developmental pathways – and aims

at pedagogical improvements in the teaching of writing to beginners. The author provides an in-depth analysis of the writing of two beginner L2 German learners (with L1 English) over four semesters of collegiate language study by using developmental profiling techniques. The study design is longitudinal, with multiple and relatively dense data-collection waves (fourteen measurement occasions corresponding to the fourteen units seen in the textbook used in class).

Vyatkina explores variation in terms of the frequency of some complexity features (such as coordinate, nominal and non-finite verb structures) and, to do so, she uses corpus analysis techniques with semi-automatic corpus annotation: initial automatic POS tagging, manual checking of the output and manual selection and count of more complex structures. The developmental dynamic is explored as sets of complexification strategies (see Ortega 2012) or repertoires of choices (as described in Ortega and Byrnes 2008a) of specific syntactic structures, used by learners at each of the fourteen measurement occasions. In so doing, Vyatkina analyses learner development in terms of multidimensional variability and non-linear relationships between the instructional progression and individual developmental paths.

The study follows the 'instruction-embedded total-sampling approach' (Byrnes et al. 2010: 165), in which writing samples were rough drafts of essays written by the students in response to curricular tasks rather than to external experimental tasks. The participants were students enrolled in a beginning German language programme at university over four sequential 16-week-long semesters. All classes were taught by graduate teaching assistants who followed a uniform syllabus and used the same textbooks. Each writing task concluded a corresponding textbook chapter and reflected the book's instructional content, including the focus on selected grammar structures. During the first three semesters, students typed each essay in class under timed conditions. They were required to write during the whole 50-minute-long class period and were allowed to use online dictionaries but not online translators, textbooks or notes. During the fourth semester, they wrote essays at home under untimed conditions and were allowed to use reference materials. The very last essay was again timed and written under controlled conditions. As argued by Vyatkina, variation in tasks and topics may affect linguistic complexity. However, the present study does not focus on these specific effects but rather on how two different learners respond to one and the same task at each time point.

The data collected was then annotated for syntactic complexity. The learner corpus was first tagged automatically for fifty distinct word classes using the *TreeTagger* for German (Schmid 1994) and the output was manually checked. The annotation of specific complex structures was then performed using a mixture of automatic and manual searches (e.g. searching coordinating conjunctions using the Concord function of

WordSmith Tools (Scott 2012) and then counting different coordinate structures manually based on the context of each retrieved example). More complex structures (e.g. clause types and infinitive constructions) were annotated manually by two independent annotators.

Without going into the details of each specific syntactic measure used in the study, it can be said that whilst the results show a general developmental trend towards increased frequency and range of syntactic complexity features, the trajectories of the two focal learners reveal divergence between the learners in the second half of the observation period. One male participant readily responds to instruction but abandons some syntactic features when progressing to the next task, whilst the other focal learner balances both previously learned and new features in her writing.

The pedagogical implications derived from the study have direct classroom relevance. Vyatkina (2013a) proposes the design of rubrics listing specific lexical and syntactic features (with examples) associated with each level-appropriate writing task, i.e. some sort of idealised writing profiles, with model texts and the inclusion of contextual use and functions in order to raise learners' awareness of what their expected developmental targets are.

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

In this article, Meunier and Littré (2013) track French learners' progress in the acquisition of the English tense and aspect system. To do so, they adopt a dual approach and use both corpus and experimental data.

The first part of the article reviews the status of longitudinal research on the acquisition of tense and aspect in SLA and explains that L2 longitudinal studies have often mirrored typical L1 longitudinal studies by tracing the first steps of language acquisition (either by children or adults learning the L2). Other common features of tense and aspect studies in SLA are that a majority of the studies focus on very few learners or have selected learners from a variety of mother-tongue backgrounds (which helps uncover universal paths of acquisition but limits further investigations into transfer effects).

In their longitudinal study, Meunier and Littré want to highlight the value of a learner corpus approach to the study of tense and aspect development (more specifically in that study, simple present vs present continuous) but also insist on the importance of combining learner corpus data with other data types (grammaticality judgement tests and grammaticality cum interpretation tasks). This combination makes it possible to go beyond the identification of difficulties (or lack thereof) in the learners' acquisition over time and to try and uncover the cause(s) of some remaining language difficulties even after many years of exposure to, instruction in and use of English.

The authors have decided to focus on more advanced levels of acquisition by analysing the written productions of a cohort of thirty-eight French-speaking English language and literature students at the University of Louvain, with each participant having contributed three argumentative essays (one per year over a three-year span). These essays are part of the *LONGDALE* project (presented in Section 2.2). Whilst a total of thirty-eight learners might not appear very impressive in comparison to corpus standards, the study nonetheless has a much larger sample size than most of the existing longitudinal studies on tense and aspect, which generally have an average number of informants lower than ten (and often lower than five). For the very few studies on tense and aspect carried out on more than twenty subjects (see, for instance, Klein and Perdue 1992; Bardovi-Harlig 1992), the informants came from a mixture of native-language backgrounds.

The results of the multi-level regression analysis show that the time predictor has a positive effect on the decrease in tense and aspect errors produced by learners over a period of three years, both at group and individual levels. It is interesting to note that various statistical models were also fitted on more data (to test the impact of attrition, taking into account essays produced by learners at year 1 and year 3 but not at year 2, for instance). These models included list-wise deletion (i.e. selecting only the thirty-eight participants who wrote three essays), taking only a subset of participants (i.e. only the groups that did not significantly differ from other groups were included, meaning that participants who dropped out between year 1 and year 2 and those who joined the study in year 3 were rejected) or taking all participants (i.e. all participants including those with attrited data). Irrespective of the models tested, time has a significant effect on the reduction of tense and aspect errors (ranging from about 16–27% depending on the model fitted). The top ten error pairs were reorganised into three categories, namely aspect-only errors (i.e. when the tense is correct and the aspect is not), tense-only errors (i.e. when the tense is incorrect and the aspect correct) and mixed tense and aspect errors (i.e. when both tense and aspect are incorrect). The results show that more than 50% were aspect-only errors, with progressive present/simple present ranking at the top and accounting for 25% of all the tense–aspect errors.

The experimental data analysis, carried out to help trace the reason for the persistence of errors related to the progressive aspect, shows that whilst learners master the most salient elaboration of the progressive (viz. ongoingness), their understanding of less core uses (e.g. planned events) is much less precise.

Subsequent guidelines for classroom teaching are proposed in the study, among which the idea that teachers of advanced French learners of English should not necessarily review the whole range of uses of tenses and aspect from year to year (as is commonly done). They should

no longer teach the most prototypical uses to more advanced learners and should spend more time on less frequent or less core uses (e.g. the modal meanings of the present progressive).

4 Critical assessment and future directions

The discussion of the core issues and the presentation of representative studies have pointed to the fact that learner corpora are solid and reliable data sources to trace learners' proficiency development in an L2. As pointed out in Section 1, many variables recorded as metadata in LCR can be used as dependent variables, potential predictors or dynamic factors impacting SLA. As for the learners' productions, they can be analysed as being representative of larger groups or populations (on the basis of the variables encoded in the corpus) but within-group variability and individual trajectories can also be accessed. The linguistic focus of the studies presented is also extremely varied and encompasses all dimensions of the complexity, accuracy and fluency paradigm (Housen et al. 2012).

In the last paragraphs I would like to stress what I consider as essential steps to be taken for a sound and healthy development of longitudinal learner corpus studies. First, whilst proxies (such as proficiency level) can undoubtedly be used to circumvent the difficulties inherent to the collection of longitudinal data (see Section 2.2), it is essential that sustained efforts be devoted to the collection of longitudinal data as 'longitudinal designs can uniquely help researchers document the lengthy trajectories of adults who strive to become multicompetent and multicultural language users' (Ortega and Byrnes 2008a: 18). The collection of longitudinal data goes hand in hand with the need for new practices/requirements in learner corpus data collection. I would, for instance, plead for the collection of information related to proficiency in the learners' L1. This would enable researchers to be much more specific in their future analyses and interpretations of bi- and multi-literacy practices. L1 production data should also ideally be collected as it would enable an integrated comparison of the learners' proficiency levels in their L1 and L2 and would greatly enhance the interpretation of the results for individual trajectories (including access to features of learning disabilities such as dysorthographia or dysgraphia for writing and dyslexia for speaking). Another requirement is the ongoing/dynamic collection of metadata. The importance of metadata in learner corpus studies is paramount and perhaps even more so in longitudinal designs as researchers follow learners over a much longer period of time. This naturally implies that the initial metadata collected at, say, Time 1 will not be self-sufficient. Researchers will want to document the learning paths (courses taken, stays abroad, amount and type of language practice, etc.). Such rich and dynamic

metadata will be essential for a refined understanding and interpretation of future research results.

Second, in terms of linguistic features analysed, lexis (single or multi-word units) and grammar have occupied pride of place in LCR. The lexis–grammar interface and the patterned nature of language have also been central. Communication strategies, in contrast, have been the poor relation. Future research should also focus on these strategies, viz. how language learners maintain communication, make meaning and negotiate meaning. If such issues have been partially addressed, mainly in LCR studies using computer-mediated communication, they nonetheless deserve more attention in the future.

A third issue that I find essential to consider is that of corpus size and representativeness as measured, among other things, by the number of subjects included in a developmental study. Case-study approaches are not always particularly valued in corpus-linguistic circles; Ortega and Byrnes (2008a: 9) even speak of 'the contested legitimacy of the approach in certain social science circles, including sectors of applied linguistics'. That said, the highly valued exponential growth of corpus sizes as exemplified by the collection of what Davies calls 'second generation mega corpora'[18] and the numerous benefits that can be gained from the analysis of such corpora may not necessarily be the desired path to follow when it comes to longitudinal learner corpora. Whilst 'big is beautiful' is still a valid motto in corpus circles, smaller but much 'denser' longitudinal learner corpora should also be valued, and collected. To quote Polat (2011: 3754), 'despite its time-consuming and labor-intensive collection process, the use of a dense developmental corpus seems to be a very promising research approach, especially if paired with more qualitative analyses'. As mentioned in Section 2.1, learner corpus collection implies a constant trade-off between the density of the data and the number and/or representativeness of the subjects whose language data is collected. More subjects are typically involved in cross-sectional data-collection designs, which lend themselves more naturally to quantitative analyses. In longitudinal designs, in contrast, fewer subjects are often involved but more qualitative studies can be performed as individual trajectories can be analysed in detail.

In order to combine the strengths of various approaches, a mixed approach can be used. Johnson et al. (2007: 112) state that mixed-methods research (MMR) is 'becoming increasingly articulated, attached to research practice, and recognized as the third major research approach or research paradigm, along with qualitative research and quantitative research'. The promotion of MMR in LCR is the fourth issue that I would like to identify. Bergman (2008: 1) defines MMR as 'the combination of at least one qualitative and at least one quantitative component in a single research project or program'. As for Ibbotson (2013: 2), he explains

[18] See http://davies-linguistics.byu.edu/ling485/for_class/corpora_notes.htm (last accessed on 13 April 2015).

that no matter whether 'the focus is on language processing, acquisition, or change', in usage-based linguistics knowledge of a language 'is based in knowledge of actual usage and generalizations made over usage events' and the complexity of language 'emerges not as a result of a language-specific instinct but through the interaction of cognition and use'. LCR, with its exclusive focus on (semi-)authentic language use and analysis, constitutes one of the usage-based paradigms and, as such, lends itself well to MMR. It must be kept in mind that a quality standard for MMR is, however, as Hashemi and Babaii (2013: 828) state, to achieve high degrees of integration at various stages of the study. This integration has, for instance, been achieved in Rosi's (2009) study on the acquisition of aspect in Italian L2. Her research involves work on a native Italian corpus and on longitudinal learner corpus data, which is supplemented by the analysis of experimental data; there is constant to-ing and fro-ing between quantitative and qualitative analyses which are carried out and interpreted within a cognitive and connectionist framework.

Finally, the last issue concerns availability: more data should be made available to a larger research community. This wish, expressed repeatedly by Myles (2008 and other publications), goes beyond making the data available and includes making statistical codes, task prompts or coding systems available in order to favour replication studies and enhance the expertise in analysing longitudinal data. As argued by Littré (2014), it may help other researchers better understand the choices made in a specific study but it also represents a commitment to the openness and transparency that is central to the scientific endeavour. The author adds that whilst a necessary balance must be struck between openness and other practical considerations (time needed to collect the data and priority for analysing it, for instance), it would also allow other research teams to identify potential errors and improve upon previous analyses.

Key readings

Ortega, L. and Byrnes, H. (eds.) 2008b. *The Longitudinal Study of Advanced L2 Capacities*. New York: Routledge.

This edited volume, whilst not focusing exclusively on learner corpora, is a must-read for any researcher interested in longitudinal research. It provides key theoretical and methodological reflections on the longitudinal study of advanced capacities and includes chapters that report on empirical longitudinal investigations of various types (descriptive, quasi-experimental, qualitative and quantitative). Chapter 4 in the volume, written by Florence Myles, is more specifically dedicated to the investigation of learner language development with electronic longitudinal corpora.

Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.) 2012. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: Benjamins.

This volume provides an overview of current research on the use of learner corpora perceived from developmental and cross-linguistic perspectives. Eleven chapters of the book focus on the proficiency development of young learners of English as an L2 on the basis of the *International Corpus of Crosslinguistic Interlanguage* (*ICCI*). The other articles present studies carried out on spoken learner corpora and on learner corpora of languages other than English (French and Japanese).

Hasko, V. and Meunier, F. (eds.) 2013. *Capturing L2 Development through Learner Corpus Analysis*. Special issue of *The Modern Language Journal* 97 (S1).

This special issue of *The Modern Language Journal* is entirely devoted to the role that learner corpora can play in uncovering the developmental processes in L2 learning. The introductory chapter offers a critical discussion of the aspects in which the disciplines of LCR and SLA would benefit from closer interdisciplinary engagement. The six articles included in the volume address syntactic complexity, contiguous and discontiguous multi-word unit use, the numeral classifier system, tense and aspect acquisition and the development of L2 accuracy learner profiles, and illustrate various learner corpus research designs.

Belz, J. A. and Vyatkina, N. 2008. 'The pedagogical mediation of a developmental learner corpus for classroom-based language instruction', *Language Learning and Technology* 12(3): 33–52.

This article explores the pedagogically mediated use of a learner corpus in language teaching and in the developmental analysis of second language acquisition. It also addresses the issue of authentication in corpus-driven language pedagogy. The authors illustrate how an ethnographically supplemented developmental learner corpus may contribute to second language acquisition research via dense documentation of micro-changes in learners' language use over time.

Castello, E., Ackerley, K. and Coccetta, F. (eds.) In press. *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*. Bern: Peter Lang.

This edited volume contains five articles devoted to longitudinal corpora. The authors of these articles have all used *LONGDALE* data and address the following topics: tense and aspect issues in an

error-annotated corpus of, respectively, French and German learner writing; the evolution of word classes in a variety of written assignments produced by Dutch learners and individual learners' differences that can affect vocabulary and syntactic control; French learners' ways of expressing attitudinal stance in oral communication; metadiscursive features, more specifically the evolution of the use of *it*-extraposition in the reading reports and argumentative essays produced by Italian learners; and finally the use and short-term impact of corpus literacy practices and data-driven learning activities in first-year Italian language students.