11 Learner corpora and grammar

Tom Rankin

1 Introduction

This chapter aims to provide an overview of analyses of properties of second/foreign language (L2) grammar(s) using learner corpora. Learner corpus research (LCR) on grammar is relatively underrepresented in the literature in comparison to studies of lexis. This may be due to the difficulties related to the automatic identification and analysis of syntactic structures in corpora (see Gilquin 2002). Nevertheless, various studies have examined a range of grammatical phenomena in a variety of learner corpora. These reflect various L1 (first language)–L2 pairings at various levels of proficiency. However, given the social, educational and economic position of English as a second or foreign language, there is a tendency for English to predominate as a target language. Given the development of the field of LCR, advanced levels of proficiency have been more extensively studied using learner corpora than have lower proficiency levels. This contrasts with studies of grammar in the wider field of second language acquisition (SLA).

In addition to this empirical diversity, a similar situation exists with respect to the theoretical and conceptual landscape of LCR on grammar. There is no theoretically monolithic approach to what might be called ‘learner corpus grammar’ and learner corpus studies of grammar have come from diverse theoretical perspectives. Nevertheless, given developments in the wider field of corpus linguistics, there is a certain predisposition towards broadly functional, usage-based models of grammar in LCR (see Meunier 2008). While the discussion which follows includes consideration of theoretical issues, the focus will be on methodological and empirical issues in LCR without a specific commitment to any one theoretical position. The survey should thus serve as a guide to corpus research that has been (and can be) carried out to study L2 grammar.
Before embarking on the survey, however, it is useful to situate learner corpus research on grammar at the interface of neighbouring fields. Granger (2008b: 272) points out that LCR ‘sits at the crossroads’ between corpus linguistics, second language acquisition and foreign language teaching and that while it has found a secure foothold in corpus linguistics, it still has only a limited influence in the other fields (see Chapter 14, this volume, on the link between LCR and SLA). This somewhat uncertain status of LCR with respect to neighbouring fields is particularly relevant for the study of L2 grammar. The study of L2 grammar has long been a central concern in SLA. Production data from L2 learners has, in turn, been an important empirical resource in the field. Granger (2004: 123) points out that ‘[t]here is nothing new in the idea of collecting learner output. Both FLT [foreign language teaching] and SLA researchers have been collecting learner output for descriptive and/or theory-building purposes since the disciplines emerged’. One might then wonder where to draw a dividing line (and indeed whether there even is a dividing line) between SLA production studies of grammar and LCR on grammar.

A brief sketch of some major SLA data-collection projects will serve as an initial point of comparison. The design of these projects and research on the data illustrate where there are methodological differences with LCR. I confine the discussion here to just a brief consideration of the most important such studies in order to juxtapose these with LCR discussed in the following sections.

The ZISA1 project collected and transcribed longitudinal oral data from five adult learners with low proficiency L2 German with Italian and Spanish as their native languages (see Clahsen et al. 1983). The European Science Foundation Project (ESF, Perdue 1993) collected data for five target languages (English, German, Swedish, French and Dutch) from forty learners with Arabic, Finnish, Italian, Punjabi, Spanish or Turkish as L1. The learners were acquiring the target languages naturally in the L2 environment. They provided spoken data from naturalistic interaction as well as more controlled story-retelling tasks. The data remains publicly available through the CHILDES database (MacWhinney 2000).

This data has formed the basis of many important approaches to L2 grammar. On the basis of the ESF data, Klein and Perdue (1997) formulated the concept of the Basic Variety, which states that naturally acquired interlanguage is constrained by phrasal, semantic and pragmatic constraints. This functional perspective sees learner languages as constituting varieties in their own right, largely independent of specific grammatical properties of L1 and L2. L2 grammatical organisation is efficient for communicative purposes. ESF and ZISA data has also been used to support claims for different models of the L2 initial state in generative approaches to SLA.

---

1 ZISA: Zweitspracherwerb Italienischer und Spanischer Arbeiter ("Acquisition of German by Italian and Spanish Workers").
data from low-proficiency learners provides evidence relevant to access to universal grammar and the extent of L1 parametric transfer to the L2; for example, Schwartz and Sprouse’s (1994) Full Transfer / Full Access hypothesis and Vainikka and Young-Scholten’s (1996) Minimal Trees hypothesis. While these have been landmark projects in SLA, they are not normally considered learner corpus studies and are generally smaller in scope than the learner corpus projects which emerged in the 1990s and 2000s.

2 Core issues in LCR on grammar

This is obviously not an appropriate forum in which to delve deeply into general issues surrounding the relationship between corpus linguistics and theories of grammar. Such issues have been the subject of long-standing and ongoing debate (see, for example, Aarts 2000). But given that this discussion persists in the wider field of corpus linguistics, a brief consideration of issues is nevertheless helpful to situate the approach to grammar in LCR, and set the scene for the discussion to follow. Section 2.1 concentrates on conceptual issues with respect to theories of grammar. Section 2.2 then examines methodological trends and some of the main findings in LCR.

2.1 Corpus linguistics and theories of grammar

Corpus linguistics is fraught with definitional and conceptual issues. Taylor (2008) poses the question ‘what is corpus linguistics?’ and explores how the term is defined and used by scholars in corpus linguistics research papers. She concludes that in the literature ‘there are radical differences in the representation and understanding of what corpus linguistics is’ (Taylor 2008: 196). Corpus linguistics is variously conceptualised as ‘a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm (theoretical or methodological), or a combination of these’ (Taylor 2008: 180).

Such conceptual issues about the status of corpora and corpus linguistics are particularly acute with respect to grammar. There has been a certain tension between corpus linguistics and grammatical theory (especially generative theory). Chomsky has indeed claimed that corpus linguistics ‘does not exist’ (Aarts 2000: 5). This perhaps goes some way to explaining why SLA of a generative theoretic bent has shied away from using corpora and corpus-linguistic methodology.

The corollary of this is that corpus linguistics is more readily identified with certain other theories of grammar. Gries (2012a), for example, observes that there are many points of convergence between corpus linguistics, on the one hand, and cognitive linguistics and construction grammar, on
the other. He formulates a proposal for corpus linguists ‘to assume as the main theoretical framework within which to explain and embed [their] analyses a psycholinguistically-informed, (cognitively-inspired) exemplar/usage-based linguistics’ (Gries 2012a: 56–7).

Irrespective of whether one agrees with Gries’s views or not, the empirical insights from corpus linguistics studies have certainly been a driving force in the development of approaches to grammar which would be broadly in line with a cognitive-linguistic/constructionist perspective. Examples of such approaches include Sinclair’s (1991) Idiom Principle, Hunston and Francis’s (2000) Pattern Grammar and Sinclair and Mauranen’s (2006) Linear Unit Grammar. Such frameworks assume specific conceptual approaches to grammar which share broad affinities with cognitive/constructionist models. For example, they tend to investigate corpora without hypotheses derived from a theory of grammar. The syntactic regularities emerge from the study of large data sets of natural language use. There is no strict dichotomy between the lexicon and syntax; rather, particular syntactic patterns are associated more or less closely with lexical items or specific classes of lexical items.

Given these affinities between the wider field of corpus linguistics and cognitive/constructionist theories of grammar, it is also possible to identify a certain predisposition to such models of grammar in LCR (see Wulff and Gries 2011; Gilquin 2012). Meunier (2008) discusses these affinities in terms of the possible application of cognitive and corpus linguistics in pedagogical grammar. She points to the fact that corpus linguistics and cognitive linguistics both reject dichotomies of performance vs competence and syntax vs lexicon. Similarly, frequency plays a central role in both fields. It is a key element in cognitive-linguistic notions of entrenchment and conventionalisation, while in corpus linguistics, recurrence is ‘a key to typicality, and forms the basis of analyses of lexical and/or grammatical features of varieties of a language’ (Meunier 2008: 93).

However, learner corpus studies of grammar have come from heterogeneous conceptual and theoretical perspectives (see below). A tendency for corpus linguists to take a cognitive/constructionist line does not preclude the use of learner corpora as an empirical foundation by researchers studying L2 grammar from other perspectives (for example Oshita 2000, 2004; Lozano and Mendikoetxea 2010). It can therefore be claimed that regardless of controversies in corpus linguistics more generally, learner corpora are a methodological resource in second language research rather than an independent theoretical approach. This is in line with Aarts’s (2000: 7) claim that corpus linguistics refers to a methodology in linguistics and while ‘theoretical linguists couch their work in the terms of a particular theoretical model[,] corpus linguists may or may not use models, the important point is that they use data from corpora in their claims about language’.
2.2 Defining core learner corpus research on grammar

As already noted, there is nothing new in investigating L2 grammar on the basis of production data. The question posed in the introduction was what distinguishes production studies from learner corpus studies. There is, in fact, no clear dividing line between the two; rather, there are several aspects which could be considered more typical of one or the other approach. So, while production studies tend to draw on data from learners acquiring an L2 naturalistically and usually collect spoken data, LCR studies’ data tends to be written production by instructed learners. LCR also concerns itself more with pedagogical issues connected to the teaching of grammar, while production studies have tended to use the data to answer questions about the nature of an L2 grammatical system with respect to predefined linguistic theories. In connection with this, LCR shows a preoccupation with usage, i.e. how grammatical forms pattern lexico-grammatically and are influenced by register variation, and the pragmatic use to which they are put by learners in comparison to native speakers. In addition, issues of quantitative over- and underuse of features of the target language in comparison to native usage have a central importance rather than just qualitative patterns of divergent or target usage.

Finally, there is also a tendency for the methodological approach to differ. Again, however, this is not a clear dividing line. This can be illustrated by drawing on Gries’s (2006: 4–5) category and prototype definition of corpora, whereby there are prototypical and more peripheral members of the category of ‘corpus studies’. ‘There are a few criteria that are – though not individually necessary – shared by much, if not most, work within corpus linguistics’ (Gries 2006: 4). These core criteria include the following:

1. The analysis is based on machine-readable corpora of naturally occurring language.
2. The corpus is intended to be balanced and/or representative of the variety the study is aimed at.
3. The analysis aims to be systematic and exhaustive.
4. The analysis uses statistical data to cover the middle ground between what is possible/grammatical and what is not.
5. The analysis proceeds on the basis of frequency lists, concordance lines in which the word of interest is shown in its natural context, and collocations.

(Gries 2006: 4)

Based on this framework, Gilquin and Gries (2009: 6) claim that ‘there is actually no strict corpora-experiments dichotomy. Rather, just as linguistic data in general forms a continuum of naturalness of production/collection, so do corpora: they vary along the above dimensions [in 1–5], which results in a continuum ranging from prototypical corpora via less
typical corpora to corpora whose compilation is distinctly experimental in nature’ (see also Chapter 2, this volume). So, ‘prototypical’ learner corpus research tends to use corpus technology such as taggers, parsers and concordancers. In addition, the corpus methodology itself may be discussed as an object of study. In contrast, SLA production work on grammar concentrates on the morphosyntactic findings relevant to particular theoretical questions, without necessarily explicitly discussing issues of corpus collection, annotation and analysis. Indeed, Granger (2004: 134) has pointed out that the term ‘corpus’ is ‘rarely found in SLA books and articles’.

2.2.1 Trends in learner corpus research on grammar
While ‘core’ learner corpus studies, as defined above, tend to have properties in common which distinguish them from other production studies, there is no core theoretical or conceptual issue which unifies learner corpus studies of grammar. It is, therefore, difficult to provide a coherent narrative of the development of the field. The following aims to highlight those common elements in approach and methodology which define the core field of learner corpus research.

A strong methodological tendency in learner corpus studies of grammar is the use of Contrastive Interlanguage Analysis (CIA, Granger 1996), which permits, among other things, the identification of transfer by comparing L2 production from different L1 groups. It also identifies patterns of divergent production as well as over- and underuse by employing comparisons with native-speaker production. Various studies discussed below (and in Section 3) illustrate variations on the use of this methodology.

A major defining feature of LCR is the prominent role of methodological discussion and issues connected to the design and analysis of corpora (see also Chapters 2 and 3, this volume). This applies to corpus studies of L2 grammar to the extent that the automatic identification and analysis of complex grammatical features poses a significant computational problem. There has, therefore, been discussion of the procedures which may be used to facilitate analysis of L2 morphosyntax. Meunier (1998) provides a general overview of tools for (semi-)automatic analysis of learner corpus data. However, she points out that there is a dearth of automated quantitative and qualitative studies of English syntactic structures, and ‘if they are long overdue for native English, they are virtually non-existent for ESL [English as a Second Language] and EFL [English as a Foreign Language]’ (Meunier 1998: 35). Gilquin (2002: 195) similarly asserts that parsing ‘has not quite reached a mature level’ and that studies of grammar have thus had to turn to lexical search methods on POS (part-of-speech)-tagged or raw corpora to study syntactic phenomena. She also illustrates how search algorithms on POS-tagged corpora can be used to extract complex grammatical phenomena (causative uses of make in this case). As with native corpora of English, there remains a gap in the
Learner corpora and grammar

learner corpus landscape for parsed corpora (see Chapter 5, this volume). Therefore, studies have in the main had to use POS-tagged or raw corpora with a great deal of manual editing of results (see Section 3).

Examples of the use of POS-tagged learner corpora are provided by Aarts and Granger (1998) and Borin and Prütz (2004). Aarts and Granger (1998) use the L1 Dutch, Finnish and French subcorpora of the *International Corpus of Learner English* (*ICLE*, Granger et al. 2009) in comparison with the *Louvain Corpus of Native English Essays* (*LOCNESS*). The corpora were tagged with the *TOSCA* tagger and tag trigrams were extracted. Statistical comparisons showed that the frequencies of trigrams in each of the learner corpora differed from the native corpus in similar ways and similar patterns were over- or underused; for example, the ways in which learners start sentences differ from the native speakers, with a greater frequency of sentence-initial connectors and an underuse of subject-initiated sentences. Borin and Prütz (2004) build on this type of study by expanding the range of n-grams investigated to 1- to 4grams in the *Uppsala Student English Corpus*, the *Stockholm-Umeå Swedish Corpus* and the *British National Corpus Sampler*. However, comparison with Aarts and Granger’s (1998) results illustrates issues with this sort of research. Although no clear pattern emerges, it is suggested that contrastive POS n-gram studies can yield insights into differences and similarities between different language varieties. Borin and Prütz (2004: 82) also caution that ‘POS tag sequences are of course not syntactic units; they merely give better clues to syntax than word level investigations are able to provide, so that the picture we get of learner (and native speaker) language syntax is distorted and needs careful interpretation to be usable’.

In addition to straightforward comparisons between learners’ native language, target language and interlanguage, more complex comparisons have been undertaken in corpus studies of L2. The issue of usage and frequency is emphasised where a register perspective is taken into account. Biber and Reppen (1998: 145) point out that ‘corpus-based analyses show that linguistic association patterns are generally not valid for the language as a whole; rather, there are usually striking differences in the use of grammatical features across registers’. Drawing on a study of complement clauses in the native English *Longman Grammar Corpus* and *Longman Learners’ Corpus* (academic prose by L1 French, Spanish, Chinese, Japanese learners of English), they show that patterns of usage are similar across the learner corpora. These patterns are also similar to native usage in conversation and fiction, but different from native academic prose.

Issues of pedagogy are never far from these sorts of studies and a concern with pedagogy appears in various guises in a range of studies of grammar. A motivation for Biber and Reppen’s (1998: 147) study, for example, is to fill a gap left by EFL/ESL reference works and textbooks, which tend not to address questions such as the relative frequency of different types of complement clauses, factors which condition the choice between structural
variants, etc. This reflects a strong tendency in the wider field of corpus linguistics to connect language description with the development of reference or instructional materials (e.g. the Cobuild and Longman dictionaries and reference works). Much learner corpus research on grammar has sought to expand on this by adding a learner perspective. The basic insight is that a learner corpus study provides relevant information about grammatical difficulties at different proficiency levels in different L1–L2 pairings. Such data can then be used to inform pedagogical practices or materials (see Chapter 22, this volume). Hinkel (2004) illustrates this approach on the basis of tense and aspect usage in L2 English. She shows that issues with respect to usage of verbal forms persist at higher levels of proficiency in learner academic writing. This could be addressed by dealing with grammatical issues with respect to tense, aspect and passive usage in the relevant academic discourse context.

This approach has also been used to address those areas where there may be a gap in the standard instructional materials. Palacios-Martínez and Martínez-Insua (2006) point out that the usage of there-constructions is not usually the subject of specific attention in grammar teaching, but they tend to pose a problem in L1 Spanish–L2 English. Rankin and Schiftner (2011) investigate sentence structures with marginal prepositions in L1 German–L2 English and use learner corpus data to develop teaching materials which address their usage in order to fill a gap in the available instructional materials.

Error analysis may also be used to address issues in acquisition and grammar pedagogy. Chuang and Nesi (2006) base their error analysis of L1 Chinese–L2 English on the finding that university students attending pre-session academic English courses in Britain want more explicit grammar teaching than is included in the courses. They use error analysis to identify the types of grammatical errors that are most prevalent in the written production of the students. Errors in the article system are most consistently problematic for the learners. This information is used to develop teaching units for an online grammar course (see also Chapter 22, this volume).

The pedagogical aspect also occurs as an explicit component of the CIA model, where pedagogical materials are included as an independent point of comparison with learner and/or native production. Rankin (2010) compares materials used to teach adverb placement in an ESP (English for Specific Purposes) course with the production of adverb placement patterns in the writing of L1 German–L2 English learners who take part in the course. Tono (2004) investigates the acquisition of subcategorisation frame patterns of English verbs by L1 Japanese speakers on the basis of spoken and written production from the Japanese English as a Foreign Language Learner (JEFLL) Corpus and a native Japanese corpus. He also compiles a corpus of Japanese EFL textbooks aimed at different educational levels. He finds strong effects of the frequency of occurrence of structures.
in the textbooks and over- and underuse of patterns by learners, illustrating the importance of teaching effects and the complex interaction of factors in instructed SLA.

2.2.2 L2 grammar and learner corpora for SLA research

The previous sections compared SLA production studies of L2 grammar with prototypical learner corpus studies. Myles (2005, 2008) outlines criticism of the approach of each of these fields in isolation from the other and calls for more synergy between them (see also Chapter 14, this volume). In essence, she proposes that SLA would benefit from using larger, machine-readable databases supported by computational tools, while LCR would benefit from greater engagement with SLA theory. Myles (2005: 380–1) sums up her thoughts as follows:

Most of the studies using corpora make little use of software other than concordancing, and remain for the most part rather unambitious in their use of new technology. They also often remain rather descriptive, documenting differences between learner and native language rather than attempting to explain them, and the developmental dimension is almost totally lacking. Corpus-based L2 studies are also often not sufficiently informed by SLA theory, and tend to assume that finding out differences in use between learners and native speakers will have direct pedagogical implications, which is of course not necessarily the case. Such research is useful nonetheless, as we need to have good descriptions of learner language in order to inform our understanding of what shapes its development, but it is now time that corpus linguists and SLA specialists work more closely together in order to advance both their agendas.

This criticism could be seen as a more specific elaboration of the long-recognised problem that merely describing language production does not tell us a great deal about the grammatical competence which underpins that production. This is particularly the case with written production, which may bear more of the hallmarks of learned meta-grammatical knowledge and monitoring processes, making spoken corpora more appropriate for fundamental SLA work on grammar (Myles 2005: 375). LCR on grammar must always beware of merely providing a descriptive snapshot of some grammatical phenomenon as produced by a (group of) learner(s) at a specific point in time. In this light, one can sympathise with the call for engagement with explanatory theories, greater use of developmental data, and more use of technology to facilitate fine-grained analysis of grammatical properties.

Myles’s research groups have sought to address these problems in the form of the FLLOC and SPLLOC projects (French/Spanish Learner Language Oral Corpora). These are corpora of longitudinal oral production of instructed

---

2 For more on the make-up of the corpora, see www.flloc.soton.ac.uk and www.splloc.soton.ac.uk (last accessed on 13 April 2015).
L2 French and Spanish by L1 English speakers. They are available through the TalkBank database and follow the transcription and storage conventions of CHILDES. The data from these projects have been used in a variety of studies of L2 grammatical development (see the project web pages for full publications lists). However, the data collected include production elicited as the result of experimental techniques, illustrating the more experimental end of the corpus–experiment spectrum outlined above.

Myles’s criticisms, of course, have never applied wholesale to LCR. On the issue of a lack of SLA theory, for example, Oshita (2000, 2004) draws on generative SLA theories in studies of passivisation and null expletives in English in corpora of learner writing. The lack of a developmental perspective does not apply to work by Housen (2002), who studies the development of English tense and aspect forms and usage across four proficiency levels by native speakers of French and Dutch. What is more, Housen uses transcripts of spoken production, thereby addressing the issue of overreliance on written corpora.

More recent work can also be seen to address the problems Myles alludes to. Collentine and Collentine (2013) make ambitious use of new technology in their study of L2 Spanish structural alignment in interaction. Their corpus is made up of archived chat from learner interactions in synchronous computer-mediated communication. It should be noted also that this approach could provide an interesting solution to the spoken/written problem given the more spontaneous and interactional nature of online chat. Collentine and Collentine (2013) also draw on interactional research in SLA theory. Their study thus makes innovative use of technology and is informed by SLA theory.

The same applies to work by Gries and Wulff (Wulff and Gries 2011; Gries and Wulff 2013). They have studied the dative and genitive alternations in English by German, Dutch and Chinese speakers as well as patterns of gerundial and infinitival complementation. They adopt the theoretical framework of construction grammar to explain the patterns they identify. In Gries and Wulff (2013), they also make use of powerful statistical techniques to move beyond more simple patterns of learner over- or underuse of grammatical variants compared to native speakers.

Finally, Meunier and Littré (2013) have addressed the developmental perspective in work which uses data from LONGDALE. They examine the development of tense and aspect in L2 English writing, and then use the data from the corpus results to feed into an experimental acceptability-judgement study. In this way, the issue of relying solely on production can be circumvented as the corpus–experiment divide is bridged by using production data as a cue for experimental variables which might yield further explanatory insights into the nature of the learners’ competence.

3 Longitudinal Database of Learner English (see www.uclouvain.be/en-cecl-longdale.html, last accessed on 13 April 2015).
To conclude this section, it is obvious that LCR on grammar has recently advanced beyond the core learner corpus approach of studying written L2 production in comparison to similar native speaker production. Projects such as LONGDALE, FLLOC and SPLLOC provide more ‘traditional’ typical corpus data as well as experimentally elicited data, including production and judgement evidence which targets specific morphosyntactic features. It remains to be seen whether and how this might stimulate further deeper collaboration between corpus linguists and SLA theorists in the study of L2 grammar.

3 Representative studies

The choice of the three studies discussed below reflects an attempt to illustrate in greater detail some issues discussed briefly so far. So, the range of studies serves to exemplify the diverse theoretical backgrounds of learner corpus studies: Osborne (2008b) explores learner usage without any explicit commitment to a predefined theory; Lozano and Mendikoetxea (2010) test learner corpus production against hypotheses from generative SLA; Bartning (2000) interprets corpus findings from the perspective of Processability Theory (Pienemann 1998).

Irrespective of theoretical outlook, a strand which runs through the studies is the issue of usage and frequency, highlighting the importance of this for learner production (especially in Osborne 2008b). The studies illustrate the use of learner corpus work on grammar with respect to SLA theory (Bartning 2000; Lozano and Mendikoetxea 2010). They also demonstrate the computational tools which have been used in learner corpus work on grammar, simultaneously showing that the lack of syntactically parsed learner corpora is not a bar to work on clause-level grammatical phenomena (cf. Gilquin 2002). Osborne (2008b) and Lozano and Mendikoetxea (2010) demonstrate how clause-level syntactic phenomena can be captured by using tagged corpora or by taking a lexical approach, both of which, however, require more labour-intensive methodologies to identify and clean the data. In addition, the grammatical phenomena in the studies below, i.e. word order and agreement, have been of particular interest in the wider field of SLA. This provides an interesting point of comparison to other studies of L2 grammar which have used different production or experimental methodologies.

Obviously, a restricted number of studies cannot do justice to the full range of approaches to LCR and types of learner corpora that have been used in the field. While attempting to illustrate the variety of LCR on grammar, the studies below also represent some major trends in the field. Thus,
English is the target language in two of the three studies. Although a range of other target language learner corpora is available, it must be recognised that, given the position of English in the educational landscape, the majority of learner corpus studies have examined English as a Foreign Language. In addition, both Osborne (2008b) and Lozano and Mendikoetxea (2010) use ICLE. Again, this reflects the reality of the field, where ICLE has been a major resource that is very widely used. CIA has also been the methodology of choice in a large number of studies, which tend to compare the production of different L1 learner groups with the production of native speakers, and this also applies to Osborne (2008b) and Lozano and Mendikoetxea (2010). All of the studies deal with advanced learners, which again reflects a strong general tendency in the field.


The inspiration for Osborne’s (2008b) study comes from findings in generative SLA, where word order and adverb placement have been a central concern in discussions of parametric transfer. While Osborne does not explicitly couch the corpus results in terms of any one theoretical school, he clearly sympathises with a non-generative model which assumes a syntax–lexicon continuum rather than a clear syntax–lexicon distinction. He thus presents a detailed account of the patterning of adverb placement in L1 French–L2 English production, comparing this with the production of learners of other L1 backgrounds to point out that adverb (mis)placement in L2 writing seems to be constrained by lexical factors as well as the purely syntactic factors that have been examined in the generative tradition.

The syntactic analysis and SLA studies of adverb placement relate to the verb movement parameter. In movement languages like French, lexical verbs move to the left of sentential adverbs as in (1a) vs (2a). In non-raising languages like English, lexical verbs occur to the right of adverbs, so that adverbs may not normally surface between a verb and its direct object.

(1a) Je mange souvent des pommes ~ (1b) *I eat often apples.
(2a) *Je souvent mange des pommes ~ (2b) I often eat apples.

Speakers of French who are learning L2 English tend to allow verb movement in English and so produce non-target adverb placement. Such patterns are variable and touch on issues of optionality in L2 grammars. While features such as adverb placement are variable in English (and many other languages), second language learners will tend to permit a greater degree of optionality in their grammatical system. Non-target
options such as (1b) may continue to exist in the L2 grammar, even after the target distribution has apparently been acquired (see Sorace 2005 on L2 optionality).

Osborne’s study is exploratory in that it does not set up particular hypotheses to test but rather seeks to establish the way that adverb placement patterns in learner production (Osborne 2008b: 129–30). CIA is adopted as the methodology to establish patterns of L1 influence using ICLE (2.8 million words), the Chambéry Corpus (1.1 million), Essay Bank (165,000) and LOCNESS (263,000). The focus of attention is the influence of L1 French, given that French–English has been the most studied language pair with respect to adverb placement.

In order to identify the relevant grammatical structures, the corpora were tagged using Brill POS tagging. Batch search methods were used in MonoConc (Barlow 1999) to extract occurrences of adverbs in sentence-initial, sentence-final and preverbal positions, as well as verb-adverb-object order (VAO). Given that the corpora are not parsed, this required several search routines as the first element in an NP object may be a noun, adjective, determiner, quantifier, etc. The resulting search results also required manual cleaning of the data to remove occurrences which were not relevant to VAO sequences, for example, where the adverb modified an element in the following NP as in speak exactly the same language.

Having identified the structures of interest, chi-square statistical measures were used to test for significant differences in the frequency of occurrence of VAO versus preverbal placement in the L1 French production compared to the native corpora. There was, however, only a restricted usage of statistical measures, as results are not reported for the other L1 groups, nor for specific usages of different VAO patterns. VAO occurred at an average rate of 6.1 per 100,000 words in the native corpora. In the L1 French learner corpora, VAO occurred at a rate of 25.4 in Chambéry, and 22.5 in ICLE-FR. The differences in frequency between the two NNS (non-native speaker) corpora are not significant, but each differs significantly compared to the native-speaker (NS) corpora, possibly indicating a reliable effect of L1 French on the use of such patterns. The French corpora fall into a wider pattern for the different L1s in ICLE; it is possible to distinguish three groups (Osborne 2008b: 134):

1. where VAO is markedly more frequent than NS corpora (French, Italian, Spanish)
2. where VAO is less frequent than NS (Dutch, German, Swedish)
3. where VAO is comparable, or just slightly higher than NS (Finnish, Bulgarian, Russian, Czech, Polish).

While the higher frequency in the production of speakers of verb-raising Romance L1s might seem to indicate syntactic transfer, the occurrence
of VAO in the production of speakers of non-raising L1s suggests other factors need to be considered. In addition, speakers of verb-second L1s (Dutch, German, Swedish) produce VAO even less frequently than native speakers even though it is a possible structure in their L1s. This leads Osborne to explore the conditions which trigger VAO as an infrequent option in both native speaker and learner production.

As well as NP shift and verb–adverb collocations, which influence VAO production in native English, Osborne points out that VAO in learner writing might be the result of lexical deficiencies and adverb scope. It seems that heavy-NP shift, whereby an object NP may be moved to clause-final position after adverbs, also conditions VAO in the learner corpora, but with a lower threshold for the triggering conditions. Osborne calculated the weight of object NPs in VAO structures and classified NPs of 1–3 words as ‘light’, 4–6 words as ‘medium’ and 7 or more words as ‘heavy’ NPs. In the native corpora, there was a clear tendency for objects in VAO structures to be heavy (48% of objects in VAO). However, in the learner corpora, either there were roughly equal proportions of heavy vs light NPs, or there was a tendency for NPs in VAO to be light. So, even where VAO was produced less frequently by learners, as was the case for German, the qualitative pattern may still be divergent. For example, only 22 per cent of VAO produced by the German learners involve heavy NPs. Osborne (2008b: 136–7) observes that ‘even when, quantitatively, learners do not produce more V-Adv-O sequences than natives, those that they do produce are likely to appear strange because they have non target-like characteristics’.

A strong semantic link between verbs and adverbs may also facilitate VAO in English; for example, to take seriously functions almost like a multi-word verb, where objects may then occur directly after the adverb. In the learner corpora, Osborne finds that 20 per cent of all VAO sequences have verbs of perception or cognition with an ‘enhancing’ adverb, e.g. know well. Such lexical issues may in turn be linked to more general deficiencies in that verb–adverb sequences may be used by learners as semi-lexicalised units to compensate for lexical deficiencies. So, learners tend to use verbs which are semantically unspecific and complete their meaning with an adverb. They thus avoid more specific and less frequent single verbs, e.g. affect positively rather than improve or ameliorate.

Finally, adverb scope plays a role, and again, this seems to some extent to be lexically specific and concerns also, which is particularly problematic for the learners. While V-also-NP occurs in the native corpora, this is motivated by narrow scope of the adverb over the following NP. In the learner corpora, the structure occurs more frequently and is qualitatively divergent in that also appears to have wide scope over the whole clause rather than narrow scope of the following object, see (3) and (4).
Overall, Osborne’s (2008b) study illustrates how a detailed consideration of the lexico-grammatical features which regulate the actual production of a specific grammatical phenomenon may lead to a reconsideration of the nature of the syntax-lexicon interface. In contrast to syntactic approaches, he proposes that the conditions which regulate adverb placement are not ‘exclusively syntactic’, but adverb placement is ‘also influenced by semantic and phraseological associations among components of the VP’ (Osborne 2008b: 143). On the methodological side, it is also pointed out that ‘learner corpora are invaluable for tracking apparently unsystematic aspects of L2 usage such as these’ (Osborne 2008b: 143). However, one must also sound a note of caution. Osborne cites Lamb (2000: 92), who states that ‘some patterns … can be explained by the ingenuity of the linguist in finding patterns even in chaos’. One must therefore be careful to establish robust measures to identify systematic differences in learner vs native production by using statistical measures of the distribution of grammatical phenomena. The tendencies identified by Osborne would need stronger support, possibly from experimental studies which could systematically manipulate the lexico-grammatical features which seem to regulate production in writing.

In common with Lozano and Mendikoetxea’s (2010) work, Osborne’s (2008b) study serves to illustrate several general issues with research on L2 grammar using ICLE. Firstly, ICLE is a corpus of L2 writing and, as Myles (2005) points out (see above), spoken production probably provides a better reflection of grammatical competence. One must, therefore, be cautious to draw conclusions about the patterning of grammatical features in L2 writing rather than making strong claims about the nature of L2 grammars per se. Furthermore, ICLE is cross-sectional and provides a snapshot of learner ability at quite an advanced stage of development, rather than a developmental or longitudinal perspective. Connected to the issue of proficiency, Osborne (2008b: 130–1) highlights a further issue with any large-scale learner corpus in terms of proficiency measures. While the proficiency level is described as ‘post-intermediate’ or ‘proficient user’ on the Common European Framework, the actual range in such a large population of learners is likely from B2 to C2. This illustrates a difficulty with corpora, where proficiency is often impressionistically assessed or shows a wide range of variability (see also Chapter 2, this volume). In many instances, it would be necessary to have a more closely defined proficiency level in order to make robust claims about L2 grammatical development.

Lozano and Mendikoetxea (2010) also investigate word order in L2 English. They investigate L2 English produced by L1 Spanish speakers. As with Osborne’s study, the grammatical phenomenon under consideration has been important in wider SLA theory (see Sorace 2005). Unlike Osborne, Lozano and Mendikoetxea (2010) use learner corpora specifically to test hypotheses stemming from SLA theory. The Interface Hypothesis (Sorace and Filiaci 2006) states that advanced L2 learners are more likely to experience problems with grammatical phenomena at the ‘external’ interfaces of syntax and discourse than with ‘internal’ interfaces of syntax and lexicon or semantics.

The syntactic analysis of inversion and pro-drop is intricate and beyond the scope of what is necessary for present purposes (see Lozano and Mendikoetxea 2010 for detailed discussion). Summarising, English permits postverbal subjects as in (5) in a restricted range of interface-constrained contexts.

(5) Into the room came a stranger.

The verb in this sort of inversion structure must be an unaccusative verb of appearance or existence, the postverbal subject is focus (i.e. discourse-new and/or hearer-new), and postverbal subjects will tend to be heavier than preverbal subjects. The constraints on postverbal subjects in Spanish only partially overlap with English as inversion is possible with all verb classes rather than just unaccusatives. As with English, inversion is used to mark focalisation. Drawing on these cross-linguistic differences and findings from previous SLA research, Lozano and Mendikoetxea (2010: 482) formulate hypotheses for interface constraints on L1 Spanish–L2 English production of postverbal subjects: (1) the lexicon–syntax interface: as in native English, postverbal subjects will be produced with unaccusative verbs; (2) the syntax–discourse interface: as with native English, learners will place unaccusative subjects in postverbal position when they are focus, in preverbal position when they are topic; (3) the syntax–phonology interface: as in native English, learners’ unaccusative postverbal subjects will (typically) be heavier than their preverbal counterparts.

These hypotheses are tested in the Spanish subcorpus of ICLE and WriCLE (Written Corpus of Learner English, collected at the Universidad Autónoma de Madrid). These were combined into one corpus as WriCLE followed the same procedures as ICLE (giving a total of 264,212 words). LOCNESS was used as the native control corpus for comparison.

Once again, the methodology illustrates the difficulty of automatically identifying complex grammatical structures in corpora which are not parsed (and an example of how to work around such difficulty using concordancing tools). A total of seventy-two lemmas (31 unaccusative verbs, and 41 unergative) was concordanced using WordSmith Tools 4.0 (Scott 2004). Importantly, possible orthographic and morphological non-target forms were also concordanced. In studies where lexical search routines
have to be used to identify grammatical structures, it is important to be able to identify the grammatical phenomena, regardless of non-target performance. Concordances containing postverbal (VS) and preverbal (SV) sequences with these verbs were filtered according to fifty-one criteria (Lozano and Mendikoetxea 2010: 483 provide an example of some of the criteria used). Subjects were coded for phonological ‘weight’ (light vs heavy) by comparing numbers of words with a measure of syntactic complexity and also coded for discourse status (topic or focus). A total of fifty-eight inverted structures were identified in the learner corpus and sixteen in the native. These were compared to a subset of the uninverted SV structures.

While unmarked SV order is the norm in both learner and native production (92.9% of concordances vs 97.8%), the learners produce postverbal subjects significantly more often than do native speakers (7.1% vs 2.2%). Aside from the overuse of inversion structures, the hypotheses concerning the constraints on inversion are confirmed. In terms of verb classes, only unaccusatives occur in VS sequences. Inverted subjects are consistently focus information (98.3% in learner corpus, 100% in native). In both the learner and native corpora, inverted subjects were usually ‘heavy’ on the scale used (81% for learners and 81.3% for native speakers). The interface features which license subject–verb inversion were the same in the native and in the learner corpora.

However, despite the similar interface conditions, learners continue to produce ungrammatical structures, due to the wrong choice of preverbal element. The most frequent type of VS structure in the learner corpus, at 44.1 per cent, involves it-insertion, as in (6), which is ungrammatical in English. Possible postverbal structures become ungrammatical due to the dropping of expletive there, as in (7). Ungrammatical structures also occurred with no overt initial element, as in (8). So, altogether two-thirds of VS-structures in the learner corpora are ungrammatical as they are structurally impossible in English.

(6) In the name of religion it had occurred many important events.
(7) And from this moment begins the avarice.
(8) It is difficult that exist volunteers with such a feeling against it.

Lozano and Mendikoetxea (2010) analyse the results in terms of the Interface Hypothesis. The patterns seem to indicate that learners experience difficulty in integrating syntactic features with features external to the grammar. So while the interfaces constrain production of VS for learners in the same way as for native speakers, competition from the L1 and processing limitations may mean that learners cannot always map the interface information onto appropriate syntactic structures.

In addition to providing data relevant to theoretical models of L2 grammar, the use of learner corpora by Lozano and Mendikoetxea (2010), as in Osborne’s study, illustrates the importance of frequency and lexical
issues in learner corpus studies. While the constraints on production of inversion were similar for natives and learners, it was noted that inversion occurred significantly more frequently in the learner corpus. This seems to be caused by a lexical bias in the learners’ production. Half of the inversion structures produced by learners contained the verb *exist*, which would usually have a postverbal subject in Spanish. This feature does not transfer directly to L2 English as there are more instances of SV with *exist* than inverted structures. It is the general overuse of the lexical item ‘exist’ which serves to inflate the rate of inversion structures. It is clear, then, that at least when dealing with L2 production, one must take into account frequency effects stemming from lexical biases, even where the overall conceptual approach to grammar does not necessarily admit a grammar-theoretical role for such features.


Bartning’s (2000) study differs from the previous two in a number of ways. In this case, French is the target language (L1 Swedish), there is a developmental aspect from learners at different proficiency levels, and the evidence comes from spoken corpora, i.e. the InterFra corpus. However, there is a certain empirical price to be paid for using longitudinal, spoken learner corpora. Obviously the collection of such corpora is more time consuming and technologically demanding than the collection of written corpora as it requires recording and transcription of spoken production as well as tracking learners over a longer time period to give a longitudinal perspective. So, the price for using spoken, longitudinal corpora is a smaller sample size. Bartning draws on longitudinal data from interviews carried out over four terms with six advanced university-level learners. Her pre-advanced data comes from interviews and narrations by nine secondary school-level learners (compare with ICLE, which contains written data from more than 3,000 learners).

As with Lozano and Mendikoetxea, Bartning uses her data to test SLA theory in the form of Processability Theory (Pienemann 1998). This states that L2 grammatical development is constrained by the processability of grammatical structures. A learner progresses through hierarchical developmental stages which reflect the types of grammatical structures which may be processed. For example, the starting point is the word, so grammatical relations such as agreement will not be marked at Stage 1. The important stages for Bartning’s study are Stages 3 and 4. At these stages, phrasal information and inter-phrasal information may be processed, respectively. Thus, one would predict that phrasal morphology...
Learner corpora and grammar

as required by adjective agreement in *la maison verte* (the.FEM green.FEM house) will be acquired before inter-phrasal morphology as in *la maison est verte* (the.FEM house is green.FEM). To summarise, in L2 French, agreement will be marked where adjectives are used attributively within the NP before agreement is marked on predicative usages at the clause level.

French marks gender agreement on determiners (masculine *le/un*, feminine *la/une*). Adjectives also display gender agreement by adding a consonant to masculine forms with vocalic variation (e.g. *petit/petite*), with simple vocalic variation (e.g. *bon/bonne*), with complex variation (e.g. *vieux/vieille*), or by changing the final consonant (e.g. *neuf/neuve*). These forms are examined in the learner production against the predictions that such gender agreement marking should appear in attributive contexts (*la maison verte*) before predicative contexts (*la maison est verte*). The processability predictions are extended to include usage of such grammatical features and the extent to which they are automatized in production at more advanced levels of proficiency.

The data was made up of 1,352 target and non-target forms of adjectives extracted from the *InterFra* corpus, which were submitted to statistical testing using chi-square to test for differences between two population proportions. Interestingly, there is no information about how the grammatical structures of interest were identified and extracted from the corpora. There is no discussion of specific corpus linguistics tools, concordancers, frequency lists, etc. The study therefore falls into a similar category to general SLA production studies, where results are presented in terms of their relevance to SLA rather than as a learner corpus study per se.

It is found that pre-advanced learners have more difficulty than advanced learners with marking gender agreement on determiners. The pre-advanced learners have equal difficulty with different types of determiners; the accuracy rate of gender marking is 74% on definite determiners and 72% on indefinite determiners. For the advanced learners, there is, however, a significant difference in accuracy rates between definites (93%) and indefinites (83%). The advanced learners also overgeneralise the masculine definite form *le* to feminine contexts. Using different proficiency levels illustrates that there is development as pre-advanced learners apparently use the two genders randomly while advanced learners show more signs of improvement with definite determiners.

Learners’ production of adjectival agreement more directly addresses the predictions of Processability Theory. Data from the advanced group contradicts the predictions as frequency of accurate agreement on anteposition\(^6\) attributive adjectives (74%) is significantly lower than in

\(^6\) Attributive usage of adjectives is further subdivided in French between anteposition, where the adjective occurs before the noun it modifies (*un petit chien*, ‘a little dog’) and postposition, where the adjective occurs after the noun it modifies (*le journal suédois*, ‘the Swedish newspaper’).
predicative position (84%). However, the issue seems to be specific to feminine forms as these are replaced with non-target masculine forms in all contexts. So it seems that the syntactic position of the adjective matters less than the retrieval and application of L2 morphophonological forms. The pre-advanced group’s performance is more in line with the predictions as they have more problems with agreement in predicative position. However, Bartning cautions that the pre-advanced group do not produce many adjectives at all, which makes it difficult to establish clear, significant patterns of usage either within the pre-advanced group or between the groups.

Overall, Bartning’s study concentrates on the grammatical patterns relevant to L2 development and does not emphasise the corpus side of the study. This illustrates the usage of learner corpora as a methodology in SLA research, rather than learner corpora taking centre stage as a separate object of study. It highlights the use of spoken corpus evidence rather than written production, and so addresses issues with how representative of grammatical competence written production can be. The trade-off, of course, is a smaller sample size as the collection of the corpus is more labour intensive. In addition, using spoken corpora cannot be seen as a simple panacea to address issues with grammatical competence. In the case of Bartning’s study, the grammatical accuracy of the form hinges crucially on pronunciation. As an example, the distinction between the feminine adjective *neuve* and the masculine *neuf* relies on a voicing distinction on the final segment. One could imagine scenarios where learners’ phonetic and phonological systems may prevent them from making such distinctions in production even though they have acquired the relevant grammatical properties. Spoken corpora may, therefore, pose a different range of problems in the study of L2 grammar and care is obviously necessary in the design and annotation of spoken corpora (see Chapter 6, this volume).

4 **Critical assessment and future directions**

As pointed out at the beginning, Granger (2008b) has observed that while learner corpus research is at the crossroads of corpus linguistics, SLA and FLT, its most secure foothold has traditionally been in corpus linguistics. This is particularly evident in the case of LCR on grammar, in which the approach to grammar from the wider corpus linguistics field has also tended to inform LCR. Thus, LCR, as with corpus linguistics more generally, often goes hand in hand with a more usage-based, constructionist approach to grammar in implied (or overt) opposition to a generative tradition, which assumes distinctions between syntax/lexicon, competence/performance and acquisition/learning. An aim of this chapter has, however, been to show that learner corpora can be
used productively to contribute to a range of theoretical positions on L2 grammar and are not necessarily inextricably linked to specific theories. It would be beneficial, surely, for LCR to continue to contribute empirically to a variety of frameworks rather than becoming narrowly identified with just one conceptualisation of grammar. Learner corpora could then consolidate their place as a flexible methodological tool available for any researcher interested in questions of L2 grammar acquisition/learning. While this has, of course, always been the case, conceptual differences between corpus linguists and theoretical syntacticians have perhaps been reflected in the fact that those interested in theories of L2 syntax have not immediately turned to learner corpora as a resource. As Granger (2004) noted, there is a long history of L2 grammar research drawing on production evidence. The recent growth in availability of learner corpora and corpus analysis tools has not been reflected in their use in SLA research on grammar.

There are perhaps signs that this is changing. The FLLOC and SPLLOC corpora have been used extensively to address questions of development of L2 grammar. Interestingly, these have in common with LONGDALE that they include experimental evidence in addition to purely corpus production resources. Indeed, LONGDALE is called a database rather than a corpus precisely because it includes a wider range of different data types in addition to production. As such resources with spoken production and additional experimental evidence become more readily available, their value for L2 grammar research will perhaps become more obvious to SLA specialists than has thus far been the case.

Looking forward, it is likely that technological advances will make corpus collection, annotation and analysis easier, allowing a wider range of researchers in addition to dedicated corpus linguists to engage with corpus techniques. Collentine and Collentine (2013) provide an excellent example of the innovative use of technology and corpus-linguistic techniques to build and analyse a corpus of learner online interaction. Note that this was not a corpus linguistics study, but rather an investigation of interactional alignment within an interactionist approach to SLA, which just happened to use corpus-linguistic techniques. This perhaps points the way for other scholars of L2 grammar by showing how corpus tools can be used to address research questions which emerge from theoretical approaches to L2 grammar learning.

In addition to communication tools such as online chat, which lend themselves readily to corpus collection, increasingly powerful statistical tools are being brought to bear on questions of L2 grammar. Gries and Wulff (2013) highlight how more sophisticated statistical measures applied to constraints on the genitive alternation in L2 English can provide new insights into the distribution of variant grammatical forms. Such insights would not become evident from more traditional learner corpus measures of over- and underuse (Gries and Wulff 2013: 232). Future more
robust statistical studies offer the potential to gain further fresh insights into patterns of usage of L2 grammar (see also Chapter 8, this volume). Beyond theoretical and methodological issues with respect to SLA, learner corpus research on grammar has the potential to inform language pedagogy. As has been outlined, a pedagogical perspective is often a central concern in learner corpus research (see Chapters 20 and 22, this volume). This is usually conspicuously absent from theoretical SLA work on grammar. At the interface of SLA and L2 pedagogy, learner corpus research thus has the potential to inform new methods or tools which link insights from the description and analysis of L2 grammar to teaching materials or activities. Learner corpus research on grammar has found indirect pedagogical applications, in reference materials design, for instance. Rundell and Granger (2007) outline how insights from error analysis in ICLE have been used to design additions to the Macmillan English Dictionary for Advanced Learners (2007) to promote accuracy and fluency. Grammar presents more practical problems than lexical issues for the direct pedagogical use of corpora in the language classroom. The identification of more complex grammatical properties requires annotation, which demands more expertise on the part of teachers and more training for learners to be able to use such corpora effectively. This is no bar per se to the direct application of corpora for the teaching of grammatical properties. It may, however, require greater collaboration between teachers and researchers in future in order to implement it effectively.

Given Granger’s characterisation of LCR as being at the interface between corpus linguistics, SLA and FLT, it has the potential to act as a bridge between theory and practice. Despite potential theoretical problems of over-relying on production evidence to study grammar, production is, after all, what is tested and assessed in language teaching. As speech and language processing continues to advance and possibly facilitate the analysis of grammar in corpora, LCR on grammar is well placed to play a key role at the interface between theoretical understanding of the development of L2 grammar and applications in the teaching and assessment of foreign languages.

**Key readings**


Biber and Reppen compare the usage of complement clauses in the Longman Grammar Corpus and Longman Learners’ Corpus against the background of information presented in reference and pedagogical materials. They note that aspects of usage such as the frequency of
different structural variants, the frequency of matrix verbs and register distribution of different forms are disregarded in reference materials. Across a range of factors such as the frequency of occurrence of different matrix verbs and frequency of that-omission, the learners’ written production more closely mirrors patterns in native speech. This illustrates the importance of genre and register variation as a factor in addition to formal grammatical properties.


Gries and Wulff study the genitive alternation in L2 English (i.e. squirrel’s nest vs the nest of the squirrel). They use the German and Chinese components of ICLE. Rather than relying on over-/underuse of the variants in learner vs native production, they code a number of possible grammatical, phonological and processing predictors for the use of the variants. These were submitted to a multifactorial analysis. Results indicate that processing factors constrain native and non-native production. The use of multifactorial analysis captures complex networks of interactions which would not be evident from a less sophisticated statistical analysis.


Meunier and Littré examine the development of the tense and aspect system in L1 French–L2 English. Longitudinal data comes from LONGDALE. Tense and aspect errors were tagged and analysed in a mixed-effects model taking number of errors at each point of time as a variable to identify developmental patterns. The most persistent error was replacement of present simple with the progressive. The results were used to develop an acceptability-judgement task to test why this error persists. This differentiated between learners’ ability to use the progressive for ongoing events and a less precise understanding of present progressive for planned events.


Oshita is a generative study of passive unaccusative errors of the type What was happened? in L2 English. Ten verbs which have been observed to occur frequently in such errors were used to extract
clauses with potential errors from the Italian, Spanish, Japanese and Korean components of the *Longman Learners’ Corpus*. Token clauses were classified for (non-)target syntactic patterns. The prevalence of the *Three boys were arrived* pattern in non-target structures led Oshita to suggest that this error type is an overgeneralisation of the passive morphosyntax of English. It is pointed out that naturally produced interlanguage data is an important complementary source for theory construction in addition to judgement evidence.


Thewissen exploits an error-tagging system to track the development of different error types at different proficiency levels. She addresses the issue of proficiency by using a subset of essays from the French, German and Spanish components of *ICLE* which were independently rated and categorised according to Common European Framework descriptors B1–C2. These subcorpora are then used in a quasi-longitudinal study to establish developmental patterns of errors across proficiency levels. Measures of the number of errors as a function of how often a potential error could have occurred identifies strong, weak and non-progressive developmental patterns.