22

What can corpora tell us about language learning?

Pascual Pérez-Paredes and Geraldine Mark

1 What is language learning?

In this chapter we take a two-pronged approach to considering what corpora can tell us about language learning: (1) the "can of affordability/facility" of corpora – what they represent and what they enable/allow us to look at and (2) the "can of possibility" – the potential for a broader understanding and representation of language learning.

What distinguishes a corpus linguistics perspective from other approaches to language learning research is its emphasis on the study of language as a product. Language, as a semiotic resource, is situated at the micro level of social activity where L2 users engage in the completion of tasks in instructed learning contexts or in real life, both face-to-face and virtual, interactions with other speakers and increasingly more so, with automated agents through voice and text interaction. In learner language research, corpora are used as proxies of usage, that is, as providers of evidence of communication, spoken or written. The question is, once evidence of this usage has been gathered, what do researchers do with it and what does it tell us about learning?

As Tyler and Ortega have put it, learning a language 'is one of the most complex accomplishments humans achieve. This is true for the first language learner and perhaps even more so for the second language learner' (2018: 3). Amid this complexity, expert views both on the nature of language and the nature of learning differ, and here we note the existence of at least two approaches to how the term has been conceptualised by researchers in the broader field of applied linguistics.

Inclusive conceptualisations are mindful of the heterogeneity in the language learning endeavour, encompassing different languages across space, social contexts, lifespan, language status and roles of instruction. Mitchell, Myles and Marsden advocate a broad understanding of the term that includes 'the learning of any language, to any level, provided only that the learning of the "second" language takes place sometime later than the acquisition of the first language' (2013: 1). The Douglas Fir Group (a transdisciplinary group led by the belief that language acquisition is too complex to be explained by one or two theoretical approaches) (2016: 19) claim that the broader language learning field is interested in 'school-aged children, adolescents, and adults [and how

DOI: 10.4324/9780367076399-22

they] learn and use, at any point in life, an additional language', including second, foreign, indigenous, minority or heritage languages. Inclusive conceptualisations accept, therefore, that different language learning theories co-exist and compete with each other to explain specific areas that are considered relevant when accounting for how language learning occurs: 'language inextricably involves cognition, emotions, consciousness, experience, embodiment, brain, self, human interaction, society, culture, mediation, instruction, and history in rich, complex, and dynamic ways' (Douglas Fir Group 2016: 39).

When specific approaches are concerned with a limited set of learner types, contexts of learning and scope of theorisation, we are dealing with what we may label as *narrow conceptualisations* about language learning. These may range from, for example, young learners' language learning in classroom contexts through explicit instruction to language learning in immersive instructional or naturalistic contexts.

Narrow conceptualisations tend to prioritise one aspect of learning over others and as a result can expose a fragmentation of the phenomenon into cognitive, social and emotional elements. We note a tendency in learner corpus research (LCR) to conceptualise language learning as a cognitive phenomenon where language is primarily observed as a bounded system of formal rules. We argue, following Mitchell *et al.* (2013), that language learning needs to account for quite a long list of evidence and theorisation around the nature of language and language use; the learning process itself; the role of variability in L2 learning, as well as the target language models pursued; the role of the L1(s) and L2(s); and, among others, the role of the language learner and learner differences.

When it comes to analysing the contribution of corpora to language learning, it is necessary, therefore, to try to understand what language learning is being contributed: Are there specific views of language learning that have been embraced by CL researchers? If so, how have they bridged that gap between theory and empirical research? To address these questions, we will look at how learner language has been represented in corpora, how these corpora have evolved and what methods and approaches we use to investigate them. We note here that while both L1 and L2 corpora have been used in language learning to teach languages (see Chapter 30, this volume) as part of a research resource in second language acquisition (SLA) (see Chapter 23, this volume) to build teaching materials, coursebooks, dictionaries, etc., and in data-driven learning (see Chapters 24–31, this volume), the main focus of this chapter is on corpora that contain learner language.

2 Learner language in corpora

History and focus

In corpus linguistics-inspired research, the term "learner language" is used to denote the body of language produced by L2 users. Within SLA studies, Ellis and Barkhuizen have defined the scope of learner language as 'the primary data for the study of L2 acquisition' (2005: 4) and situate their discussion in the context of the *limitations* that the collection of natural use imposes. While they cite a wide range of data collection and analysis methods, the study of language, as conceptualised from the corpus linguistics camp, with its well-defined set of methods and data collection procedures, has not necessarily been conceived as mainstream in traditional SLA research (Dornyei 2007).

The emergence of learner corpora claimed a new turn in providing 'systematic collections of authentic continuous and contextualised language use by foreign/second (L2) learners' (Callies 2015: 35). Pre-corpus work on learner language is often associated first with learner error, with errors as a window into the learning process, as evidence of learner strategies and processes, departing from the view of error as "bad habits" (Corder 1967), and then with "interlanguage" (Selinker 1972), and the idea of learner language as an independent system worthy of analysis in itself. The Louvain-born *International Corpus of Learner English* (ICLE) project marked a major landmark in the collection and study of learner language with the emergence of the Contrastive Interlanguage Analysis (CIA) tradition (Granger 1994) and the gathering and analysis of large-scale learner data from a variety of L1 backgrounds. The ICLE project (see Table 22.1) was built within the design matrix for the L1 *International Corpus of English* (ICE) (Greenbaum 1990) and brought with it the wealth of methodological approaches already afforded by corpus linguistics. Many new learner corpora were built following the ICLE design (Tono 2003; Tono and Díez-Bedmar 2014), and new error-coding and tagging systems were developed. A perceived strength of this design at the time was that it allowed researchers to identify those learner "errors" that were universal to all learners from all L1 backgrounds as well as those that were L1-specific. The CIA approach has had such a lasting impact on the field of learner corpus research that most corpus-based studies have adopted a comparative design, either contrasting learner language with an L1 benchmark corpus or to another L2 dataset. LCR, following in the coat-tails of L1 corpus research, brought with it a broadening of the topic of analysis in learner language, with a shift from the traditional SLA focus on morphology to attention on lexis and phraseology and register and the wide range of variables afforded by metadata and tagging systems.

Learner corpora continue to gather for a myriad of research purposes (see https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html).
Much of the pioneering work in corpus development originated as a result of commercial and academic partnerships (e.g. the seminal work of the BNC and the COBUILD project), with commercial interests leading the way. A similar impetus can be seen in learner corpus development, e.g. the *Cambridge Learner Corpus* (CLC), the *Longman Learners' Corpus* and, most recently, the *Trinity Lancaster Corpus* (TLC), which have all been developed as commercial resources, for example, by publishers and exam boards for the production of published materials and related academic research.

From description to understanding, from error to competence

Methodological issues of comparative studies, control of variables, subjectivity in the assignment and categorisation of both learner-related features (e.g. proficiency level, variation) and annotation (e.g. error tagging) have been widely debated (Tono 2003; Ädel 2015; Granger 2015; Gablasova *et al.* 2017; McEnery *et al.* 2019) and are issues which endure in corpus design to the present day. While acknowledging the impact made by LCR in mobilising the creation of learner corpora and *describing* learner language, there has long been a desire 'to test some of the current hypotheses [in SLA] on larger and better constructed datasets, as has happened in L1 acquisition' (Myles 2005: 376) and still more recently a call for LCR to engage beyond the contrastive descriptive paradigm and to realise the promise for its usefulness in *understanding* language learning (McEnery *et al*. 2019).

Frequency and comparability have persisted as a central focus of LCR. Descriptions of learner language use are often still provided in deficit terms of overuse, underuse and misuse or error in relation to a target L1 or comparable L2 dataset. Recent developments in the field are nudging the focus away from error-based to competency-based analysis, describing the data in terms of what learners can do rather than what they are getting "wrong", with a view to understanding development and the "learning" process. Emergence of aspects of this pursuit is seen *inter alia* in the developmental studies of "criterial features" – the linguistic properties characteristic of a given level of competence (Hawkins and Buttery 2010), patterns of L2 accuracy (Thewissen 2013), L2 morphemes (Murakami 2014) and, more recently, formative usage-based studies (Ellis *et al.* 2016; Römer and Garner 2019; Pérez-Paredes *et al*. 2020; see also Chapter 23, this volume). These studies represent to varying degrees the coming together of SLA and LCR research, all motivated by the study of language learning though with diverse goals and methodologies. Increasingly, both fields are seeing the relevance of the data and methods of the other, particularly when creating triangulated research designs. They represent some effort to shift the focus in both the design and analysis of learner corpora from description to interpretation, but the movement is slow (Myles 2015).

Design, data and focus

A declared aim of corpus designers is to contribute to our understanding of language learning, and Table 22.1 outlines three areas that define how research designs are approached. These areas relate to tasks, language collection considerations and the learners participating. Although learner corpora vary in scope, most consider all three elements, while often prioritising them differently.

Learner corpora in general, as represented by a sample in Table 22.2, are intended to shed some light on learner language usage in essays or other types of written tasks and spoken tasks, either monologic or dialogic, at different points in time and across different groups of L1 speakers.

Table 22.1 Corpus design elements and considerations

Research design element	*Considerations*
Mode/Tasks	Task types: the degree of semi-experimental control during the collection of these tasks varies across corpora. Tasks dictate written/spoken mode, degree of interaction, time allowed for task, etc.
Collection	Cross-sectional: data collected at one time point from multiple groups (see Chapter 23, this volume). Dynamic, longitudinal: data collected from the same group at multiple time points. Pseudolongitudinal or quasilongitudinal: a mix of data collected from the same learners at different points in time and from learners at different proficiency levels at the same point in time.
Learners	Learners are usually grouped according to their L1 and, depending on the corpus, their proficiency level or year of study. The number of learners involved and the sampling strategies vary.

Table 22.2 A sample of learner corpora and their research purposes and content

International Corpus of Learner English (ICLE) https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html Argumentative essays written by upper intermediate to advanced learners of English from several L1 backgrounds. Allows for the regular inclusion of new subcorpora, highlighting the fundamentally dynamic nature of the ICLE project. **Key elements:** L2 English, 25 L1 backgrounds, written, cross-sectional, 5.5 million words, high-level proficiency.
Louvain International Database of Spoken English Interlanguage (LINDSEI) https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html Spoken counterpart to ICLE, produced by advanced learners of English from different L1 backgrounds. **Key elements:** L2 English, 11 different L1 backgrounds, spoken, cross-sectional, 1 million+ words.
Trinity Lancaster Corpus (TLC) https://www.trinitycollege.com/about-us/research/Trinity-corpus A large corpus of learner (and examiner) speech which can be used in a wide range of research contexts, including SLA, language testing, L2 pedagogy and materials development, etc. **Key elements:** L2 English, nine different linguistic and cultural backgrounds, spoken, pseudolongitudinal, proficiency levels B1 to C2.
Corpus Escrito del Español como L2 (CEDEL2) http://cedel2.learnercorpora.com/#section6 Investigating how people learn Spanish grammar (morphology and syntax). A comparative corpus of L1 Spanish – L2 English, called WriCLE (Written Corpus of Learner English) was also created. **Key elements:** L2 Spanish, L1 English, written, average age 20 years; 512,873 words, cross-sectional.
Longitudinal Corpus of Chinese Learners of Italian (LOCCLI) https://www.unistrapg.it/cqpweb/doc_corpora/LOCCLI_documentation.pdf 350 essays written by 175 Chinese learners of Italian, who attended Italian language courses in Perugia for 6–8 months, and it was collected in 2016. Data was collected in two different points in time, from the same learners. **Key elements:** L2 Chinese, L1 Italian, 175 learners; longitudinal
Guangwai-Lancaster Chinese Learner Corpus (GLCLC) http://cass.lancs.ac.uk/wp-content/uploads/2016/05/Poster_GLC-small.pdf A balanced sample that covers three proficiency levels: beginner, intermediate and advanced, providing a unique insight into L2 Chinese lexical and grammatical development. **Key elements:** L2 Chinese; written and spoken, different proficiency levels, longitudinal, 80 countries represented, over 1 million.

The design and collection of learner corpora are subject to the same gamut of considerations, processes, affordances and pitfalls as any other corpus; they are representative only of what they contain, whatever efforts are made for balance and representativeness. There are more learners of English than any other language globally,

and therefore there are strikingly more corpora of L2 English than of any other language. Ease of collection may tend inevitably towards the written over spoken and cross-sectional over longitudinal and also to a disproportionate number of corpora of high-proficiency-level L2 users studying a target language at university over lower-level and young L2 learners (see Chapter 23, this volume). There is a host of variables specific to learner corpora (e.g. L1 background, L1 context, proficiency level, learning history, task effect, type of language instruction/input) to consider and accommodate. Added to this, Buttery and Caines (2010) point out that for fair accounts of learner language to be made, "opportunity of use" must also be controlled for. Learners may be using their linguistic resources strategically, not necessarily displaying what they know, constrained by the limitations of the context.

A detailed list of individual studies (c.2000 references at the time of writing) can be found in the learner corpus bibliography (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html). The objects of focus of these studies can be categorised as follows:

> *Discrete features*: e.g. individual parts of speech; adverbials; articles; cohesion; discourse markers; formulaic language; lexical bundles; tenses; verbs, etc.;
> *Composite features*: e.g. measures of phraseological sophistication; measures of clausal and phrasal complexity, etc.;
> *Constructs*: e.g. metadiscourse features; involvement; information packaging; grammar and language learning, etc.

We note that not all areas of potential study receive equal attention. Using the learner corpus bibliography, Paquot and Plonsky (2017) have shown that lexis (including single words and multi-word units) is, by a long way, the main focus of study, accounting for 65 per cent of all studies, in contrast with discourse and pragmatics, accounting for 30 per cent and 10 per cent, respectively. Given the range of targets chosen for analysis, currently available learner corpora offer a window on a narrow conceptualisation of language learning, which may explain some of the challenges to making corpora more relevant in the broader field of language learning research and practice.

3 Learner corpus research in practice

Here we zone in on three case studies, representing different types of data, approaches and findings, and consider how they have contributed to our understanding of language learning.

The first study looks at exploratory longitudinal corpora, encouraging the building and analysis of small-scale data with a focus on discrete and composite features of language, including syntactic complexity. It demonstrates the potential for a triangulated approach, complementing corpus tools and statistical analysis with natural language processing tools. The second is a study of adverb use in spoken English in cross-sectional data across a variety of tasks. The third study investigates grammatical development across proficiency levels, in a pseudolongitudinal corpus of written exams.

Case study 1: examining the evolution of language learning

Here we examine two robust approaches to corpus longitudinal designs. The first approach uses a longitudinal corpus to track the development. Vyatkina (2013) looked at the development of syntactic complexity in two beginning L2 German learners over four

semesters in a US college. She collected data every three or five weeks and gathered 19 pieces of both timed and untimed essays from each student. She looked at syntactic structures such as coordinate and complex nominal structures per clause, using developmental profiling techniques (Figure 22.1) – a combination of POS tagging with manual checking and annotation, using concordance software.

This approach can reveal the point when L2 target features emerge in the writing after focused instruction. Vyatkina suggests that language teachers compile their own small learner corpora and develop learner developmental profiles for learning, feedback and assessment purposes.

The second approach (Siyanova-Chanturia and Spina 2019) uses the *Longitudinal Corpus of Chinese Learners of Italian* (LOCCLI) (see Table 22.1) to understand the effects of a six-month instructional period on the acquisition of noun + adjective word combinations by L1 Chinese learners of Italian (n = 175) at three CEFR levels (A1, A2 and B1). The learners wrote two essays at the beginning and end of the teaching period. The researchers used a variety of measures to analyse 1,401 observations of [N + Adjective] combinations and mixed-effects modelling to examine changes between the two collection points. They did not find a correlation between time effects and the frequency of the aforementioned measures; that is, after six months the learners did not produce more word combinations. The authors argue that L2 collocational knowledge is slow and uneven, though they found that the decrease in use of word combinations was more significant in the A1 level. The authors conclude that the learning of collocations is a process 'fraught with difficulties… in which more exposure and higher proficiency may not necessarily lead

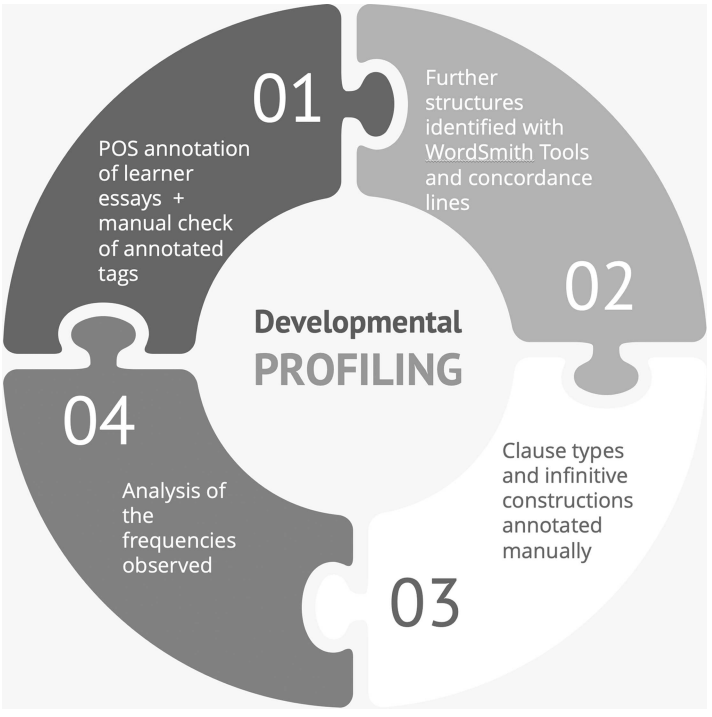

Figure 22.1 Developmental profiling techniques used in Vyatkina (2013)

to a more idiomatic and target-like output and may, indeed, result in lower levels of idiomaticity and greater reliance on lower frequency combinations' (2019: 452).

Case study 2: researching the use of stance adverbs in a spoken corpus

Pérez-Paredes and Díez-Bedmar (2019) provide evidence to suggest that task selection and design needs further attention in LCR. They examined the use of *really*, *actually* and *obviously* in two populations of Spanish L1 speakers of English in the TLC (see Table 22.2). Frequency differences were found at B1 and B2 levels in discussion tasks. L1 Mexican and European Spanish speakers used these stance adverbs in significantly different ways. At B2 level, this difference was also significant in both conversation and discussion tasks. The use of *really* and *actually* as device-making devices in dialogic discourse is more frequent as language learners' communicative competence increases. These uses are, additionally, less epistemic. Usage differences in the learner groups who share the same L1 but have a different cultural and geographical background (Mexico and Spain) also point to the influence that different varieties of L1 English may have on learners. This research lends further evidence (Mark and Pérez-Paredes 2018; Pérez-Paredes and Bueno 2019) to the way in which tasks influence the elicitation of language.

Case study 3: using a large-scale corpus of written exam data to examine competence and development

A large-scale corpus project, *The English Profile* (www.englishprofile.org), used the CLC to develop detailed profiles of learner language use at each of the CEFR levels (Harrison and Barker 2015). O'Keeffe and Mark (2017) looked at grammatical features across six proficiency levels, resulting in an open access online description of learner use. Other strands focused on vocabulary (Capel 2012) and communicative functions (Green 2012).

The 55-million word CLC is a pseudolongitudinal corpus compiled from written exams. Each document is tagged for L1 (143 L1s), nationality, level of education, age, gender, exam taken, performance achieved, CEFR level, task type, style and format, allowing for a developmental view of a linguistic feature across proficiency levels, as well as a cross-sectional view across any of the tagged elements. O'Keeffe and Mark (2017) devised an iterative design, combining cycles of analysis to facilitate an aggregated picture of usage, moving from a single element (e.g. frequency of form) to a breakdown of detail of usage across the range of metadata. Automatic part-of-speech (POS) tagging (in the *Sketch Engine* toolkit) allowed for analysis of structural sequences using a corpus query language (CQL) search facility. The iterative methodology involved first a quantitative view based on frequency of use and distribution across levels, L1s and tasks, complemented with a close qualitative analysis of the concordance lines. By way of example, past simple relative frequencies across the proficiency levels are retrieved using a CQL search on a past simple tag. A quantitative view of forms and normalised frequencies shows a fairly even distribution of usage across all six proficiency levels (between 24,391 and 28,601). However this hides a complexity of use, revealed when these frequencies are explored qualitatively through detailed analysis of concordance lines and collocational patterns at each level. What emerges is a growing repertoire of functions and associated lexis, summarised in Table 22.3.

Table 22.3 Development of past simple form and functions across proficiency levels

	A1	*A2*	*B1*	*B2*	*C1*
FORM: AFFIRMATIVE (limited range of verbs)					
USE: EVERYDAY EVENTS AND STATES					
FORM: AFFIRMATIVE (increasing range of verbs)					
FORM: NEGATIVE (limited range of verbs)					
FORM: QUESTIONS: yes/no, wh- (limited range of verbs)					
FORM: WITH 'WHEN'					
FORM: AFFIRMATIVE (wide range of verbs)					
FORM: NEGATIVE (increasing range of verbs)					
FORM: QUESTIONS: yes/no, wh-, negative forms (limited range of verbs)					
USE: HABITUAL STATES OR ACTIONS					
USE: IMAGINED SITUATIONS AFTER 'IF'					
USE: ORDERING OF PAST EVENTS					
USE: REGRET					
FORM: NEGATIVE					
FORM: QUESTIONS: yes/no, wh-, negative forms (increasing range of verbs)					
FORM: WITH SUBORDINATING CONJUNCTIONS					
FORM: WITH TIME ADJUNCTS					
USE: POLITENESS AFTER 'IF'					
USE: POLITENESS: 'I WONDERED' AND 'I WANTED'					
FORM: INVERSION WITH 'NOT ONLY… BUT ALSO'					
FORM: QUESTIONS: yes/no, wh-, negative forms (wide range of verbs)					
USE: COMPLEX ORDERING OF PAST EVENTS					
USE: FOR EMPHASIS, WITH 'DID'					
USE: POLITENESS: 'I THOUGHT'					

This illustrates a developmental pathway, one of the key findings of the study, that learners know a syntactic pattern or structure early on in their learning, and their competence increases as a developing repertoire of lexis and functions and lexis, including pragmatic meaning. Table 22.4 shows an illustration of this development from http://www.englishprofile.org/english-grammar-profile/egp-online at levels A1, A2, B1 and B2.

Table 22.4 Extracts from the *English Grammar Profile* of past simple development

Level	*Guideword, Can-do statement and Examples*
A1	USE: EVERYDAY EVENT AND STATES Can use the past simple to talk about everyday events or states. *The people were very polite* (A1, French). *Every time I went to Hendon Park, I saw so many people there* (A1, Polish).
A2	FORM: AFFIRMATIVE Can use the affirmative form with an increasing range of verbs. *I forgot to tell you some details about tomorrow evening* (A2, Turkish). *I spent £35 on all those clothes* (A2, Spanish - European).

(*Continued*)

Table 22.4 (Continued)

Level	*Guideword, Can-do statement and Examples*
B1	USE: ORDERING OF PAST EVENTS Can use the past simple to order sequences of events in the past, in the context of narratives. *I remember her as a shy girl but I read she became a manager, she went on lots of trips, she met a lot of new friends, she got married and then she got divorced* (B1, Italian). *We waited for the suitcase for two days and then we bought new dresses* (B1, Czech).
B2	USE: POLITENESS: 'I WONDERED', 'I WANTED' Can use the past simple with 'I wondered' and 'I wanted' as politeness structures, when making polite requests and thanking. *So, I wondered if you could introduce me to somebody who knows how to use a camera, so that I will be able to record my trip for you* (B2, Greek). *I wanted to know if the rooms are single or double, if they have showers and if there is room service?* (B2, Spanish - European).

Another key insight from the study was evidence of competence at odds with traditional expectations of grammar teaching and learning. For example, learners demonstrated use of 27 different structures to express conditionality across the levels, in contrast with the prescriptive paradigm of first, second and third conditionals typically expected and taught in grammar teaching syllabi.

4 The story so far: what has corpora told us about language learning?

Corpora so far have managed to offer researchers and teachers powerful insights into usage that has widened our understanding of constructs such as interlanguage. As Granger (2015) has observed, CIA and learner corpora have advanced several new dimensions of study, particularly in the areas of lexis and discourse. Language learning professionals now have access to studies that give them a comprehensive picture of L1 and L2 varieties through the analysis of discrete language features, collocations, lexical bundles, colligations and collostructions, which have considerably enhanced our knowledge of the L2 phrasicon. Corpus linguistics research methods offer important analytical affordances to those interested in the analysis of lexicogrammatical features with an emphasis on frequency and distribution of language items. The main contribution of corpora to language learning, therefore, is an increased understanding of the linguistic outputs produced by learners when engaged in a wide range of tasks and interactions across different media of delivery. We can use these outputs as evidence of a variety of learning factors such as the effect of L1 input, frequency of certain items in L2 input, saliency of discrete linguistic features in both L1 and L2 language and L1 influence. The evidence found in the corpora can inform language learning pedagogy and practice in many different ways (as demonstrated by other chapters in this section) and thus promote reflexivity in learners and teacher education and practice.

The three case studies in this chapter exemplify some important insights with implications for learning. The two longitudinal approaches showcased in the first case study display some of the most interesting contributions of corpora to language learning: tracking development. Whether development is located at the grammatical, the lexical or

the lexicogrammatical level is for the language teachers and researchers to decide. In any case, they represent the need for a solid understanding of the methods used and show how small-scale approaches are useful in combining automatic analysis with manual tagging. A common thread through these representative studies is that they provide a glimpse of learning; they reveal a tip of the iceberg and raise many more questions not just about language learning theories, e.g. order of acquisition theory (Murakami 2014), explicit vs. implicit knowledge, teaching and learning (Ellis 2015), but also language learning contexts, e.g. classroom instruction, data-driven learning (see Chapter 29, this volume), the importance of input and its dynamic nature and availability in the digital age. Murakami (2014), which challenges the long-held belief of fixed L2 order of acquisition, is one such example demonstrating the central role that LCR is now playing in our understanding of learning. The second case study illustrates that the use of well-designed corpora allows researchers to understand monologic and dialogic communication as they reveal aspects of frequency, collocation, colligation, function and speaker variation that would otherwise remain hidden. The range of roles played by adverbs and their wider linguistic contexts reveal speakers' codification of meaning that goes from boosting to hedging, from disagreement to topic shift, from minimising to the expression of solidarity. Lack of exposure and awareness of the colligational frames where adverbs occur in conversation (Carter and McCarthy 2017; Hunston 2019) expressing stance-related meanings may contribute to L2 speakers' lack of understanding of the meanings construed and the most common lexical items integrating them. These studies both exemplify and emphasise the importance of the form–function relationship in language learning and challenge long-held beliefs of what to teach when. McCarthy (2020) describes the wider pedagogical impact of this kind of research, highlighting evidence of "grammaring" (Larsen-Freeman 2003) and the acquisition of grammar as a constant creative process.

We have seen the affordances offered by learner corpora and now turn to future possibilities. Does the nature of the data, the predominance of cross-sectional over longitudinal corpora, and the methodologies applied, the contexts of collection, inevitably result in studies offering narrow, static descriptions of language use, heavily reliant on quantitative approaches? What are we missing? Are there specific views of learner corpus research that have led us down the path of description but fallen short of what this means for learning, and how we might apply this to pedagogical settings?

5 Ways ahead: harnessing the potential

One of the main criticisms of existing corpora is that they fail to provide fine-grained data that can account for how the *learning* of the attested language actually happened, for example, in terms of the implicit vs. explicit nature of language learning, the presence of prototypical features in the L2 input or cognitive-related features associated with the processing of the L2. This criticism was echoed by McEnery *et al.* (2019), who voiced the aspiration to compile learner corpora that can resonate with a wider range of researchers and theoretical positionings.

LCR is relatively young; it is and should be a work in progress. The design elements in Table 22.2 may paint a picture of language learning that is biased towards a narrow conceptualisation. Speech and its unfolding development are the primary focus of L1 acquisition studies. Most L2 corpora are of written language. In the minority of spoken learner corpora available, there is little attempt to align spoken forms with anything other than

written norms of the target language. Most learner corpora assume monolingual realities by grouping learners attending to one native tongue and by identifying one L2 target. This may point to a monolingual bias in the way the data are collected. Alternatively, we may think that the multilingual turn (Douglas Fir Group 2016) will take some time to be visible in how learner language is collected and how it is factored into corpus design. Also, having groups of L1 speakers learning an L2 immediately visualises the native language of the learners as the independent variable in most research designs, leaving aside other considerations that are crucial in language learning, as we have seen in the previous paragraphs. For example, if we decided to embrace the view that language learning is a complex adaptive system (Beckner *et al.* 2009), which explains the nonlinearity nature of language learning and the existence of so-called phase transitions, existing corpora may only be capable of providing some of the data that are required to understand these changes. It is expected that future developments in artificial intelligence (AI), machine learning and data processing will bring together new ways of looking at learner language development, offering more opportunities to analyse learner language using larger datasets and adaptive systems (e.g. Ballier *et al.* 2020). Expected outcomes will include more attention to adaptive learning, self-directed language learning and automatic assessment.

The emphasis on the comparison with "native-speaker" data has probably driven most of the findings in CIA and LCR to a narrow conceptualisation of language learning that is not currently massively endorsed by SLA researchers or L2 teaching professionals.

Perhaps this is the main challenge ahead for learner corpora: to imagine a method (a discipline perhaps?) that builds on Granger's (2015) reconceptualisation of CIA, but at the same time can successfully relate to current language learning characterised by the overlapping of material and digital social contexts in a multilingual world (Douglas Fir Group 2016; Ortega 2013) across language learning theories. For example, O'Keeffe (2021) points out how learner corpora are increasingly being explored so as to identify the process of construction development across levels of competence in usage-based accounts of language learning (see Chapter 23, this volume). Complementarity with other research methods and theories presents a good opportunity to increase the spread and the relevance of corpora in language learning. And this must be done in a context where learning goals are re-assessed in order to capture the multilingual nature of both learning and communication, both spoken and written, including attention to the learner as a person, the formation and transformation of identities (Leung and Scarino 2016) and the recognition of the situatedness of the learning and use beyond the analysis of errors.

Our current and future challenges are discovering ways of capturing, representing and understanding this in its broadest, deepest and most inclusive conceptualisations. Even when it is possible to measure general tendencies and find recurrent patterns in large-scale data and look at general frequency, distribution, collocational and colligational pattering, variation is always below the surface. McEnery *et al.* (2019: 84) call for a move beyond the 'immediately discoverable'. We suggest embracing this position as the starting point to look ahead.

Further reading

Carter, R. A. and McCarthy, M. J. (2017) 'Spoken Grammar: Where Are We and Where Are We Going?', *Applied linguistics* 38 (1): 1–20. (This review argues for a re-thinking of grammatical description based on spoken corpus evidence and considers development in digital communication and implications for description and pedagogy.)

Díez-Bedmar, M. B. (2018) 'Fine-Tuning Descriptors for CEFR B1 Level: Insights from Learner Corpora', *ELT Journal* 72(2): 199–209. (A conceptualisation of CEFR performance levels using learner data and a reformulation of the grammatical accuracy descriptor at the B1 level to raise learner awareness of frequent errors.)

Kyle, K. and Crossley, S. (2018) 'Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices' *The Modern Language Journal* 102(2): 333–49. (A study of the predictive validity of three types of syntactic complexity indices related to clausal and phrasal complexity in the TOEFL exam, which suggests that more attention should be paid to phrases, in particular noun phrases, in language education.)

References

Ädel, A. (2015) 'Variability in Learner Corpora', in S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, pp. 379–400.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T. and Gaillat, T. (2020) 'Machine Learning for Learner English: A Plea for Creating Learner Data Challenges', *International Journal of Learner Corpus Research* 6(1): 72–103.

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Ke, J., Christiansen, M. H., Larsen-Freeman, D., Croft, W., Schoenemann, T. and Five Graces Group (2009) 'Language is a Complex Adaptive System: Position Paper' *Language Learning* 59: 1–26.

Bley-Vroman, R. (1983) 'The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity', *Language learning* 33(1): 1–17.

Buttery, P. and Caines, A. (2010) 'Normalising Frequency Counts to Account for "Opportunity of Use" in Learner Corpora', in Y. Tono, Y. Kawaguchi and M. Minegishi (eds) *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, Amsterdam: John Benjamins, 187–204.

Callies, M. (2015) 'Learner Corpus Methodology', in S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, pp. 35–55.

Capel, A. (2012) 'Completing the English Vocabulary Profile: C1 and C2 Vocabulary', *English Profile Journal* 3(1): 1–14.

Carter, R. A. and McCarthy, M. J. (2017) 'Spoken Grammar: Where Are We and Where Are We Going?', *Applied linguistics* 38(1): 1–20.

Cook, V. (1991) 'The Poverty-of-the-Stimulus Argument and Multicompetence', *Second Language Research* 7: 103–17.

Corder, S. P. (1967) 'The Significance of Learners' Errors', *International Review of Applied Linguistics* 5: 161–70.

Dornyei, Z. (2007) *Research Methods in Applied Linguistics*, Oxford: Oxford University Press.

Douglas Fir Group (2016) 'A Transdisciplinary Framework For SLA in a Multilingual World', *Modern Language Journal* 100: 19–47.

Ellis, N. C. (2015) 'Implicit and Explicit Language Learning: Their Dynamic Interface and Complexity', in P. Rebuschat (ed.) *Implicit and Explicit Learning of Languages*, Amsterdam: John Benjamins, pp. 1–24.

Ellis, N. C., Römer, U. and O'Donnell, M. B. (2016) *Usage-Based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*, Malden, MA: Wiley.

Ellis, R. and Barkhuizen, G. P. (2005) *Analysing Learner Language*, Oxford: Oxford University Press.

Gablasova, D., Brezina, V. and McEnery, T. (2017) 'Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence', *Language Learning* 67(S1): 155–79.

Granger, S. (1994) 'The Learner Corpus: A Revolution in Applied Linguistics', *English Today* 10(3): 25–33.

Granger, S. (2009) 'The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching', in K. Aijmer (ed.) *Corpora and Language Teaching*, Amsterdam: John Benjamins Publishing, pp. 13–32.

Granger, S. (2015) 'Contrastive Interlanguage Analysis: A Reappraisal', *International Journal of Learner Corpus Research* 1(1): 7–24.
Green A. (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range. English Profile Studies 2*, Cambridge: Cambridge University Press.
Greenbaum, S. (1990) 'Standard English and the International Corpus of English', *World Englishes*, 9(1): 79–83.
Harrison, J. and Barker, F. (eds) (2015) *English Profile in Practice. English Profile Studies 5*, Cambridge: Cambridge University Press.
Hawkins, J. A. and Buttery, P. (2010) 'Criterial Features In Learner Corpora: Theory And Illustrations', *English Profile Journal* 1(1): 1–23.
Hunston, S. (2019) 'Patterns, Constructions, and Applied Linguistics', *International Journal of Corpus Linguistics* 24(3): 324–53.
Larsen-Freeman, D. (2003) *Teaching Language: From Grammar to Grammaring*, Boston, MA: Heinle.
Leung, C. and Scarino, A. (2016) 'Reconceptualizing the Nature of Goals and Outcomes in Language/s Education', *The Modern Language Journal* 100(S1): 81–95.
Mark, G. and Pérez-Paredes, P. (2018) 'Examining High Frequency Adverbs in Learner and Native Speaker Language: Some Implications for Spoken EFL Learning and Teaching', *13th Teaching and Language Corpora conference (TaLC 2018)*, Faculty of Education, University of Cambridge, July 2018.
McCarthy, M. J. (2020) *Innovations and Challenges in Grammar*, London: Routledge.
McEnery, T., Brezina, V., Gablasova, D. and Banerjee, J. (2019) 'Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use', *Annual Review of Applied Linguistics* 39: 74–92.
Mitchell, R., Myles, F. and Marsden, E. (2013) *Second Language Learning Theories*, London: Routledge.
Murakami, A. (2014) *Individual Variation and the Role of L1 in the L2 Development of English Grammatical Morphemes: Insights from Learner Corpora*, unpublished PhD thesis, University of Cambridge.
Myles, F. (2005) 'Interlanguage Corpora and Second Language Acquisition Research', *Second Language Research* 21(4): 373–31.
Myles, F. (2015) 'Second Language Acquisition Theory and Learner Corpus Research', in S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research*, Cambridge: Cambridge University Press, pp. 309–31.
O'Keeffe, A. (2021) 'Data-Driven Learning - A Call for a Broader Research Gaze', *Language Teaching* 54(2): 259–72
O'Keeffe, A. and Mark, G. (2017) 'The English Grammar Profile of Learner Competence. Methodology and Key Findings', *International Journal of Corpus Linguistics* 22(4): 457–89.
Ortega, L. (2013) 'SLA for the 21st Century: Disciplinary Progress, Transdisciplinary Relevance, and the Bi/Multilingual Turn', *Language Learning* 63: 1–24.
Paquot, M. and Plonsky, L. (2017) 'Quantitative Research Methods and Study Quality in Learner Corpus Research', *International Journal of Learner Corpus Research* 3(1): 61–94.
Pérez-Paredes, P. (2019) 'English Language Teacher Education and Second Language Acquisition', in S. Walsh and S. Mann (eds) *Routledge Handbook of English Language Teacher Education*, London: Routledge, pp. 253–67.
Pérez-Paredes, P. and Díez-Bedmar, B. (2019) 'Certainty Adverbs in Spoken Learner Language: The Role of Tasks and Proficiency', *International Journal of Learner Corpus Research* 5(2): 253–79.
Pérez-Paredes, P. and Bueno, C. (2019) 'A Corpus-Driven Analysis of Certainty Stance Adverbs: *Obviously, Really* and *Actually* in Spoken Native and Learner English', *Journal of Pragmatics*, 140, 22–32.
Pérez-Paredes, P., Mark, G. and O'Keeffe, A. (2020) *The Impact of Usage-Based Approaches on Second Language Learning And Teaching*, Cambridge: Cambridge University Press, retrieved from: https://www.cambridge.org/us/educationreform/insights.

Römer, U. and Garner, J. R. (2019) 'The Development of Verb Constructions in Spoken Learner English: Tracing Effects Of Usage and Proficiency', *International Journal of Learner Corpus Research* 5(2): 206–29.

Selinker, L. (1972) 'Interlanguage', *International Review of Applied Linguistics* 10(3): 209–31.

Siyanova-Chanturia, A. and Spina, S. (2019) 'Multi-Word Expressions in Second Language Writing: A Large-Scale Longitudinal Learner Corpus Study', *Language Learning*, 70(2): 420–63.

Thewissen, J. (2013) 'Capturing L2 Accuracy Developmental Patterns: Insights from An Error Tagged Learner Corpus', *The Modern Language Journal* 97: 77–101.

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

Tono, Y. and Díez-Bedmar, M. B. (2014) 'Focus on Learner Writing at the Beginning and Intermediate Stages: The ICCI Corpus', *International Journal of Corpus Linguistics* 19(2): 163–77.

Tyler, A. and Ortega, L. (2018) 'Usage-Inspired L2 Instruction an Emergent, Researched Pedagogy', in A. Tyler, L. Ortega, M. Uno and H. I. Park (eds) *Usage-Inspired L2 Instruction: Researched Pedagogy*, Amsterdam: John Benjamins, pp. 3–26.

Vyatkina, N. (2013) 'Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus', *The Modern Language Journal* 97(S1): 11–30.