1 Corpus-based Research in Second Language Spanish[1]

AMAYA MENDIKOETXEA

1.1 Introduction

Second Language Acquisition (SLA) is a diverse field, both conceptually and empirically. Conceptually, it draws from several disciplines (linguistics, psychology, sociology, etc.) and encompasses a variety of theoretical frameworks. It relies on data types drawn from different data elicitation techniques and a variety of methodological approaches. From a cognitive perspective, the main objective of SLA research is to build models of the underlying systems of knowledge that learners have at a particular point in the SLA process (their interlanguage) and to provide a principled account of how that knowledge is acquired and how it develops. As Myles (2005, 372) points out, ''the language produced by learners, whether spontaneously or through various elicitation procedures, remains a central source of evidence for these mental processes, and the success of SLA research therefore relies on having access to good-quality data.'' Learner language is primary data for the study of SLA and learner corpora (a special type of corpora containing second language (L2) learners' written or oral language samples, see Section 1.2) should occupy a central role in SLA research.

Methodologically, L2 researchers have traditionally, but not exclusively, relied on (quasi)experimental and introspective data (see overviews in, e.g., Gass and Mackey 2007; Mackey and Gass 2005; Mitchell and Myles 2004; White 2003). While the use of large-scale corpora has become standard practice in first language (L1) acquisition research, large L2 corpora are still scarce and relatively little use has been made of corpora in L2 research. In this paper, I discuss the use of learner corpora in the study of L2 Spanish acquisition. It is not my intention to provide a comprehensive survey of the corpora available and related work but to describe the most relevant projects as examples of what learner corpora can contribute to the field of SLA. In Section 1.2, I define learner corpora and learner corpus research. In Section 1.3, two currently available L2 Spanish learner corpora are described: a spoken corpus, SPLLOC *(Spanish Learner Language Oral Corpus)*, and a written corpus, CEDEL2 *(Corpus Escrito del Español como L2)*, as well as the research

The Handbook of Spanish Second Language Acquisition, First Edition. Edited by Kimberly L. Geeslin.

carried out with them. A brief overview of corpus-based research in L2 Spanish is also provided. In Section 1.4, I point out the way forward for corpus-based SLA research.

1.2 Learner Corpora and SLA

1.2.1 What learner corpora are and why we need them

Based on Sinclair's (1996) definition of language corpora, Granger (2002) defines learner corpora as:

> Electronic collections of authentic F(oreign) L(anguage)/S(econd) L(language) textual data according to explicit design criteria for a particular SLA/FLT(eaching) purpose. They are encoded in a standardised and homogeneous way and are documented as to their origin of provenance. (Granger 2002, 7)

The compilation and exploitation of a learner corpus requires a wider range of expertise than is required for native language corpora (Granger 2009, 15). On the one hand, researchers need to be familiar with the methodology of *corpus linguistics*[2]: corpus design, corpus annotation, automated data extraction and analysis, and so on. This has the additional complication that most available tools have been designed for native corpora and are therefore not fully suitable for learner corpora; for example, Part-of-Speech (POS) tagging, which, as Granger (2009, 15) points out, is affected by the high rate of errors in learner language. On the other hand, a good background of *linguistic theory*, as well as SLA theory, is necessary for analyzing and interpreting the data. These two types of expertise are not often found together: ''many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them [SLA researchers] effectively, while SLA researchers typically know little about what corpora can do for them'' (Tono 2003, 806).

One of the main contributions of learner corpora is that they provide a much wider empirical base than has previously been available. SLA studies are often conducted on the basis of a very limited number of subjects, which raises questions about whether results can be generalized (Granger 2002, 6). These studies have served the purpose of hypothesis-building in SLA research, but there is an increasing awareness of the need to test hypotheses on larger and better constructed databases (see Myles 2005, 2007a, 2007b). Moreover, corpora are often used in an exploratory fashion: to discover sets of data not normally found in small studies, which can become crucial to inform current debates in SLA, and to discover patterns of use, as well as for quantitative studies (e.g., frequency). The latter is especially useful for usage-based approaches and input-driven models of SLA (see Gries 2008), but corpora may also be used to inform current debates on the role of input in more formal approaches to SLA.[3] Finally, the use of corpora, by the very nature of the data (contextualized discourse), enables researchers to tackle some previously neglected aspects of SLA, such as lexis, phraseology, information structure, and so on, instead of just morphology and syntax, which are the traditional focus of SLA research (Granger 2009, 17).

Some caveats are necessary regarding two apparent dichotomies emerging from the definition of learner corpora and the discussion above: authentic vs. non-authentic (elicited) data and corpus vs. experiments in SLA. It is actually difficult to define what constitutes ''authentic'' data when dealing with learners' production. Granger (2002, 8) defines authentic learner data in instructional settings as data resulting from authentic classroom activity: texts that are produced for pedagogical reasons and for the corpus, but that use procedures exerting very little control. Compositions guided by pictures and typical experimental data resulting from elicitation techniques are not, according to this author, ''authentic'' language samples. For Nesselhauf (2004, 128): ''Since the distinction between more or less controlled is, naturally, not clear-cut, such collections might be considered peripheral parts of learner corpora''. Sinclair (1996) claims that data collected through major intervention by the linguist form ''experimental'' corpora. It turns out that learner corpora are most often either semi-authentic or experimental, rather than fully natural, but still a highly valuable source of learner language. The position in the scale of naturalness depends on the degree of control researchers wish to exert over the data and this is, in turn, dependent on their research questions. Given that corpora contain data with varying degrees of naturalness, there is in fact no strict corpora-experiments dichotomy (see Gilquin and Gries 2009, 6). Additionally, a growing number of researchers are arguing for combining data extracted from different sources (see Sections 1.3.2 and 1.3.3). However, this has to be done in a systematic way in order to obtain reliable conclusions on converging evidence.[4] As there is no direct access to learners' interlanguage, we need to triangulate from all available sources of data: investigating different types of behavior may help us narrow down the range of possibilities. Combining naturalistic and experimental data is crucial for this purpose.

1.2.2 Learner corpus research in SLA

Learner corpus research is L2 research which uses learner corpora as the main source of data. Interest in learner corpus research was sparked mostly by the publication of the first version of ICLE (*International Corpus of Learner English*, Granger, Dagneaux, and Meunier 2002), the starting point in the exploitation of large-scale learner corpora. ICLE consists of 2.5 million words of argumentative essays by L2 English university students, organized in different subcorpora according to the learners' L1: Spanish, Italian, French, Russian, etc.[5] Most of the studies done with ICLE have analyzed lexical aspects of learner language, probably due to limitations in concordancers and query software. Certainly, some researchers have gone *beyond the word* by analyzing phrases and structures (Fitzpatrick 2007), collocations (Nesselhauf 2005), and word order alternations (Lozano and Mendikoetxea 2008, 2010).[6] Research within the ICLE tradition is inherently contrastive. *Contrastive Interlanguage Analysis* (CIA) (see, e.g., Granger 1996; Gilquin 2001) is the term used for a research paradigm which establishes comparisons between (i) two (or more) interlanguage varieties (e.g., L1 Spanish – L2 English vs. L1 Italian – L2 English), and (ii) L1 and L2 grammars, by comparing native and non-native corpora.[7]

As a whole, studies using learner corpora in SLA fall within two categories: (i) *hypothesis-driven/corpus-based* studies and (ii) *hypothesis-finding/corpus-driven* studies (see Barlow 2005; Granger 1998; Tognini-Bonelli 2001). This reflects the tension between *deductive* vs. *inductive* approaches in language acquisition research (see Myles 2007b for an overview and discussion), with most studies falling within category (ii) (e.g., Aston, Bernardini, and Stewart 2004; Granger, Gilquin, and Meunier, forthcoming; Granger, Hung, and Petch-Tyson 2002).[8] However, learner corpus research is a relatively young but very active field. As pointed out by Díaz Negrillo and Thompson (forthcoming), in the last decade we have seen an increasing number of resources, a broadening of the uses that learner corpora are put to, and a wider diversity of users. While research within this field is also increasingly important, the contribution of learner corpus research has been much more substantial in description than interpretation of SLA data (Granger 2004, 134–135), with very little reference to current debates, hypotheses, and theories of SLA (Myles 2005).[9, 10]

1.2.3 Corpus design: deciding on criteria for learner corpora

Most available learner corpora are of L2 English. Most are written and only a few of them are spoken. Where different proficiency levels are represented in a corpus, most are cross-sectional, offering a cross-section of the learner population (containing texts from groups of learners at different proficiency levels collected at the same point in time), and very few are longitudinal, following learners' development over a period of time (containing texts from the same learners at different stages of acquisition).[11]

For a corpus to be useful for linguistic analysis, it is crucial to have strict design criteria that follow standard practices (see, for instance, Wynne 2005). In fact, most learner corpora are opportunistic; researchers collect data which are readily available and do not require vast investments of resources and time. Some are designed following *ad hoc* methodology, that is, to elicit particular types of structures or lexical items (against what is considered good practice in corpus design, according to Sinclair 2005). The collection of texts gathered in a corpus has to be (i) representative, with a high degree of inclusiveness and a low degree of language bias, so that the corpus could potentially contain all likely morphosyntactic forms and a variety of language structures and vocabulary items (see Sinclair 2005; Gries 2008), and (ii) balanced, containing a fair and equally proportioned sample of each of the language varieties it is supposed to be representative of (e.g., a roughly equivalent number of words for each of the proficiency levels in cross-sectional and longitudinal corpora).

In addition, to be useful for SLA research, learner corpora should incorporate a reliable measure of learners' proficiency (see Tono 2003) to allow for contrastive analyses of learners' interlanguage at different proficiency levels, as well as for developmental research. This, together with other types of background information (e.g., L1, length of exposure, learning environment, etc.), is essential to conduct L2 research concerning interlanguage grammars, as well as, for instance, critical period effects, language use patterns, likely cross-linguistic effects, stay abroad effects, and so on. Finally, decisions have to be made about annotation, which is often manual and semiautomatic. Though standard annotation would be desirable, researchers tend to adopt their own annotation schemes to suit their research purposes.[12] Rutherford and Thomas (2001) argue in favor of reexamining the procedures and tools of the CHILDES project, originally conceived for L1 acquisition, to explore their potential for learner corpus analysis: the set of transcription conventions in CHAT (Codes for the Human Analysis of Transcripts) and the CLAN (Computerized Language ANalysis) suite for POS tagging.[13] Using the same annotation scheme facilitates sharing and comparing results.

1.3 Spanish Learner Corpora

This section offers a brief overview of L2 Spanish learner corpora, focusing on two cross-sectional corpora: CEDEL2 and SPLLOC. These are (relatively) large corpora which represent different types of learner corpora (written vs. spoken). They have been developed according to strict criteria that make them suitable for SLA research and are freely available for the research community.[14]

1.3.1 A brief overview of Spanish L2 corpora

We are just beginning to see the emergence of research projects based around the compilation and exploitation of L2 Spanish corpora. Most are pedagogically oriented for Spanish FLT (see Campillos Llanos 2012, Section 1.3.1 for an overview). Pedagogically oriented corpora are often based on error analysis, for example, CORANE *(Corpus para el Análisis de errores de aprendices de E/LE)* (Cestero Mancera *et al*. 2001), a cross-sectional corpus containing over a thousand compositions by learners from different L1 backgrounds, recently published in CD-ROM format (Cestero y Penadés 2009). A similar corpus is CATE (*Corpus de Aprendices Taiwaneses de Español*, Lu 2012),[15] developed as part of a multilingual corpora project, with almost 2,000 compositions and over 300,000 words. The largest written corpus, to this author's knowledge, which has been designed with clear SLA research purposes, is CEDEL2 (*Corpus Escrito del Español como L2*, Lozano 2009a; Lozano and Mendikoetxea, forthcoming) (see Section 1.3.3).[16]

Corpus-based research into L2 Spanish oral production often involves case studies or small numbers of learners, given the difficulties of obtaining and transcribing the data. Campillos Llanos (2012, 23–25) offers a detailed account of studies based on error analysis of L2 Spanish oral interlanguage. This author has collected an oral corpus containing interviews with 40 learners from a variety of L1 backgrounds and different proficiency levels, which has been transcribed using the CHAT format and has been fully tagged for phonetic, lexical, grammatical, and pragmatic errors.[17] The only multimodal corpus, to our knowledge, belongs to the Spanish *Proficiency Level Training* (SPT) project at the University of Texas, which has been developed to train teachers to assess proficiency levels on the basis of oral speech production (Koike 2007).[18]

A number of studies based on oral production have been conducted to investigate aspects such as null subjects, inflection, word order, and so on which are reflections of parametric differences between Spanish and other languages, notably English. The project *Beyond Parameters* (University of Ottawa) involved the collection of interviews for longitudinal/cross-sectional studies (among others, Liceras *et al*. 1997; Liceras and Díaz 1998, 1999; Liceras, Valenzuela, and Díaz 1999). On the basis of this project, Díaz (2007) has compiled a longitudinal corpus collecting interviews taken at one- or two-month intervals involving learners with L1 German, Swedish, Icelandic, Chinese, and Korean. The largest freely available Spanish learner oral corpus which has been designed specifically with SLA research purposes in mind is SPLLOC (*Spanish Learner Language Oral Corpus*, Mitchell *et al*. 2008), following the same design criteria as its predecessor FLLOC *(French Learner Language Oral Corpus)*.[19] In what follows we offer more details about this corpus (Section 1.3.2), as well as CEDEL2 (Section 1.3.3). We then look at other corpus studies of L2 Spanish which focus on the form-function relationship (Section 1.3.4).[20]

1.3.2 SPLLOC

1.3.2.1 The SPLLOC program The creation of a database of spoken L2 Spanish in the SPLLOC program constitutes a landmark in L2 Spanish acquisition research and SLA as a whole. Data has been collected from L1 English L2 – Spanish learners in an instructional setting at a variety of levels (from beginners to advanced), as well as from native speakers. Both the sound files and their transcriptions (in CHILDES format) are available through the SPLLOC website (see note 14). SPLLOC designers, like their predecessors in FLLOC, focus on spoken, rather than written, data, under the hypothesis that spontaneous speech produced in face-to-face interaction is likely to provide more direct evidence about the L2 learner's interlanguage, thus minimizing the effects of self-correction and monitoring. Given the variability of learners' oral production data and the tendency to avoid structures in speech, the design of SPLLOC involves a variety of genres (narrative, interview and picture description, peer discussion). There is also a substantial speech sample from each individual participant (40–60 minutes), with varying interlocutors, and a balance of open-ended and focused elicitation tasks (see Mitchell *et al*. 2008 for details).

1.3.2.2 Work with SPLLOC Two independent research projects have been undertaken using this corpus. SPLLOC1 investigated the acquisition of central morphosyntactic features, such as word order (Domínguez and Arche 2008) and clitic pronouns (Arche and Domínguez 2011), from a developmental perspective and focusing on acquisition at the interfaces. The purpose of these studies was to test current hypotheses in SLA: the *Impaired Representation Hypothesis* (Hawkins and Chan 1997; Tsimpli and Roussou 1991, among others) vs. the *Missing Surface Inflection Hypothesis* (Lardiere 1998; Prévost and White 2000, among others) in the clitics study, and the *Interface Hypothesis* (Sorace 2005; Sorace and Serratrice 2009; Tsimpli *et al*. 2004, among others) in the word order study. Both studies were based on experimental data from SPLLOC to facilitate hypothesis-testing research.[21] The results in Arche and Domínguez (2011) of the clitic production and interpretation tasks show that accuracy in performance correlates with level of proficiency (beginner, intermediate, and advanced). However, while high rates of production correlate with high rates in the comprehension task for the advanced group and, conversely, low rates in production correlate with low rates in comprehension for beginners, intermediate level learners score very high in the comprehension task (higher than 80%) but show very low usage of clitics (around 20%). As for word order, the results show that the acceptability of verb-subject (VS) orders is in strict correlation with learner proficiency levels. Subject-verb inversion (an option not allowed in the learners' L1) is not selected by learners in the beginner and intermediate groups, but is correctly preferred by the advanced group. In addition, Domínguez and Arche (2008) argue that the optionality shown by advanced learners should be understood as an intermediate stage showing grammar restructuring, rather than a case of a pragmatic deficit.

Both studies mentioned used the focused elicitation tasks in SPLLOC1. Activities prompting learners' production and/or interpretation are widely used in linguistically oriented SLA. They serve the purpose of pushing learners to produce particular target structures of interest to researchers, which learners may avoid in natural production, and allow researchers to infer ''not only what learners know is correct in the second language, but also what learners know is not possible'' (Gass and Mackey 2007, 73). Including focused elicitation tasks in the corpus alongside more open-ended tasks being undertaken by the same L2 participants creates the possibility for triangulation across different data types. Further research within SPLLOC1 should then involve exploring the constructions in the focused tasks in the open-ended tasks of the corpus.

The more open-ended tasks (picture description and interviews) in SPLLOC1 have been used to undertake a comparative investigation into lexical progression amongst school learners of Spanish and French (using equivalent data from FLLOC) (Marsden and David 2008). Their analysis supports the idea that an important indicator of development in inflectionally rich languages like Spanish is increased morphological (inflectional and derivational) variation, not just accuracy (see also Collentine 2009). It also shows that as learners' linguistic competence develops, they start to produce more verbs than nouns, and as they produce more verbs, they also start to produce more adjectives. As pointed out by Asención-Delaney and Collentine (2011, 302), it is logical to think that these changes in lexical and grammatical production are parallel to changes in the discourse types that learners produce (see Section 1.4.4).

The combined approach (looking at both experimental and corpus data) is adopted in SPLLOC2, which investigates the development of the tense-aspect system in L2 Spanish in order to understand the route of acquisition of past tense forms in an instructional context. Domínguez *et al*. (2012) use three oral tasks with varying degree of experimental control to show that the emergence of temporal markings is determined mainly by the dynamic/non-dynamic contrast (event vs. state). While this is an important finding regarding the role lexical aspect plays in the acquisition of the tense-aspect system in L2 Spanish, the main contribution of this paper is in the methodology employed, and in particular, in the combination of naturally occurring corpus data with more controlled production data and an experimental comprehension task in order to obtain a fuller picture of the learners' interlanguage grammar. As mentioned above, there is a growing awareness among corpus linguists and L2 researchers that combining different types of data is essential in trying to determine the linguistic competence of L2 grammars (see, e.g., Gilquin and Gries 2009; Mendikoetxea and Lozano, forthcoming).

1.3.3 CEDEL2

1.3.3.1 CEDEL2 and the WOSLAC program CEDEL2, (Lozano 2009a; Lozano and Mendikoetxea, forthcoming) is a written L1 English – L2 Spanish corpus sampling learners of all proficiency levels. It originated within the WOSLAC (*Word Order in Second Language Acquisition Corpora*) research group at the Universidad Autónoma de Madrid, which had a double objective: (i) to explore the role of the *Interface Hypothesis* in L2 grammars using corpus data (Lozano and Mendikoetxea 2008, 2010), and (ii) to compile two comparable learner corpora, which are suitable for L2 research. Thus, together with CEDEL2, *Wri*CLE *(Written Corpus of Learner English)* has been created: an L1 Spanish – L2 English corpus of approximately the same size as CEDEL2 and compiled according to the same principles (see Rollinson and Mendikoetxea 2010), under the assumption that well-researched parametric contrasts between Spanish and English, as observed in corpus data, are crucial to inform the current debate on deficits at the interfaces.

The fact that CEDEL2 and *Wri*CLE can be used to study transfer in both directions is already an important contribution, but as a new source of data CEDEL2 represents an advance in L2 Spanish research for several reasons. While a deductive approach is followed in SPLLOC (i.e., the corpus is designed to elicit specific linguistic constructions to test specific research questions, see Myles 2007b), CEDEL2 has a more exploratory, inductive approach. It crucially follows Sinclair's (2005) ten standard principles recommended for corpus. So, CEDEL2 is designed to potentially tackle any L2 research question concerning any linguistic structure and the fact that it is a written corpus allows for a larger number of words (c. 750,000 words to date, and aiming at 1 million words in the near future, coming from c. 2,400 participants). Like SPLLOC, it contains a similarly designed Spanish native speaker subcorpus serving as a control group, which allows for the reliable contrast of interlanguage data against the native norm under equally comparable conditions.[22] Unlike other L2 learner corpora that do not include a reliable measure of learners' proficiency, CEDEL2 learners (mostly students in a university context in English-speaking countries) were administered a standardized grammatical placement test.[23] This is essential to conduct reliable studies of SLA and interlanguage development, as well as contrastive analyses of learners' interlanguage at different proficiency levels.[24]

CEDEL2 learners write a brief composition for which they choose from a range of twelve titles, graded according to complexity. The fact that most available (native and learner) corpora are written is often considered to be a limitation of learner corpus research. While the validity of spoken learner production data is undoubted, this is not to say that written data cannot be used for the investigation of L2 grammars. There are likely to be fewer performance errors in the written language and the errors found are those that escape monitoring, indicating grammatical or lexical gaps in the learners' mental grammar. Learners tend to use more complex structures when they are writing, which could be more revealing in terms of their competence than the simplified language often found in oral language. Furthermore, written corpora are particularly suitable for studying the interlanguage of advanced learners, especially in comparison with similar L1 corpora. Learner corpus research in the ICLE tradition shows that advanced learner texts are a valuable source of data to study aspects such as modality, degree adverbs, tenses, collocations, phraseology, causativity, information structure, clefts, anaphora, and so on. Written corpora can also be used in hypothesis-testing studies: passivized structures and expletives (Oshita 2000, 2004), and subject inversion in L2 English (Lozano and Mendikoetxea 2008, 2010).

1.3.3.2 Work with CEDEL2 CEDEL2 samples have been used, for instance, in published research on the acquisition of pronominal subjects (Lozano 2009b; see also Lozano 2011), unaccusative predicates, and *se* (Escutia 2010, 2012), and learner collocations (Alonso *et al*. 2010a, 2010b).[25] Comparative work is currently being undertaken extracting data from both *Wri*CLE and CEDEL2 on the production of expletive subjects in L2 English and L2 Spanish (Ferrandis, in progress).

Corpus data are particularly valuable to explore the *Interface Hypothesis*, since the relevant structures are embedded within the larger discourse context. An example of this is the choice between null/overt subjects in L2 Spanish, their presence/absence in native Spanish being governed by discourse principles and information structure (namely, topic and focus). Regarding pronominal subjects, recent studies reveal that learners of L2 Spanish are sensitive to the formal syntactic mechanisms licensing overt and null pronominal subjects from early stages of acquisition, but show residual deficits when their distribution is constrained by the notions topic and focus at the syntax-discourse interface, even at advanced levels. Lozano (2009b) uses data from CEDEL2 to explore this issue and reveals that deficits are selective because they affect third-person animate features only, while the rest of the pronominal system remains stable. In a following study, Lozano (2011) explores the production of third person subjects (full NPs, pronominals, and null subjects) in topic-continuity and topic-shift contexts comparing native vs. learner production. Again, the corpus study reveals what experimental data has kept hidden: not all discursive features are equally vulnerable at the syntax-discourse interface. Learners are more sensitive to the pragmatic constraints of topic-shift than to those of topic-continuity, with a tendency to be redundant rather than ambiguous.

As for collocations, phraseology is one of the areas that have figured prominently in learner corpus research (see Ellis 2008; Granger and Paquot 2010, among others). Using data from CEDEL2, Alonso *et al*. (2010a, 2010b) focus on the technical aspects related to the processing of learner collocation errors in a corpus: (i) analysis of the corpus and derivation of a collocation error typology; (ii) definition of a tag set to annotate the corpus; and (iii) tagging the corpus, with the ultimate purpose of developing an advanced Natural Language Processing (NLP)-based computer-assisted language learning (CALL) environment for learning collocations in Spanish.

1.3.4 Exploring the form-function relation in corpus-based research[26]

Corpus-based SLA research is particularly appropriate to explore the ways in which form is mapped to function and vice versa. The use of corpora has also advanced our understanding of the acquisition of specific aspects of Spanish grammar, which require knowledge from different linguistic areas. Some of the examples of work given above using CEDEL2 and SPLLOC can illustrate this approach (e.g., the studies based on word order or the null/overt realization of pronominals). Corpus-based research has also been employed to study the acquisition of the copula verbs *ser* and *estar* (see also note 16). It is well known that copula choice is an area of great difficulty for L2 learners of Spanish. Geeslin (2000) is the first study to depart from the error-analysis approach to understanding the path of copula acquisition. She investigates the linguistic features affecting copula choice in SLA using data obtained in semi-structured interviews, a picture description task, and a contextualized questionnaire. A similar approach (see also Geeslin 2003) is followed by Cheng, Lu, and Giannakouros (2008) in their analysis of the semantic, pragmatic, and lexical characteristics of copula use in free written essays using data from the CATE corpus mentioned in Section 1.3.1. One of their main findings regards the influence of essay type on copula choice: a higher *estar* usage rate was likely in exploratory and descriptive essays relative to a narrative essay baseline. The results of their research suggest that the investigation of forms whose meaning difference appears subtle to L2 learners must benefit from the use of similar multifactorial analyses, based not only on linguistic variables, but also on text types and text length. This is the approach adopted by Collentine and Asención-Delaney (2010) in an analysis of how discourse type influences copula choice in L2 Spanish.

Gathering large corpora of digitized learner production has also proved essential to study complexity and semantic density (see Housen and Kuiken 2009, and all papers in that volume). Ortega (2000) uses a corpus of L2 Spanish (intermediate) learners to investigate reliable ways of measuring syntactic complexity. The best predictors for syntactic complexity in her analysis were clause length, amount of subordination, and phrasal elaboration (see also Ortega 2003 for other findings). Collentine (2004) makes use of corpus-based techniques in a comparison of morphological and lexical complexity between L2 Spanish learners in two different learning contexts: in-class and study abroad. The corpus comprises oral segments produced by learners in an oral proficiency interview before and after the experimental period (semester). A quantitative discourse analysis of the corpus indicated that the study abroad learners had a more complex narrative (as shown, for instance, in their use of past tenses and public verbs, e.g., *decir que* ''to say that'') and could produce language that was more semantically dense. The in-class context facilitated the development of discrete grammatical and lexical features, with learners producing a higher concentration of nouns and adjectives. More recently, Collentine (2011) has examined interlanguage complexity and accuracy in an analysis of L2 learner production resulting from the input received during a CALL (Computer Assisted Language Learning) task that involved making choices in a 3D environment. The results suggest that both the learners' choices, termed autonomous moves, and the subsequent input they receive affect their production in terms of accuracy and complexity.

Finally, Asención-Delaney and Collentine (2011) present a multidimensional analysis of a written L2 Spanish corpus examining how L2 learners of Spanish (second- and third-year university students) combine lexical and grammatical features and structures to generate different discourse types in what constitutes a first attempt at characterizing learner discourse in L2 Spanish using a type of analysis which has already been employed for native corpora (see Biber *et al*. 2006; Parodi 2005, 2007[27]). The corpus comprises written samples for course assessment purposes: letters, narratives, descriptions, summaries, and argumentative essays. A multidimensional analysis combines quantitative and qualitative research, technological tools, exploratory factor analysis, and a qualitative analysis of texts. Their analysis uncovers four significant clusters that can be considered to be distinct discourse types, characterized by two main stylistic variations: narrative (with a concentration of verbal features) and expository (with a concentration of nominal features), thus providing good insight into how L2 communication occurs in relatively extended discourse, where learners have to combine morphological, grammatical, and lexical features. Unfortunately, their corpus did not include samples of more spontaneous writing tasks, such as emails, chat transcripts, and so on, which are crucial to understand L2 discourse under stronger communicative pressure.

1.4 Corpus-based Research: The Way Forward

SLA has been slow to incorporate corpus-based techniques for interlanguage analysis (see Myles and Mitchell 2004). However, in Granger's words (2009, 28), learner corpus research is ''slowly but surely being integrated into SLA,'' a movement which is due to the recognition of the contribution of corpus-based methods to SLA research and the corresponding recognition among learner corpus researchers that their findings have to be integrated within SLA theories and hypotheses. However, if corpus-based research is going to make a significant contribution to the field of SLA, new, well-designed corpora need to be made available to the research community, representing a much wider variety of registers, tasks, learners, L2s, and so on. Such corpora should be compiled according to standard design criteria, which make them maximally useful for SLA research, and furthermore, they should be compiled by SLA researchers (or in collaboration with them), to ensure that they are not simply opportunistic and are based upon formal measurements of proficiency. In addition, corpora must be fully documented, and it should be possible to select texts from subcorpora or to filter out texts that do not meet certain criteria. As Granger (2009, 28) points out, there is a special need for longitudinal corpora if one is to approach the developmental problem of SLA (see note 20).

Significant developments in corpus analysis are also needed. Tools must be developed which are suitable for learner data and are not reliant on manual tagging. Ideally, there should be a movement toward standardized annotation systems, which are both powerful and user friendly, and a way of integrating the storing, annotation, and searching of learner corpora. Together with this, methodologies have to be developed to combine corpus data with experimental data in search of converging evidence and to test aspects, which cannot be adequately tested with corpus data (see Gilquin and Gries 2009; Granger 2012; Mendikoetxea and Lozano, forthcoming). Methodological issues in corpus-based research are also concerned with the combination of quantitative and qualitative corpus analysis in order to increase generalizability of results. Finally, there is a clear need for a closer relationship between (learner) corpus linguists and SLA researchers, with more hypothesis-testing, explanatory studies (see Granger 2004). This will only be possible if corpus design and methodologies are useful for SLA purposes. Given the interdisciplinary nature of the field, to fully exploit the potential of corpus-based SLA research there is a need for multidisciplinary teams that integrate corpus linguists, SLA specialists, computer scientists, experts on language assessment, and language pedagogy and teaching practitioners (see also Granger 2009, 28).

NOTES

1 This research has been partially funded by research grants FFI2008-01584 and FFI2011-23829 from the Spanish Ministry of Science and Innovation, which I gratefully acknowledge. I am also thankful to the WOSLAC team for discussion on many of the issues presented here, and especially to Cristóbal Lozano for the compilation of CEDEL2 *(Corpus Escrito del Español como L2)*, as part of this research program. I also wish to thank the editor of this volume and two anonymous reviewers for their insightful comments and references, which have greatly improved this paper. All remaining errors are mine.

2 Corpus linguistics involves the use of corpora as the central element for linguistic analysis. On the impact of corpus linguistics on all areas of linguistic inquiry, see Lüdeling and Kytö (2008), O'Keefe and McCarthy (2007), McEnery and Hardie (2012), among others.

3 As pointed out by Díaz Negrillo and Thompson (forthcoming), a recent development is the creation of complementary corpora of input (e.g., textbooks in an instructed learning environment). See also Meunier and Gouverneur (2009).

4 See Mendikoetxea and Lozano (forthcoming) for a proposal concerning a methodological model for the integration of corpus and experimental data.

5 An expanded version of ICLE has been recently released (Granger *et al*. 2009). *LINDSEI* (The *Louvain International Database of Spoken English Interlanguage)* (http://www.uclouvain.be/en-cecl-lindsei.html) is the oral counterpart to ICLE. Given the effort and time required in compiling and transcribing a spoken corpus, oral learner corpora are scarce and tend to be smaller than written corpora.

6 The webpage of the *Centre for English Corpus Linguistics* is an invaluable resource for learner corpus bibliography (http://www.uclouvain.be/en-cecl-lcbiblio.html).

7 For the proponents of CIA, the contrast between learner and native corpora involves a detailed analysis of linguistic features to uncover and study non-native features in the speech and writing of (advanced) non-native speakers. This includes errors, but it is conceptually wider as it seeks to identify overuse and underuse of certain linguistic features and patterns. As for the comparison of learner data from different L1 backgrounds, we can gain a better understanding of interlanguage processes and features, such as those which are the result of transfer or those which are developmental, common to learners with different L1s (see Granger 2002: 12–13 and references cited therein). Regarding the comparison between native and learner corpora, only ''comparable'' corpora can be used for that purpose. Thus, the CECL team has compiled a native English corpus of novice writers, LOCNESS (the *Louvain Corpus of Native English Essays*), containing argumentative essays written by British and American university (and A-level) students. See http://www.uclouvain.be/en-cecl-locness.html for details.

8 In addition, an important number of large L2 corpora have been created over the past few years to meet the needs of EFL materials designers (e.g., *Longman Learner Corpus* and the *Cambridge Learner Corpus*). Smaller projects have also involved the creation and exploitation of corpora for the teaching of Spanish as an L2 (see Section 1.3.1).

9 For further discussion on the use of learner corpora, see also Myles (2005, 2007a, 2007b), Thoday (2008), Rutherford and Thomas (2001), and Tono (2003).

10 Some SLA researchers have collected and analyzed relatively large amounts of naturalistic learner data. Lardière (1998), for instance, uses data from an English learner, Patty, coming

from email exchanges collected over several years. These studies allow for a detailed analysis of interlanguage development, but conclusions cannot be extrapolated to other learners.

11 On the need for corpora other than written and cross-sectional, see Barlow (2005, 349), Myles (2005, 388), and Granger (2009, 28).

12 See Díaz-Negrillo and Thompson (forthcoming) for an overview of annotation tools.

13 The largest collection of naturally occurring data is the *Child Language Data Exchange System*, CHILDES (MacWhinney 2000), which has become an international benchmark in the study of L1 acquisition and bilingualism and has also been recently employed in SLA research. It contains over 44 million words in over 30 subcorpora sampling different languages, most of which are grammatically tagged. At least 3,200 research papers have used CHILDES as their source of data.

14 SPLLOC can be downloaded from http://www.splloc.soton.ac.uk/ and Talkbank (CHILDES). As for CEDEL2, the webpage to search and download the corpus is currently under construction (more information on http://www.uam.es/proyectosinv/woslac/cedel2.htm).

15 For more information see http://corpora.flld.ncku.edu.tw/#

16 There are also smaller corpora such as *The Anglia Polytechnic University Learner Spanish Corpus*, which has been used to study the acquisition of the verbs *ser/estar/haber* (Ife 2004).

17 More information on the Spanish Learner Oral Corpus can be found at http://cartago.lllf.uam.es/corele/index.html.

18 Information on this corpus can be found at http://www.laits.utexas.edu/spt/.

19 Detailed information on FLLOC can be found at http://www.flloc.soton.ac.uk/index.html.

20 A longitudinal written and oral corpus of L2 Spanish is currently being collected at the University of Southampton by members of the SPLLOC and FLLOC team under a study abroad research program called LANG-SNAP *(Languages and Social Networks Abroad Project)*. The specific aims of the project are to document the development of students' knowledge and use of the target language over a 23-month period including a 9-month stay abroad. More information on this project can be found at http://langsnap.soton.ac.uk/.

21 It is questionable whether data resulting from elicitation tasks like those used in these studies can actually be considered corpus data (see the discussion in Section 1.2.1).

22 The native speakers' subcorpus contains about 25% of the total number of words in CEDEL2. Participants are mostly university students who contribute to the corpus under the same conditions as the learners and write on the same topics. For more information, see http://www.uam.es/proyectosinv/woslac/collaborating.htm.

23 The placement test used was The University of Wisconsin College-Level Placement Test, which can be easily administered online (see Lozano and Mendikoetxea, forthcoming, for more details).

24 There is no standard measure of learners' proficiency in SPLLOC; learners are classified into three levels according to age and number of years studying Spanish, corresponding to institutional levels (Year 9, Year 11, and so on).

25 A full list of publications using CEDEL2 can be found at http://wdb.ugr.es/~cristoballozano/?page_id=64.

26 I am grateful to an anonymous reviewer for bringing to my attention some of the references mentioned in this section.

27 Biber *et al.* (2006) provide the first multidimensional analysis of native Spanish by analyzing a 20 million-word corpus, with written and spoken data in 19 registers. Parodi (2005, 2007) uses a 2.5 million-word corpus to study the differences between written and spoken (native) Spanish.

REFERENCES

Alonso Ramos, Margarita, Leo Wanner, Nancy Vázquez Veiga, Orsolya Vincze, Estela Mosqueira Suárez, and Sabela Prieto González. 2010. ''Tagging Collocations for Learners.'' In *eLexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELEX2009, Cahiers du CENTAL 7*, edited by Sylviane Granger and Magali Paquot, 375–380. Louvain-la-Neuve: Presses Universitaires de Louvain.

Alonso Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. 2010. ''Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora.'' In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valetta, Malta: Language Resources Evaluation. Available at http://www.lrec-conf.org/proceedings/lrec2010/index.html.

Arche, María J., and Laura Domínguez. 2011. ''Tracking Morphology and Syntax Dissociation in SLA: Evidence from L2 Clitic Acquisition in Spanish.'' In *Morphology and its Interfaces*, edited by Alexandra Galani, Glyn Hicks, and George Tsoulas, 291–319. Amsterdam and Philadelphia: John Benjamins.

Asención-Delaney, Yuly, and Joseph Collentine. 2011. ''A Multidimensional Analysis of a Written L2 Spanish Corpus.'' *Applied Linguistics*, 32(3): 299–322.

Aston, Guy, Silvia Bernardini, and Dominic Stewart, eds. 2004. *Corpora and Language Learners*. Amsterdam and Philadelphia: John Benjamins.

Barlow, Michael. 2005. ''Computer-based Analysis of Learner Language.'' In *Analysing Learner Language*, edited by Rod Ellis and Gary P. Barkhuizen, 335–357. Oxford: Oxford University Press.

Biber, Douglas, Mark Davies, James Jones, and Nicole Tracy-Ventura. 2006. ''Spoken and Written Register Variation in Spanish: A Multi-dimensional Analysis.'' *Corpora*, 1(1): 1–37.

Campillos Llanos, Leonardo. 2012. *La Expresión Oral en Español Lengua Extranjera: Interlengua y Análisis de Errores Basado en Corpus*. Unpublished PhD dissertation, Universidad Autónoma de Madrid.

Cestero Mancera, Ana M., and Inmaculada Penadés Martínez. 2009. *Corpus de Textos Escritos para el Análisis de Errores de Aprendices de E/LE (CORANE)*. CD-ROM. Alcalá de Henares: Universidad de Alcalá.

Cestero Mancera, Ana M., Inmaculada Penadés Martínez, Ana Blanco Canales, Laura Camargo Fernández, and José Simón Granda. 2001. ''Corpus para el Análisis de Errores de Aprendices de E/LE (*CORANE*).'' In *Actas del XII Congreso Internacional de ASELE: Tecnologías de la Información y de las Comunicaciones en la Enseñanza de la E/LE,*edited by Ana Gimeno Sanz, 527–534. Valencia: Editorial U.P.V.

Cheng, Ann Chung, Hui Chuan Lu, and Panayotis Giannakouros. 2008. ''The Uses of Spanish Copulas by Chinese-speaking Learners in a Free Writing Task.'' *Bilingualism: Language and Cognition*, 11: 301–317.

Collentine, Joseph. 2004. ''The Effects of Learning Contexts on Morphosyntactic and Lexical Development.'' *Studies in Second Language Acquisition*, 26(2): 227–48.

Collentine, Joseph. 2009. ''Study Abroad Research: Findings, Implications and Future Directions.'' In *Handbook of Language Teaching*, edited by Michael H. Long and Catherine Doughty, 218–34. Oxford: Wiley-Blackwell.

Collentine, Joseph, and Yuly Asención-Delaney. 2010. ''A Corpus-based Analysis of the Discourse Functions of *Ser/Estar*+adjective in Three Levels of Spanish FL learners.” *Language Learning*, 60(2): 409–445.

Collentine, Karina. 2011. “Learner Autonomy in Task-based 3D World and Production.” *Language Learning and Technology*, 15(3): 50–67.

Díaz, Lourdes. 2007. *Interlengua Española: Estudio de Casos*. Barcelona: Printulibro Intergrup.

Díaz-Negrillo, Ana, and Paul Thompson. Forthcoming. “Learner Corpora: Looking Towards the Future.” In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by Ana Díaz Negrillo, Nicolas Ballier, and Paul Thompson. Amsterdam and Philadelphia: John Benjamins.

Domínguez, Laura, and María J. Arche. 2008. “Optionality in L2 Grammars: The Acquisition of SV/VS Contrast in Spanish.” In *Proceedings of the 32 Annual Boston University Conference on Language Development*, edited by Harvey Chan, Heather Jacob, and Enkeleida Kapia, 96–107: Somerville, MA: Cascadilla Press.

Domínguez, Laura, Nicole Tracy-Ventura, María J. Arche, Rosamond Mitchell, and Florence Myles. 2012. “The Role of Dynamic Contrasts in the L2 Acquisition of Spanish Past Tense Morphology.” *Bilingualism*: *Language and Cognition*. First View, 1–20. doi:10.1017/S1366728912000363

Ellis, Nick C. 2008. “Phraseology: The Periphery and the Heart of Language.” In *Phraseology in Foreign Language Learning and Teaching*, edited by Fanny Meunier and Sylviane Granger, 1–13. Amsterdam and Philadelphia: John Benjamins.

Escutia, Marciano. 2010. “El Uso de *se* con Inacusativos por Estudiantes Avanzados de Español como Lengua Extranjera: Transferencia y Restructuración.” *RESLA Revista Española de Lingüística Aplicada*, 23: 129–151.

Escutia, Marciano. 2012. “Expletives and Unaccusative Predicates in L2A.” *Higher Education of Social Science*, 2(3): 1–14.

Ferrandis, Esther. In progress. *Cross-linguistic Interference at the Interfaces: Subjects in L2 Grammars*. Unpublished PhD dissertation, Universidad Autónoma de Madrid.

Fitzpatrick, Eileen, ed. 2007. *Corpus Linguistics Beyond the Word*. Amsterdam: Rodopi.

Gass, Susan M., and Alison Mackey. 2007. *Data Elicitation for Foreign and Second Language Research*. New York: Routledge.

Geeslin, Kimberly. 2000. “A New Approach to the Second Language Acquisition of Copula Choice in Spanish.” In *Spanish Applied Linguistics at the Turn of the Millennium: Papers from the 1999 Conference on the L1 & L2 Acquisition of Spanish and Portuguese*, edited by Ronald P. Leow and Cristina Sanz, 50–66. Somerville, MA: Cascadilla Press.

Geeslin, Kimberly. 2003. “A Comparison of Copula Choice in Advanced and Native Spanish.” *Language Learning*, 53: 703–764.

Gilquin, Gaëtanelle. 2001. “The Integrated Contrastive Model. Spicing up your Data.” *Languages in Contrast*, 3(1): 95–123.

Gilquin, Gaëtanelle and Stefan Th. Gries. 2009. “Corpora and Experimental Methods: A State of-the-Art Review.” *Corpus Linguistics and Linguistic Theory*, 5(1): 1–26.

Granger, Sylviane. 1996. “From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora.” In *Languages in Contrast: Text-based Cross-linguistic Studies. Lund Studies in English 88*, edited by Karin Aijmer, Bengt Altenberg, and Mats Johansson, 37–51. Lund: Lund University Press.

Granger, Sylviane, 1998. “The Computerized Learner Corpus: A Versatile New Source of Data for SLA Research.” In *Learner English on Computer*, edited by Sylviane Granger, 3–18. London and New York: Addison Wesley Longman.

Granger, Sylviane. 2002. “A Bird’s-Eye View of Computer Learner Corpus Research.” In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, edited by Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, 3–33.

Amsterdam and New York: John Benjamins.

Granger, Sylviane. 2004. ''Computer Learner Corpus Research: Current Status and Future Prospects.'' In *Applied Corpus Linguistics: A Multidimensional Perspective*, edited by Ulla Connor and Thomas A. Upton, 123–146. Amsterdam: Rodopi.

Granger, Sylviane. 2008. ''Learner Corpora.'' In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling, and Merja Kytö, 259–275. Berlin: Mouton de Gruyter.

Granger, Sylviane. 2009. ''The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching.'' In *Corpora and Language Teaching*, edited by Karin Aijmer, 13–23. Amsterdam and Philadelphia: John Benjamins.

Granger, Sylviane. 2012. ''How to Use Second and Foreign Language Learner Corpora.'' In *Research Methods in Second Language Acquisition: A Practical Guide*, edited by Alison Mackey and Susan M. Gass, 7–29. London: Wiley-Blackwell.

Granger, Sylviane, Estelle Dagneaux, and Fanny Meunier. 2002. *The International Corpus of Learner English*. (Handbook and CD-ROM). Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2* (Handbook + CD-Rom). Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier. Forthcoming. *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. Proceedings of LCR 2011*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson, eds. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam and Philadelphia: John Benjamins. Granger, Sylviane, and Magali Paquot. 2010. *eLexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELEX2009, Cahiers du CENTAL 7*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gries, Stefan Th. 2008. ''Corpus-based Methods in Analyses of Second Language Acquisition Data.'' In *Handbook of Cognitive Linguistics and SLA*, edited by Peter Robinson and Nick C. Ellis, 406–431. New York: Routledge.

Hawkins, Roger, and Cecilia Yuet-hung Chan. 1997. ''The Partial Availability of Universal Grammar in Second Language Acquisition: The 'Failed Functional Features Hypothesis'.'' *Second Language Research*, 13: 187–226.

Housen, Alex, and Folkert Kuiken. 2009. ''Complexity, Accuracy and Fluency in Second Language Acquisition''. In *Applied Linguistics*. *Special Issue: Complexity, Accuracy and Fluency (CAF) in Second Language Acquisition Research*, 30(4): 461–473.

Ife, Anne. 2004. ''The L2 Learner Corpus: Reviewing its Potential for the Early Stages of Learning.'' In *Applied Linguistics at the Interface*, edited by Mike Baynham, Alice Deignan, and Goodith White 91–103. London: Equinox.

Koike, Dale. 2007. *Spanish Learner Corpus and Exercises*. Austin, TX: University of Texas.

Lardiere, Donna. 1998. ''Dissociating Syntax from Morphology in a Divergent L2 End-state Grammar.'' *Second Language Research*, 14(4): 359–375.

Liceras, Juana M., and Lourdes Díaz. 1998. ''On the Nature of the Relationship between Morphology and Syntax: Inflectional Typology, *f*-Features and Null/Overt Pronouns in Spanish Interlanguage.'' In *Morphology and its Interfaces in Second Language Knowledge*, edited by Marie-Louise Beck, 307–338. Amsterdam and Philadelphia: John Benjamins.

Liceras, Juana M., and Lourdes Díaz. 1999. ''Topic-drop versus Pro-drop: Null Subjects and Pronominal Subjects in the Spanish L2 of Chinese, English, French, German and

Japanese speakers.'' *Second Language Research*, 15(1): 1–40.

Liceras, Juana M., Denyse Maxwell, Biana Laguardia, Zara Fernández, Raquel Fernández, and Lourdes Díaz. 1997. ''A Longitudinal Study of Spanish Non-Native Grammars: Beyond Parameters.'' In *Contemporary Perspectives on the Acquisition of Spanish. Vol. 1: Developing Grammars*, edited by Ana Teresa Pérez-Leroux and William R. Glass, 99–132. Somerville, MA: Cascadilla Press.

Liceras, Juana. M., Elena Valenzuela, and Lourdes Díaz .1999. ''L1 and L2 Spanish Developing Grammars and the 'Pragmatic Deficit Hypothesis'.'' *Second Language Research*, 15(2): 161–190.

Lozano, Cristóbal. 2009a. ''CEDEL2: Corpus Escrito del Español como L2.'' In *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente*, edited by Carmen M. Bretones Callejas, José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, María Elena García Sánchez, Mª Enriqueta Cortés de los Ríos, Sagrario Salaberri Ramiro, Mª Soledad Cruz Martínez, Nobel Perdú Honeyman, and Blasina Cantizano Márquez, 197–212. Almería: Universidad de Almería.

Lozano, Cristóbal. 2009b. ''Selective Deficits at the Syntax-Discourse Interface: Evidence from The CEDEL2 Corpus.'' In *Representational Deficits in Second Language Acquisition*, edited by Neal Snape, Yan-Kit Ingrid Leung, and Michael Sharwood-Smith, 127–166. Amsterdam and New York: John Benjamins.

Lozano, Cristóbal. 2011. ''When Corpus Data Reveal What Experimental Data May Hide: The Distribution of Overt and Null Pronouns in L2 Spanish (CEDEL2 corpus).'' Paper presented at the *Learner Corpus Research* conference, Université Catholique de Louvain.

Lozano, Cristóbal, and Amaya Mendikoetxea. 2008. ''Postverbal Subjects at the Interfaces in Spanish and Italian Learners of L2 English: A Corpus Analysis.'' In *Linking up Contrastive and Learner Corpus Research*, edited by Gaëtanelle Gilquin, Szilvia Papp, and María Belén Díez-Bedmar, 85–125. Amsterdam: Rodopi.

Lozano, Cristóbal, and Amaya Mendikoetxea. 2010. ''Postverbal Subjects in L2 English: A Corpus-based Study.'' *Bilingualism: Language and Cognition*, 13(4): 475–497.

Lozano, Cristóbal, and Amaya Mendikoetxea. Forthcoming. ''Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL2.'' In *Automatic Treatment and Analysis of Learner Corpus Data*, edited by Ana Díaz Negrillo, Nicolas Ballier, and Paul Thompson. Amsterdam and Philadelphia: John Benjamins.

Lu, Hui-Chuan. 2012. ''An Annotated Taiwanese Learners' Corpus of Spanish, CATE.'' *Corpus Linguistics and Linguistic Theory*, 6(2): 297–300.

Lüdeling, Anke, and Merja Kytö, eds. 2008. *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

Mackey, Alison, and Susan M. Gass. 2005. *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analysing Language (3rd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marsden, Emma, and Anabelle David. 2008. ''Vocabulary Use during Conversation: A Cross-sectional Study of Development from Year 9 to Year 13 amongst Learners of Spanish and French.'' *Language Learning Journal*, 36(2): 181–198.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Methods, Theory and Practice*. Cambridge: Cambridge University Press.

Mendikoetxea, Amaya, and Cristóbal Lozano.Forthcoming. ''Conceptual and Methodological Interfaces in SLA Research: Triangulating Corpus and Experimental Data in L2 Verb-Subject Alternations.'' *Corpus Linguistics and Linguistic Theory*.

Meunier, Fanny, and Céline Gouverneur. 2009. ''New Types of Corpora for New Educational Challenges: Collecting, Annotating and Exploiting a Corpus of Textbook Material.'' In *Corpora and Language Teaching*, edited by Karin Aijmer, 179–201. Amsterdam and Philadelphia: John Benjamins.

Mitchell, Rosamond, Laura Domínguez, María J. Arche, Florence Myles, and Emma Marsden. 2008. ''SPLLOC: A New Corpus for Spanish Second Language Acquisition Research.'' In *EUROSLA Yearbook 8*, edited by Leah Roberts, Florence Myles, and Annabelle David, 287–304. Amsterdam and Philadelphia: John Benjamins.

Mitchell, Rosamond, and Florence Myles. 2004. *Second Language Learning Theories*. London: Arnold.

Myles, Florence. 2005. ''Interlanguage Corpora and Second Language Acquisition Research.'' *Second Language Research*, 21 (4): 373–391.

Myles, Florence. 2007a. ''Investigating Learner Language Development with Electronic Longitudinal Corpora: Theoretical and Methodological Issues.'' In *The Longitudinal Study of Advanced L2 Capacities*, edited by Lourdes Ortega and Heidi Byrnes, 58–72. London and New York: Routledge.

Myles, Florence. 2007b. ''Using Electronic Corpora in SLA Research.'' In *Handbook of French Applied Linguistics*, edited by Dalila Ayoun, 377–400. Amsterdam and New York: John Benjamins.

Nesselhauf, Nadja. 2004. ''Learner Corpora and their Potential for Language Teaching.'' In *How to Use Corpora in Language Teaching*, edited by John Sinclair, 125–153. Amsterdam and Philadelphia: John Benjamins.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam and Philadelphia: John Benjamins.

O'Keefe, Anne, and Michael McCarthy, eds. 2007. *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.

Ortega, Lourdes. 2000. *Understanding Syntactic Complexity: The Measurement of Change in the Syntax of Instructed L2 Spanish Learners*. Unpublished PhD dissertation, University of Hawaii at Manoa.

Ortega, Lourdes. 2003. ''Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing.'' *Applied Linguistics*, 24(4): 492–518.

Oshita, Hiroyuki. 2000. ''What is Happened May Not Be What Appears to be Happening: A Corpus Study of 'Passive' Unaccusatives in L2 English.'' *Second Language Research*, 16(4): 293–324.

Oshita, Hiroyuki. 2004. ''Is There Anything There When *There* Is Not There? Null Expletives and Second Language Data.'' *Second Language Research*, 20(2): 95–130.

Parodi, Giovanni. 2005. ''Lingüística de Corpus y Análisis Multidimensional: Exploración de la Variación en el Corpus PUCV-2003.'' *Revista Española de Lingüística*, 35(1): 45–76.

Parodi, Giovanni. 2007. ''Variation across Registers in Spanish: Exploring the El Grial PUCV Corpus.'' In *Working with Spanish Corpora*, edited by Giovanni Parodi, 11–35. London: Continuum.

Prévost, Philippe and Lydia White. 2000. ''Missing Surface Inflection or Impairment in Second Language? Evidence from Tense and Agreement.'' *Second Language Research*, 16(2): 103–133.

Rollinson, Paul, and Amaya Mendikoetxea. 2010. ''Learner Corpora and Second Language Acquisition: Introducing WriCLE.'' In *Analizar datos > Describir variación / Analysing Data > Describing Variation*, edited by Jorge L. Bueno Alonso, Dolores González Álvarez, Úrsula Kirsten Torrado, Ana E. Martínez Insua, Javier Pérez-Guerra, Esperanza Rama Martínez, and Rosalía Rodríguez Vazquez, 1–12. Vigo: Universidade de Vigo (Servizo de Publicacións).

Rutherford, William, and Margaret Thomas. 2001. ''The Child Language Data Exchange System in Research on Second language

Acquisition.'' *Second Language Research*, 17(2): 195–212.

Sinclair, John. 1996. ''EAGLES: Preliminary Recommendations on Corpus Typology.'' http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html. Accessed January 31, 2013.

Sinclair, John. 2004. *How to Use Corpora in Language Teaching*. Amsterdam and Philadelphia: John Benjamins.

Sinclair, John. 2005. ''How to Build a Corpus.'' In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne, 79–83. Oxford: Oxbow Books.

Sorace, Anonella. 2005. ''Selective Optionality in Language Development.'' In *Syntax and Variation: Reconciling the Biological and the Social*, edited by Leonie Cornips and Karen P. Corrigan, 55–80. Amsterdam and Philadelphia: John Benjamins.

Sorace, Antonella, and Ludovica Serratrice. 2009. ''Internal and External Interfaces in Bilingual Language Development: Beyond Structural Overlap.'' *International Journal of Bilingualism*, 13(2): 195–210.

Thoday, Elizabeth. 2008. ''Issues in Building Learner Corpora: An Investigation into the Acquisition of German Passive Constructions.'' *Newcastle Working Papers in Linguistics*, 14: 145–155.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

Tono, Yukio. 2003. ''Learner Corpora: Design, Development and Applications.'' In *Proceedings of the 2003 Corpus Linguistics Conference*, edited by Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, 800–809. UCREL, Lancaster University: UCREL Technical Paper number 16.

Tsimpli, Ianthi-Maria, and Anna Roussou. 1991. ''Parameter Resetting in L2?'' *UCL Working Papers in Linguistics*, 3: 149–169.

Tsimpli, Ianthi-Maria, Antonella Sorace, Caroline Heycock, and Francesca Filiaci. 2004. ''First Language Attrition and Syntactic Subjects: A Study of Greek and Italian Near-Native Speakers of English.'' *International Journal of Bilingualism*, 8(3): 257–277.

White, Lydia. 2003. *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

Wynne, Martin. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books.