Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins, pp. 65-100.

Learner corpora and second language acquisition

The design and collection of CEDEL2*

Cristóbal Lozano and Amaya Mendikoetxea

Second language acquisition (SLA) research has traditionally relied on elicited experimental data, and it has disfavoured natural language use data. Learner corpus research has the potential to change this but, to date, the research has contributed little to the interpretation of L2 acquisition, and some of the corpora are flawed in design.
We analyse the reasons why many SLA researchers are still reticent about using corpora, and how good corpus design and adequate tools to annotate and search corpora can help overcome some of the problems observed. We do so by describing how the ten standard principles used in corpus design (Sinclair 2005) were applied to the design of CEDEL2, a large learner corpus of L1 English – L2 Spanish (Lozano 2009a).

1. Introduction

The main aim of second language acquisition (SLA) research is to build models of the underlying representations of learners at a particular stage in the process of L2 acquisition and of the developmental constraints that limit L2 production. The central source of evidence for these mental processes is the language produced by learners, whether spontaneously or through data elicitation procedures (Myles 2005: 374). The success of SLA research relies crucially on the validity and reliability of these data elicitation, and data collection, procedures.

* This research has been partly funded by research grants HUM2005-01782/FILO (Spanish Ministry of Education), FFI2008-01584 (Spanish Ministry of Science and Innovation) and FFI2012-30755 (Spanish Ministry of Economy and Competitiveness), which we gratefully acknowledge. We are also grateful to an anonymous reviewer for detailed and very useful comments on the initial manuscript. We would also like to thank the participants (both learners and researchers) in the compilation of CEDEL2 since 2005.

Much SLA research has traditionally relied on elicited experimental data while disfavouring natural language data. While the use of large-scale corpora has been standard practice in L1 acquisition research over the past twenty five years or so (CHILDES, MacWhinney 2000), large L2 corpora are still scarce, except for ICLE (Granger et al. 2002a) and other non-commercially available corpora (see Tono 2005; Granger 2008; see also Section 2 below for an overview). Consequently, relatively little use has been made of corpora in L2 research, particularly in formal approaches to second language acquisition (SLA), and many SLA researchers are still reticent about using corpus data.

The area of linguistic inquiry known as *learner corpus research* has recently come into being as a result of the confluence of two previously disparate fields: corpus linguistics and SLA (Granger 2002, 2004). On the whole, the contribution of learner corpus research so far has been much more substantial in description than interpretation of SLA data (Granger 2004: 134–135), with very little reference to current debates and hypotheses about SLA (Myles 2005), as will be illustrated in Section 2.

In this paper, we analyse the reasons why many SLA researchers are still reticent about using corpora and how good corpus design and adequate tools to annotate and search corpora could help overcome some of the problems observed. We do so by describing ten key principles (proposed by Sinclair 2005) applied to the design of a learner corpus of L2 Spanish (CEDEL2) and its contribution to SLA research.

2. Learner corpora in SLA research

In this section we will first present some of the reasons why the use of learner corpora has not been standard practice in SLA research. Then, we will offer an overview of corpora in language acquisition research, which will be followed by an introduction to learner corpora. This will provide the background in which we will justify the creation of a new learner corpus of L2 Spanish (CEDEL2), under certain design principles which will be made explicit.

2.1 A bias in second language research

Traditionally, the study of SLA from a formal perspective has typically (but not exclusively) used experimental and introspective methods such as grammaticality judgement tasks, acceptability tasks and other types of comprehension tests (see overviews in Hawkins 2001, 2007; White 2003, 2009; Mitchell & Myles 2004; Slabakova et al. 2006; Liceras et al. 2008). As pointed out by Granger (2002: 5) "much current SLA research favours experimental, metalinguistic and introspective data, and tends to be dismissive of natural language use data".

Several reasons can be given for why elicitation techniques are favoured in SLA research. For instance, Mackey & Gass (2005) provide the following reasons why metalinguistic data may be used in SLA research, as opposed to natural language use data: (i) the particular structure you want to investigate may not occur in natural production: it may be absent or there may not be enough instances, and, conversely, (ii) to answer your research question you may need to know what learners rule out as a possible L2 sentence: (a) *presence* of a particular structure/feature in the learners' natural output does not necessarily indicate that the learners *know* (i.e. have a mental representation of) the structure, and (b) *absence* of a particular structure/feature in natural language use data does not necessarily indicate that learners do not know the structure. An additional reason is provided by Granger (2002: 6): it is difficult to control the variables that affect learner production in a non-experimental context. Additionally, L2 researchers have been typically trained in (quasi)experimental methods rather than in corpus methods, except for those studies conducted with source data from CHILDES (see Myles 2007b: 386 for a discussion). The consequence of all this is that the empirical base of SLA research tends to be relatively narrow, based on the language produced by a very limited number of subjects, which, as pointed out by Granger (2002: 6), raises questions about whether results can be generalised. But the methodological future of SLA looks promising, since some researchers are currently claiming that combining both naturalistic and experimental data is crucial to gain insight into the relation between the two types of data (e.g. Gilquin & Gries 2009).

Case studies and small-scale experimental studies have greatly served the hypothesis-building endeavour in SLA research, but there are now many researchers who feel that the time has come to test hypotheses on larger and better constructed databases to see whether findings can be generalised (see Myles 2005) and to discover sets of data not normally found in small studies which can become crucial in order to inform currents debates in the SLA discipline (e.g. what aspects of grammar are more vulnerable to transfer or cross-linguistic influence, what is the role of the interfaces, for example, syntax-discourse, lexicon-syntax, syntax-phonology, in L2 acquisition, etc.). These are the main reasons for using corpora in SLA research, to which we can add another two which are common to corpus linguistics in general as a field of inquiry: to discover patterns of use and for quantitative studies (e.g. frequency).

2.2 Corpora in language acquisition research

The use of corpora in L1 acquisition research is not new, as it has been standard practice to use them in studies of child language since the 70s, though experimental methods have been also customary since the 60s. 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 since the 80s. It has also been recently employed in SLA research (see discussions in Rutherford & Thomas 2001; Myles 2005). The CHILDES collection contains over 44 million words in over 30 subcorpora sampling different languages, most of which are grammatically tagged (CHILDES, 2010). This wealth of data is reflected in the publication of at least 3,200 research papers using CHILDES as their source of data. The use of such large-scale naturalistic data in L1 research has meant a massive leap forward in our understanding of how child grammars are acquired and developed. By the same token, the use of massive naturalistic data in SLA will inevitably broaden our understanding of how learner grammars develop.

Unlike the scenario just described for L1 acquisition research, large-scale L2 corpora are rather scarce, except for ICLE (Granger et al. 2002a) and other non-commercially available corpora (see Granger 2008; Tono 2005 for an overview), which we will review below. On the whole, the contribution of learner corpus research so far 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 L2 acquisition and their implications for learner language development (Myles 2005). In other words, large-scale learner corpora have been used in pedagogical and functional approaches to SLA, with an emphasis on description over interpretation (Granger 2004: 134–135), as can be observed in recent publications (e.g. Burnard & McEnery 2000; Granger 2002; Granger et al. 2002b; Aston et al. 2004; Granger 2004; Sinclair 2004; Reppen 2006; Hidalgo et al. 2007; Aijmer 2009). Importantly, the use of such large-scale learner corpora has not been a trademark of formal approaches to investigating the acquisition and development of interlanguage grammars (see overview in Tono 2003: 805–806). This situation has started to change in the past few years as researchers are becoming increasingly aware of the benefits of analysing extensive naturalistic data to understand L2 grammar acquisition and development, as we will see in the following sections.[1]

1. Note that the use of massive naturalistic data to investigate grammatical phenomena is not new as there is a long tradition in the field of corpus linguistics Biber et al. 1998; Hunston & Francis 2000; Leech et al. 2001; Reppen & Simpson 2002; Coffin et al. 2004; Connor & Upton 2004; Baker et al. 2006; McEnery et al. 2006; Renouf & Kehoe 2006; Fitzpatrick 2007; Lüdeling & Kytö 2008). In the last two decades we have seen the emergence of large-scale native English corpora such as the BNC, *British National Corpus*, containing around 100 million words and the COCA, *Corpus of Contemporary American English*, with over 410 million words (see Davies 2010 for an overview). This has led to the creation of corpora in other native languages. For native Spanish we have the *Corpus del Español*, with approximately 100 million words (Davies 2010), the CREA, *Corpus de Referencia del Español Actual*, with 154 million words (Real Academia Española 2010a), and the CORDE, *Corpus Diacrónico del Español*, containing around 250 million words (Real Academia Española 2010b).

2.3 An overview of learner corpora and learner corpus research

The creation of learner corpora has been conditioned by a tension between an *inductive* vs. *deductive* approach in language acquisition research. Deductive (top-down) approaches depart from an initial hypothesis that will be confirmed or rejected by the data (corpus), hence the corpus is just a tool to test hypotheses. Inductive (bottom-up) approaches typically use the corpus as an exploratory tool to arrive at a hypothesis. In short, either the hypothesis is formulated and then is (dis)confirmed in the corpus, or the corpus is explored so as to formulate a hypothesis (see Myles 2007b for an overview and a discussion). Thus, studies using learner corpora in SLA fall within two categories (i) *hypothesis-driven/corpus-based* studies and (ii) *hypothesis-finding/corpus-driven* studies (see Granger 1998 and Tognini-Bonelli 2001). According to Barlow (2005: 344), the former involve using learner corpus data to test specific hypotheses or research questions about the nature of learner language generated through introspection, SLA theories, or as a result of the analysis of experimental or other sources of data. The latter involve investigating learner corpus data in a more exploratory way to discover patterns of data, which may then be used to generate hypotheses about learner language. The majority of studies within the area of learner corpus research fall within category (ii), as revealed by an analysis of the papers collected in recent edited volumes within the field (e.g. Granger et al. 2002b; Aston et al. 2004). Hypothesis-driven, corpus-based studies are hard to find.[2]

All in all, the contribution of learner corpus research so far has been much more substantial in description than interpretation of SLA data, documenting differences between native and non-native English, rather than explaining and addressing the key theoretical issues in SLA research (Granger 2004; Myles 2005). According to Granger (2004: 134–135), this is because learner corpus research has been mainly conducted by corpus linguists, rather than SLA specialists (Hasselgard 1999), and the type of learner language corpus that researchers have been most interested in (intermediate to advanced) was so poorly described in the literature that they felt the need to establish the facts before launching into theoretical generalisations. As Tono summarises:

> Many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them [i.e. SLA researchers] effectively, while SLA researchers typically know little about what corpora can do for them. (Tono 2003: 806)

2. Two examples from the volumes mentioned are Housen (2002) and Tono (2004). Our own work (see Lozano & Mendikoetxea 2008, 2010) also falls within that category.

There are SLA researchers who have collected and analysed relatively large amounts of naturalistic learner data, as is the case in Lardiere (1998), who uses data from an English learner, Patty, coming from email exchanges collected over several years. While this type of studies allows for a detailed and longitudinal study of interlanguage development, conclusions from case studies are limited as they cannot be extrapolated to other learners. The ESF project, *Second Language Acquisition in Adult Immigrants* (Perdue 1993), is one of the best-known examples of the use of corpora (with different L1-L2 combinations) to study L2 acquisition from a rather functional approach. These corpora are now part of the CHILDES database.

The publication of the first version of the *International Corpus of Learner English*, ICLE (Granger et al. 2002a) can be taken as the starting point in the exploitation of large-scale learner corpora and has inspired a growing interest in learner corpus research. Over the past few years, over 400 published L2 papers have used ICLE, though most of them are rather descriptive and/or pedagogical in nature, as discussed above.[3] The first version of ICLE (Granger et al. 2002) consists of 2.5 million words of argumentative essays written by university students with L2 English from several European countries, organized in different subcorpora divided according to L1: Spanish, Italian, French, Russian, etc. Such subcorpora allow for a new type of analysis: *Contrastive Interlanguage Analysis* (CIA), i.e. the contrast of two (or more) interlanguage varieties, e.g. L1 Spanish – L2 English vs. L1 Italian – L2 English. ICLE also allows for interlanguage vs. native language contrasts with the help of an equivalent native English corpus, *Louvain Corpus of Native English Essays* (LOCNESS), containing approximately 235,000 words coming from argumentative essays written by British and North American students. An expanded version of ICLE has been recently released (Granger et al. 2009). It contains 3.7 million words from 16 mother-tongue backgrounds, including now Chinese, Japanese, Turkish and other L1s.

The influence of ICLE can be observed in the creation of similarly designed L2 English learner corpora in Spain. The *Written Corpus of Learner English* (WriCLE)

3. See <http://www.uclouvain.be/en-cecl-lcBiblio.html> for a list of published research using the ICLE corpus. That learner corpus research is now a burgeoning field can be also seen in the publication of learner corpus studies in theoretical and descriptive journals like *International Journal of Corpus Linguistics* and *Corpus Linguistics and Linguistic Theory*, the conferences devoted to the topic (like the recent *Learner Corpus Research* conference held in Louvain-la-Neuve in 2011), the presence of learner corpus panels in corpus linguistics conference (for instance *ICAME* and *Corpus Linguistics*), the publication of methodological books on the use of (learner and native) corpora in applied linguistics (Hunston 2002; McEnery 2005), the publication of recent papers on the need to use learner corpora in second language acquisition research (Myles 2005, 2007a, 2007b), as well as in the recent publication of books on SLA research methods justifying the use of learner corpora as a valuable research tool (Brown & Rodgers 2002; Chaudron 2003; Mackey & Gass 2005; Dörnyei 2007).

(Rollinson & Mendikoetxea 2010) is being created at the *Universidad Autónoma de Madrid*. It is an L1 Spanish – L2 English written corpus whose target is one million words written by first and third year undergraduate students of English, mostly with an upper-intermediate to advanced proficiency level. Unlike ICLE, in WriCLE we can find measures of the proficiency level of the students according to the *Common European Framework of Reference for Language*s, determined by a standardized placement test. As we will discuss below, knowing the proficiency level of each learner in the corpus is essential. WriCLE contains under 700,000 words consisting of over 700 academic essays written by students. A subcorpus is currently being compiled of non-academic writing: mostly blogs (WriCLEinf) to allow comparison across different registers, as well as the study of structures not normally found in more formal, academic writing (e.g. questions).[4]

An important number of large L2 corpora have been created over the past few years to meet the needs of EFL materials designers. We will briefly mention two other large learner corpora: the *Longman Learner Corpus* (LLC) and the *Cambridge Learner Corpus* (CLC), both containing data from compositions written by L2 English learners with different L1s. None of them is commercially available since their use is restricted to the creation of pedagogical material for EFL learners by editorial staff (Pearson-Longman and Cambridge University Press respectively), though an exception to this seems to be Oshita's (2000, 2004) published research on the acquisition of intransitive structures, based on the LLC corpus as a source of data.

While the initial publication of ICLE in 2002 has popularised the use of learner corpora as a source of data in L2 research, most of the studies done with this corpus have analysed lexical aspects of learner language, probably due to the fact that the first version of ICLE was not annotated morphosyntactically, hence concordancers and query software are limited to searching lemmas and their morphological variants. Certainly, some researchers have gone *beyond the word* by analysing phrases and structures (Fitzpatrick 2007), collocations (Nesselhauf 2005) and even word order alternations (Gilquin et al. 2008; Lozano & Mendikoetxea

4. The Santiago University Learner of English Corpus (SULEC) is also a corpus of L1 Spanish – L2 English learners at an undergraduate and secondary-school level, representing all proficiency levels (elementary, intermediate and advanced), containing both spoken and written data). There is also an important spoken corpus of L2 French in the CHILDES database that has been collected at the University of Southampton, the French Learner Language Oral Corpus (FLLOC) (see Myles 2007a, 2007b). Other leaner corpora have been on the trail of ICLE, particularly in other European countries and in Asia. For an overview, see Granger (2008) and Tono (2005) and, especially, the Centre for English Corpus Linguistics webpage: <http://www.uclouvain.be/en-cecl-lcWorld.html>. This page is an excellent resource for learner corpus publications, workshops, conferences and so on.

2008, 2010). As we will see in the methodology section, some samples of CEDEL2 have been tagged for syntactic structure and collocations.

2.4 L2 Spanish learner corpora: Introducing CEDEL2

As can be appreciated in the preceding section, the development of learner corpora has followed a similar route to the development of native corpora: the creation of large English normative corpora gave rise to the appearance of L2 English learner corpora. Similarly, the creation of Spanish native corpora has led to the creation of L2 Spanish learner corpora. This is partly due to the recent world-wide interest in the study of the Spanish language. In particular, the number of published monographs, research papers and PhD theses on L2 Spanish has increased noticeably over the past few years, particularly in the USA (Pérez-Leroux & Liceras 2002; Lafford & Salaberry 2003; Montrul 2004).

The *Corpus Escrito del Español como L2*, CEDEL2, (Lozano 2009a) is a written L1 English – L2 Spanish corpus sampling learners of all proficiency levels (beginner, intermediate and advanced), plus a similarly designed Spanish native corpus for comparative purposes.[5] As of March 2011, CEDEL2 has reached around 750,000 words in electronic format, since data are being gathered via an online application.[6] While the data collection is still work in progress, some CEDEL2 samples have been used in published research on the acquisition of pronominal subjects (Lozano 2009b) and learner collocations (Alonso et al. 2010a, 2010b).

CEDEL2 originated in the WOSLAC research group (*Word Order in Second Language Acquisition Corpora*) at the Universidad Autónoma de Madrid.[7] The main aim of the WOSLAC research programme is twofold. First, we are investigating one of the much debated issues in second language research, namely, the role of the interfaces (lexicon-syntax and syntax-discourse) as a potential source of observed deficits in the development of learners' interlanguage grammars (Lozano & Mendikoetxea 2008, 2010) (for a discussion on interfaces, see Sorace 2000, 2004, 2005; Sorace & Serratrice 2009). Secondly, our aim is the completion of two comparable leaner corpora (WriCLE, see Section 2.3, and CEDEL2) to

5. L2 here refers to both 'second language' and 'foreign language'. Though the two terms have been traditionally used to refer to different acquisition settings (naturalistic vs. classroom), the distinction is not relevant for the issues that we are interested in investigating since it is standardly assumed in SLA that the (psycho)linguistic mechanisms that shape and constrain interlanguage grammars are similar independent of the learning setting (for an overview see Hawkins 2001; White 2003; Ellis 2008).

6. The online application for CEDEL2 can be seen at <http://www.uam.es/woslac/start.htm>.

7. See <http://www.uam.es/wosla>c and Chocano et al. 2007 for an overview of the WOSLAC research team.

explore and contrast the role of the interfaces, in such a way that the combinations L1 Spanish – L2 English (WriCLE) and L1 English – L2 Spanish (CEDEL2) will permit us to determine whether such deficits are a result of transfer from the learners' L1, or a by-product of input, or rather a consequence of universal developmental patterns.

Given the increasing interest in L2 Spanish acquisition research CEDEL2 is a welcome new source of naturalistic data for researchers. It complements the recently launched *Spanish Leaner Language Oral Corpus* (SPLLOC) (Mitchell et al. 2008), which has set a landmark in L2 Spanish research.[8] This is an oral L1 English – L2 Spanish corpus sampling all proficiency levels (beginner, intermediate and advanced), yet no standardized proficiency test was administered to measure learners' competence as they were classified in three levels according to age and the number of years studying Spanish. The corpus is transcribed and tagged in CHAT format, which is the standard in the CHILDES database. The SPLLOC design principles are task-based: learner data come from two types of tasks: (i) semi-natural oral tasks belonging to different genres (narratives, interviews, debates and picture descriptions), and (ii) controlled tasks in order to elicit certain structures (for instance, clitic pronouns and specific word orders in SPLLOC 1) to answer the research questions of the project (the development of L2 tense and aspect in SPLLOC2).[9] As we will see below, standard criteria in corpus design warn against designing corpora to elicit specific linguistic structures to suit the linguists' specific research questions.

CEDEL2 will be a new source of data that represents an advance in L2 Spanish research for several reasons:

1. SPLLOC uses *ad hoc* corpus design with a deductive approach (i.e. the corpus is designed to elicit specific linguistic constructions so that the researcher can test a specific research question: see Myles 2007b), but CEDEL2 is based on a more exploratory, inductive approach. It crucially follows the 10 standard design principles recommended by Sinclair (2005) for the creation of a

8. Note that it is not our purpose to provide a complete list of L2 Spanish corpora available. Researchers and practitioners are constantly creating corpora to suit their needs. Corpora are particularly suited to explore the form-function mapping and L2 research which addresses questions related to this is often based on the use of corpora (see, for instance, Asención-Delaney & Collentine 2011 and references mentioned there) (we thank one anonymous reviewers for drawing our attention to this work). L2 Spanish learner corpora are also being created for pedagogical reasons; an example of this is CORANE (*Corpus para el Análisis de Errores de Aprendices de E/LE*) (Cestero Mancera et al. 2001).

9. More information about SPLLOC can be found at http://www.splloc.soton.ac.uk/splloc2/index.html, from where the corpus can be searched and downloaded. This website also contains a list of publications about SPLLOC and the use of this corpus for particular studies.

well-designed corpus (see Lozano 2009a and Section 3 below for an overview of the principles). So, CEDEL2 is designed to potentially answer any L2 research question concerning any linguistic structure.

2. Unlike other learner corpora, it is a large-scale learner corpus (c. 750,000 words to date, aiming at 1 million words in the near future, and coming from c. 2,400 participants), so it will yield more reliable naturalistic data than traditional data.
3. It contains a similarly designed Spanish native speaker subcorpus serving as a control group, which will allow for the reliable contrast of interlanguage data against the native norm under equally comparable conditions, since, as Tono argues, "very few learner corpora incorporate L1 data as an integral part of the design. This will become more important in future learner corpora ... to identify specific features of L1-related errors or over/underuse patterns." (Tono 2003: 803).
4. It allows for *Contrastive Interlanguage Analysis* (CIA) (see Granger 1996; Guilquin 2001) since CEDEL2 (L1 English – L2 Spanish) is similarly designed to a large-scale corpus of non-native English, namely, WriCLE (L1 Spanish – L2 English) (see Rollinson & Mendikoetxea 2010). These language pairings will permit detailed analyses of transfer phenomena in both directions, together with the investigation of language-specific vs. universal influence in L2 acquisition.
5. Unlike other L2 learner corpora which do not include a reliable measure of learner's proficiency, CEDEL2 learners were administered a standardised grammatical placement test, as recommended by Tono (2003), which is essential to conduct reliable and fine-grained studies of L2 acquisition and interlanguage development. This will allow for contrastive analyses of learners' interlanguage at different proficiency levels, as well as the possibility of carrying out developmental research.
6. For each learner, CEDEL2 contains precise and detailed background information (e.g. proficiency level, age of first acquisition, length of exposure, learner's self-rating in the four skills –reading, writing, listening, speaking–, learning environment, language use patterns, etc.), which is essential to conduct L2 research concerning not only interlanguage grammars, but also critical period effects, language use patterns, likely cross-linguistic effects, residence abroad effects, self-rated proficiency vs. real proficiency, (re)sources used in composition writing, etc.

In the next section we describe how CEDEL2 was designed according to ten standard corpus design principles proposed by Sinclair (2005).

3. Design principles in learner corpora for SLA purposes: CEDEL2, a case study

As mentioned above, many learner corpora are designed following an *ad hoc* methodology, i.e. the corpus is designed according to external factors imposed by the researchers. In these cases, the language elicited from learners is semi-naturalistic, since certain tasks are designed to control for the language learners are expected to produce, e.g. in some cases learners are expected to use (morpho) syntactic structures, such as clitic pronouns or specific word orders. Detailed, SLA-informed variables about the learners' background and tasks settings are crucial. A measure of proficiency (as well as a control native corpus) is also required if the corpus is to be used for the study of interlanguage development. Additionally, a variety of learner levels is needed for developmental studies. Accessibility, making the corpus available to other researchers, is also a key factor in the success of corpus-based research in SLA.

Prior to the creation of CEDEL2, it was clear that standard good practice in corpus design had to be followed, as recommended by corpus designers (McEnery et al. 2006; Wynne 2005 and references therein). In particular, the design of CEDEL2 follows ten key design principles proposed by Sinclair (2005) in a guide to good practice for developing linguistic corpora, edited by Wynne (2005) and also Tono's (2003) suggestions for basic considerations in the design of learner corpora. According to Sinclair (2005), a well-designed and carefully-constructed corpus must be guided by certain design criteria, such as representativeness, sampling and balance. These criteria must follow ten principles, which have been applied in the design of CEDEL2.

3.1 Principle 1. Content selection

Corpus content must be selected according to *external* criteria (i.e. the communicative function of the corpus texts) and no *internal* criteria (i.e. those referring to the language of the texts), as stated in (1).

(1) "The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function."
(Sinclair 2005: 1).

As mentioned above, some learner corpus designers have followed internal criteria when using semi-natural or even controlled tasks to elicit specific linguistic structures from their learners, which are the structures those researchers are interested in. In this way, the corpus data are biased according to the corpus language content principle, since there is an imbalance in the linguistic structures elicited

(e.g. clitics) that do not correspond to their frequency of production under natural conditions. As previously stated, though the WOSLAC research team is interested in the deficits that learners show at the interfaces with certain syntactic structures, CEDEL2 was designed following strict external criteria, in such a way that all linguistic structures and lexical items could be well represented in the corpus. This principle is clearly connected to the second principle.

3.2 Principle 2. Representativeness

The corpus contents need to represent the language that it samples, as stated in the principle of representativeness, (2).

(2) "Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen" (Sinclair 2005:2).

In order to meet this principle, the learners in CEDEL2 participated voluntarily and they could freely choose to write about one composition topic out of twelve possible composition titles, as shown in Appendix 2. These composition topics were chosen from standard textbooks used in the teaching of Spanish as a foreign language and represent several degrees of difficulty, ranging from basic descriptive topics typically found in beginners' textbooks (such as *How is the region where you live?*, *Talk about a famous person*), intermediate-level topics involving the use of different verbal tenses (e.g. *What did you do last year during your holidays?*, *What are your plans for the future?*, *Describe a memorable experience*), up to advanced-level argumentative as well as descriptive complex topics requiring a wide range of linguistic structures (e.g. *Talk about the problem of terrorism*, *What do you think about the new law banning smoking?*, *Analyse the main aspects of immigration, Describe a film you have watched recently*). When selecting these topics, we strived for a high degree of inclusiveness and a low degree of language bias, in such a way that these topics could potentially elicit all likely morphosyntactic forms, different verbal aspect and tense (present, past, future), and a wide range of vocabulary. This is important since corpora like ICLE are designed around argumentative essays, which typically show a bias towards certain lexical items (e.g. verbs of opinion: *I think, I believe, I argue*) and also towards certain verbal tenses, which typically show an imbalance of present tenses over past and future tenses. Note that, while beginner-level learners typically chose relatively easy topics, intermediate and advanced learners chose all kind of topics, independently of their proficiency level, which indicates that the sampled language in the corpus is varied and proficiency-level independent.

The representativeness principle in a learner corpus refers not only to the inclusion of all possible linguistic structures and lexical items, but also to the

inclusion of all levels of competence, such that all levels of interlanguage development are represented in the corpus. CEDEL2 samples learners from all proficiency levels (beginner, intermediate and advanced). Unlike other corpora where learners are supposed to have a certain proficiency level (ICLE) or are just classified according to their educational/classroom level (SPLLOC), the learners in CEDEL2 were classified according to an independent and standardized Spanish placement test (University of Wisconsin 1998). It was initially envisaged to use DIALANG as a placement test, which is a software application based on the *Common European Framework of Reference* and dividing learners' proficiency into six standard levels (A1, A2, B1, B2, C1, C2). While its use would have been ideal in terms of reliability and contrast with other standardized levels (such as the UCLES system), it requires the download of a computer application. This means that, in most cases, our learners cannot download the software onto the language labs and computer labs where they are studying (schools and universities all over the world), since computers in most computer labs do not permit the installation of downloaded software. Hence, it was decided to use the University of Wisconsin placement test as an online application, as shown in Appendix 4.

Finally, the principle of representativeness is also related to whether the corpus is *longitudinal* or *cross-sectional*. As is standard practice in large learner corpora, it is logistically difficult to sample any given group of learners during several years as their level of proficiency increases over time, hence a cross-sectional design was implemented, whereby samples from each learner are taken at different proficiency levels (beginner, intermediate and advanced).

3.3 Principle 3. Contrast

This principle states that comparisons within a corpus can be made only if the corpus has been designed to allow for such comparisons, (3).

(3) "Only those components of corpora which have been designed to be independently contrastive should be contrasted." (Sinclair 2005: 3).

As is clear from Section 2, most learner corpus designers include an equivalent native corpus for comparative purposes. This allows for the comparison between interlanguage grammars vs. native grammars, as is standard practice in L2 research. Obviously, a different question is whether it is legitimate to compare interlanguage grammars against an ideal native norm. This is a classic issue in SLA which is out of the scope of this paper. CEDEL2 contains a comparable Spanish native subcorpus. Such comparison is legitimate since the Spanish native subcorpus follows the same design principle and the same structural criteria as the learner subcorpus, e.g. both natives and learners must answer similar background

questionnaires, they all have the same composition topics to choose from, etc. Additionally, as we mentioned in Section 2.4, CEDEL2 also allows for *Contrastive Interlanguage Analysis* (CIA) (Granger 1996; Gilquin 2001), i.e. it is possible to compare intermediate learners vs. advanced learners.

3.4 Principle 4. Structural criteria

This principle states that the criteria constraining the structure of a corpus should be few and separable, as (4) states.

(4) "Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination." (Sinclair 2005: 5).

This principle is essential in the design of large native corpora and monitor corpora such as BNC, COCA, ICE, which contain several (or even hundreds of) million words coming from different genres (literature, science reports, newspapers, dialogues, etc.) both from spoken and written language. Since CEDEL2 is a written learner corpus, its structural criteria are pre-determined by the type of corpus (Sinclair 2005: 4). Following standard practice in L2 research, the most important structural criteria in CEDEL2 is the division into three learner subcorpora (based on proficiency level) and a comparable Spanish native corpus, as stipulated by principles 2 and 3. The simplicity in corpus design can be observed in Figure 1, which shows the structural criteria and the intended target (in number of words and percentage sample size).

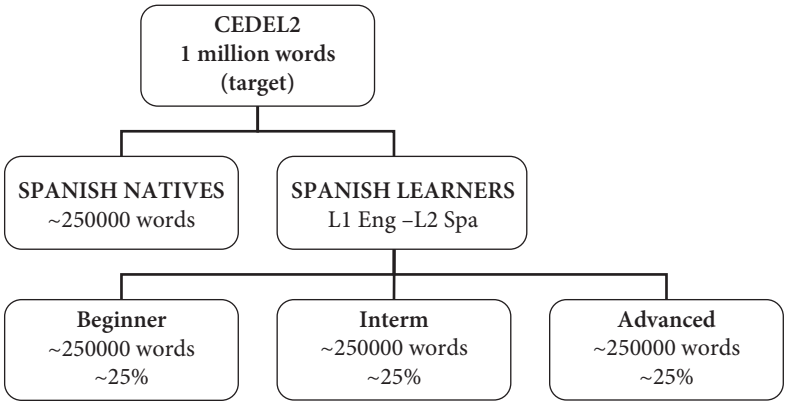


Figure 1. CEDEL2 corpus design

3.5 Principle 5. Annotation

This principle requires the raw text and the tags to be stored separately, (5).

(5) "Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications." (Sinclair 2005: 5).

As will be explained in more detail in Section 4.4, we are using the tagging and concordancing software *UAM CorpusTool* (O'Donnell 2009) which is designed to store the compositions written by learners (raw text format) separately from the tags (XML format). The XML file is the source file where the software runs the relevant commands. CEDEL2 thus meets the annotation principle, unlike other learner corpora in CLAN format used in the CHILDES database, where both raw text and tags are merged in the same file.

3.6 Principle 6. Sample size

This principle relates to the size of each text in the corpus, as stated in (6).

(6) "Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get to this target as possible. This means that samples will differ substantially in size." (Sinclair 2005: 7).

This is a crucial principle, since learner corpus researchers tend to think that each text in the corpus should be of equal length, which explains why these researchers impose a minimum-maximum word limit in the learner's composition. However, as Sinclair clearly states:

> There is no virtue from a linguistic point of view in selecting samples all of the same size. True, this was the convention in some of the early corpora, and it has been perpetuated in later corpora with a view to simplifying aspects of contrastive research. Apart from this very specialised consideration, it is difficult to justify the continuation of the practice. The integrity and representativeness of complete artifacts is far more important than the difficulty of reconciling texts of different dimensions. (Sinclair 2005: 6).

In other words, there is no linguistic justification that requires all texts to be of similar length in CEDEL2. What is crucial is for each sample to be a *complete* text, i.e. an unedited text. Following this principle, CEDEL2 contains only complete texts which vary in length. Such variability is a result of the learner's proficiency level, since some compositions are just one paragraph long (particularly those written by beginners, whose proficiency level is so low that they are unable to write

several paragraphs) to compositions containing up to several hundred words. The bottom line is that all texts in CEDEL2 are complete speech events, independent of their size.

3.7 Principle 7. Documentation

This principle states that both the design and composition of a corpus must be fully documented, (7).

(7) "The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken." (Sinclair 2005: 8).

Unlike other learner corpora, CEDEL2 contains detailed information about the structure of the corpus, as shown in the preceding sections, and, most importantly, it includes precise details about each learner (the *learning background* form and the *composition background* form, as will be explained in the data collection section, 4.1). This was done to avoid one of the typical pitfalls in corpus design, as Sinclair states:

> Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. Neither of these factors proves that there is something wrong with the corpus, because corpora are full of surprises, but they do cast doubt on the interpretation of the findings, and one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it. (Sinclair 2005: 8).

The precise information regarding each learner and each composition in CEDEL2 will allow the user to filter out or discard those texts that do not meet certain criteria or perhaps those that yield unexpected results. On closer inspection, the user may realize that those results are just an effect of any of the variables recorded in the learner's learning background profile.

3.8 Principle 8. Balance

Though the notion of balance is even vaguer than representativeness, corpus designers should strive for a well-balanced corpus, (8).

(8) "The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components." (Sinclair 2005: 9).

Sinclair is referring here to the fact that a well-balanced corpus must contain a fair and equally proportioned sample of both spoken and written language, since most of the early corpora in the 80s included more written than spoken language. While this equilibrium is desirable, Sinclair (2005: 9) notes the following: "Specialised corpora are constructed after some initial selectional criteria have been applied". Obviously, CEDEL2 is a specialized corpus which intends to be representative of written language only, hence it is well justified that it includes only written (and not spoken) language, though the corpus could be augmented in the future with spoken data that follow the same conditions and criteria as the written corpus. From an L2 psycholinguistic point of view, it may be argued that the written language of learners is more amenable to monitoring and controlled processing than their spoken language (see an overview of this classic SLA debate in general reference works such as Ellis 2008). Independently of whether the fact that learners have arguably higher opportunities for self-correction or self-repairs in written language, it has been undisputedly accepted over the past 40 years that their internalised linguistic knowledge (interlanguage) is systematic, whether such knowledge is produced in writing or in speaking. In other words, written language is as reliable as spoken language to study interlanguage phenomena, as shown by the numerous publications that have used written learner corpora such as ICLE (e.g. Granger et al. 2002; Aston et al. 2004; Hidalgo et al. 2007; Gilquin et al. 2008). The only difference between written vs. spoken texts in the study of interlanguage lies in the proportion or percentage of the observed phenomenon: some linguistic phenomena show a higher frequency in spoken than written language, and vice versa, but the phenomenon is undeniably still there. Thus, there is no principled reason to believe that written language is less reliable than spoken language in the investigation of interlanguage grammars.

3.9 Principle 9. Topic

This design principle relates to the subject matter in a corpus, (9).

(9) "Any control of the subject matter [i.e. topic] in a corpus should be imposed by the use of external, and not internal, criteria." (Sinclair 2005: 10).

This principle has been dealt with in our discussion of principles 1 and 2. As stated above, CEDEL2 was designed following external criteria and no control was exerted over vocabulary, linguistic structures or even topic, since "it seems strange to many people that it is essential that the vocabulary should not be directly controlled. But vocabulary choice is clearly an internal criterion." (Sinclair 2005: 9). The composition titles learners can choose from are varied enough to elicit a wide array of linguistic structures and lexical items which intend to fairly represent the learners' interlanguage.

3.10 Principle 10. Homogeneity

This principle calls for the homogeneity of texts in the corpus (which is of particular relevance in large normative corpora).

(10) "A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided." (Sinclair 2005: 14).

By *rogue* text he refers to odd or unusual texts "which stand out as radically different from the others in their putative category, and therefore [are] unrepresentative of the variety on intuitive grounds." (Sinclair 2005: 13). While it may seem that the avoidance of rogue texts in the name of homogeneity is an internal criterion, Sinclair argues that "The use of homogeneity as a criterion for acceptance of a text into a corpus is based certainly on the impression given by some features of its language, but is a long way from the use of internal criteria" (Sinclair 2005: 14). That is, certain texts may in principle seem to belong to a given category but, on closer inspection, may not meet the desired design criteria for the corpus. In this respect, rogue texts are avoided in CEDEL2 as they are received in the online application. Texts that do not satisfy the design criteria are discarded, e.g. compositions that have been clearly corrected in class previously and, therefore, do not represent naturalistic learner language; compositions belonging to learners whose mother tongue is other than English; compositions that are written mostly in English, and not in Spanish; compositions that are too short (just a few words) or that contain repeated structures via *copy and paste* mechanisms; compositions whose language does not clearly match the level of the learner, which probably means that the text has been taken off the internet; etc. Additionally, once the corpus data collection is finalized, researchers will examine each text to double-check that the structural criteria are met.

3.11 Conclusion

Given the ten aforementioned design criteria, we can safely claim that CEDEL2 is a well-designed leaner corpus that does not need extra design principles apart from those stated, though some adaptation is required for learner corpora. As Sinclair pointed out:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. (Sinclair 2005: 16)

Following this definition of a corpus, we are now in a position to define a *learner* corpus. Granger defines it as:

> [E]lectronic collections of authentic FL/SL textual data according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and are documented as to their origin of provenance. (Granger 2002: 7)

It is perhaps the question of what constitutes *authentic* foreign language/second language data that requires some explanation. As Granger (2002: 8) points out, learner data is rarely fully natural, especially in the case of EFL learners, who tend to learn English in a classroom. There is scale of naturalness: fully natural – product of teaching process – controlled task – scripted (Nesselhauf 2005: 128). As mentioned above, the kind of texts compiled in CEDEL2 are texts voluntarily written by learners of L2 Spanish (L1 English) and collected online (via the internet). These texts come mostly from learners who have acquired L2 Spanish in a formal/classroom setting and there is no restriction on the language/topic/content they have produced. Additionally, as stated in the learning background and composition background forms, most of the texts have been written outside the classroom. Such texts constitute authentic (written) learner data ("data resulting from authentic classroom activity", Granger 2002: 8) for the reasons detailed in the 10 design criteria (particularly, principles 1, 2, 3, 4, 8, 9, 10).

4. Current status of CEDEL2

In this section we discuss the data collection process (4.1), the amount and the distribution of the data collected so far (Section 4.2), the source and nature of the data (Section 4.3) and some preliminary tagging (4.4).

4.1 Data collection

CEDEL2 data collecting started in 2006. Data are still being collected online via electronic forms available at the WOSLAC research group webpage (see Footnote 7). Each learner must fill in three forms: (i) a *learning background*, (ii) a *placement test* and (iii) a *composition in Spanish*. At the outset of the first form learners give their consent to participate voluntarily in CEDEL2. They are informed that their data will be used only for research purposes and will be treated confidentially (learners only indicate their initials, never their full names).

In relation to the principle of documentation of corpus design (no. 7, see Section 3.7 above), learners provide detailed information about their learning background and about the composition, as about to be explained below. These details are essential for fine-grained, quantitative analyses, and in cases when the

researcher finds odd and counter-intuitive results, the different variables provided by the learner can shed some light. In particular, the details refer to two forms:

Form 1: Learning background form (see Appendix 1). Each learner provides detailed information about their L2 Spanish learning experience and Spanish native speakers fill in a similar form (a *Formación académica* form), containing:

1. *Personal details*: age, sex and information regarding the institution where they are learning Spanish, if applicable –name of the institution (school or university), course and type of studies being pursued.
2. *Linguistic details*: mother tongue, father's L1, mother's L1, language spoken at home, age of first immersion in L2 Spanish and length of stay in Spanish-speaking countries, if applicable.
3. *Self-proficiency level*: students provide their own self-rating of proficiency in each of the four skills in L2 Spanish and in other languages s/he has learnt. Note that this self-rating is not the only proficiency measure, as learners have to fill a standardized placement test (University of Wisconsin 1996), as justified above (see Appendix 4).[10]

Form 2: Composition in Spanish form (see Appendices 2 and 3). Learners and natives provide here the raw linguistic data (the composition itself), plus additional information regarding the context in which the composition was produced, namely:

1. *Background research:* learners are asked whether they have conducted any research prior to the writing of the composition; if so, they need to specify how long it took them to do the research and which instruments were employed: internet, newspapers, TV, etc.
2. *Composition title*: students can choose from a range of 12 composition titles, graded according to complexity (see discussion of Principle 2 above and also Appendix 2 for a full list of composition titles).
3. *Writing location:* learners are asked whether the composition was written in class, at home or both.
4. *Writing tools:* learners specify which linguistic tools they have used when writing the composition: bilingual/monolingual dictionaries, spellcheckers, native help, etc.

4.2 Data distribution

As shown in the time series in Figure 2, data collection started in February 2006. Around 750,000 words have been collected to date (March 2011). We can appreciate

10. *UAM CorpusTool* can be freely downloaded at <http://www.wagsoft.com/CorpusTool>

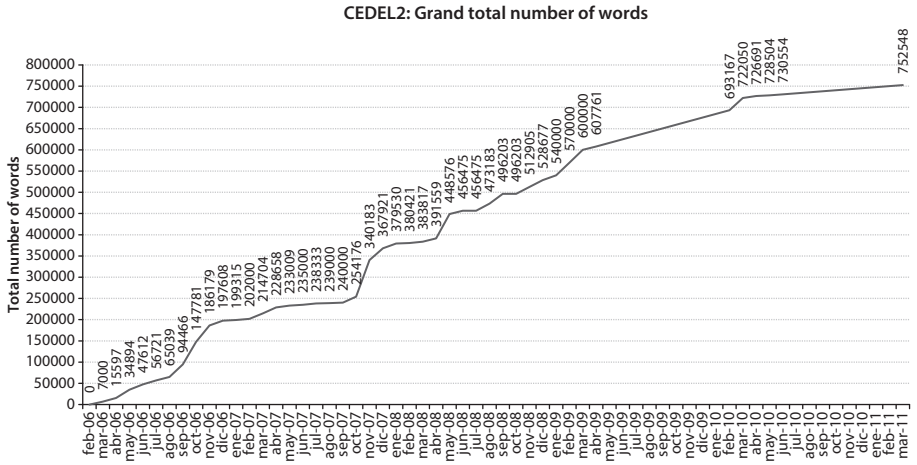


Figure 2. Evolution of CEDEL2 according to total no. of words

a mild rising trend with sporadic quick and high rises. As can be appreciated in Appendix 7, these increases are caused by a high number of learners participating in a short period as a result of (i) a call for participation published in distribution lists like the *Linguist List* or (ii) the start of the academic year. As stated earlier, the target is to reach one million words.

As described above, CEDEL2 consists of a learner subcorpus and a comparable Spanish native corpus. As shown in Figure 3, approximately ¼ of the total number of words belongs to the native subcorpus (200,326 words representing 27% of the corpus data), while the rest (552,401 words, 73%) belongs to the learner subcorpus. Assuming this proportion to remain constant until the end of the data collection (as has been the case throughout the data collection process), when

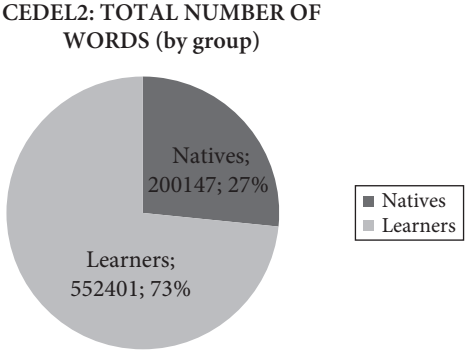


Figure 3. Proportion of words (learners vs. natives) in CEDEL2

the one million target is reached, the native subcorpus will be expected to contain c. 250,000 words, which represent an acceptable sample size for a native subcorpus (cf. LOCNESS, the English native subcorpus in ICLE version 1, which contains 235,000 words). Similarly, the learner subcorpus will eventually contain c. 750,000 words, a reasonable sample size for a learner corpus. Note that, while the grand total of the different ICLE subcorpora (version 1) reaches around 2.5 million words (Granger et al. 2002a), the Spanish subcorpus contains just over 200,000 words. Obviously, the art of sampling has a bearing on the extrapolability of the results: the larger the sample, the more reliable the findings.

While it is important to know the percentage of each corpus regarding the number of words (tokens), it is also relevant to know the percentage of participants, so as to have a rough estimation of whether each subcorpus is balanced regarding text size, i.e. the mean amount of words contributed by each participant (though, as argued in Principle 6, what is relevant here is the fact that each text be a complete artifact, independently of its size). The 711 natives who have participated represent 29% of the volume of participants (see Figure 4) and the 200,326 of words they produced represent 27% (cf. Figure 3). A similar proportion can be observed in the learner subcorpus (1,729 participants representing 71% of the total volume of participants, who contributed 552,401 words representing 73% of the total volume of words). These figures reveal that both subcorpora are balanced regarding the number of participants and the number of words they have contributed.

4.3 Source of data

As stated in Section 4.1, CEDEL2 data are being collected online via a web application (see Footnote 7). We have received data coming from volunteers all over the world. In the native subcorpus most participants have received a university

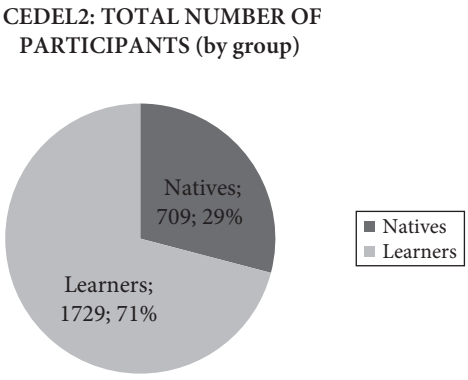


Figure 4. Proportion of participants (learners vs. natives) in CEDEL2

education. Most of them are speakers of peninsular Spanish, though we can find many participants with other varieties of Central and South American Spanish, as well as a few native Spanish speakers residing in USA.

Figure 5 shows the source of data in the learner subcorpus. The majority of data come from learners of L2 Spanish in several universities and secondary schools in USA: 1331 participants representing 77% of the total number of learners (see Appendix 5 for further details). This is followed by university students of L2 Spanish in the UK (N = 109, 6%), and by North American university students of Spanish during their stay abroad in Spain, as well as a few English native speakers residing in Spain (N = 80, 5%). A small percentage of data come from learners of Spanish in other countries (New Zealand, Australia and Canada, see Appendix 6). The remaining percentage either comes from other countries or the origin is not specified in the online form (N = 173, 10%).

Obviously, this learner background information (coupled with information on other learning variables such as institution where the learning is taking place, course level, type of studies, plus all the other linguistic background variables described earlier) provide crucial quantitative information for the researcher.

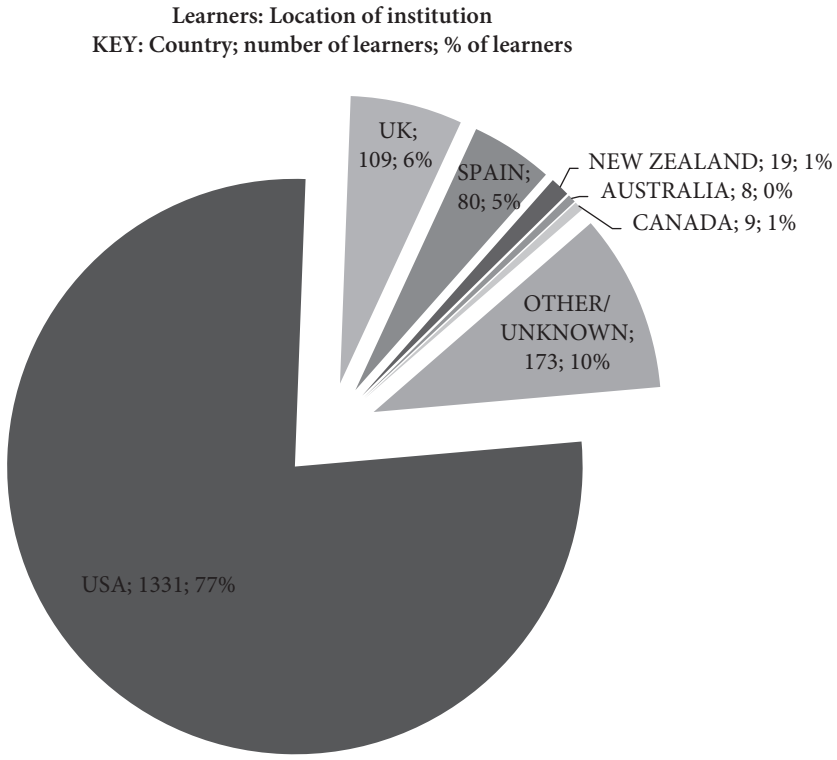


Figure 5. Source of data in the CEDEL2 learner subcorpus

4.4 Preliminary segmentation and annotation

While CEDEL2 is not fully tagged yet, some samples have been preliminarily tagged (see published work in Lozano 2009b; Alonso-Ramos 2010a, 2010b). We are using the tagging and concordancing software *UAM CorpusTool* (O'Donnell 2009), which is freely available.[11] This tool allows the user to annotate texts in different ways. In particular, the tagging process consists in selecting a segment (e.g. word, morpheme, phrase, sentence or paragraph) and assign tags to it. The tags are previously defined in the software by the user according to a scheme that can be easily designed according to the user's needs. Apart from being a tagger, *UAM CorpusTool* is also a concordancer that permits conducting descriptive and inferential statistical analyses on the corpus data. To illustrate, see the scheme in Figure 6 where pronominal subjects were annotated according to several tags designed by the researcher and implemented in the software: syntax (NP/pronoun/null), number and person (singular 1 2 3, plural 1 2 3), animacy (animate/inanimate), etc.

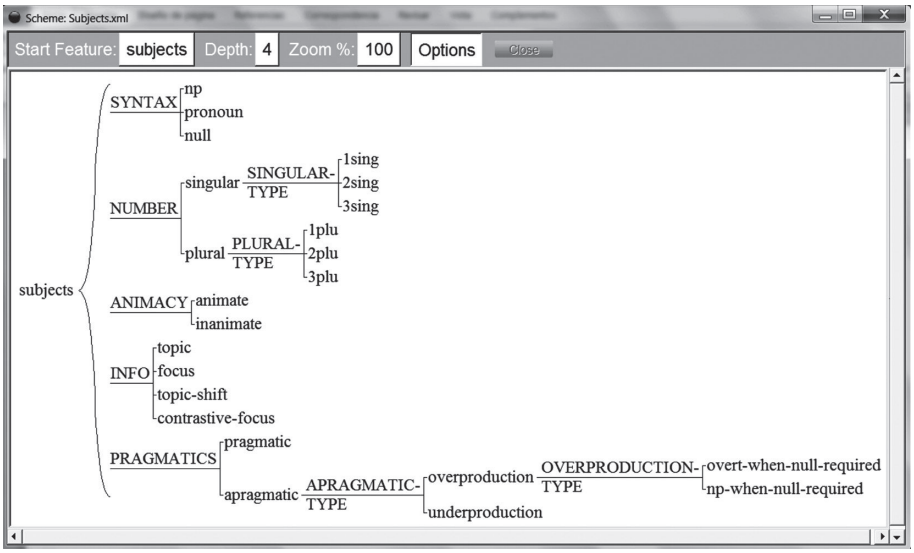


Figure 6. Scheme of tags created with UAM CorpusTools (Lozano 2009b)

11. A reviewer observes that, since CEDEL2 data are being collected online, learners may not be telling the truth about themselves in the learner/composition profiles. But note that, after data collection, each corpus text will have to be inspected manually to ensure that the learner's profile (particularly the self-rated proficiency level) agrees with the real (placement test) level. Other measures will be taken, as explained at the end of Section 3.10 when discussing rogue texts.

4.5 CEDEL2: Next steps

The next research steps for CEDEL2 are (i) to complete data collection and approach the intended target of 1 million words; (ii) to launch an online beta version of CEDEL2; (iii) to continue the tagging of the corpus with particular reference to interlanguage phenomena at the syntax-discourse interface (though future researchers will be able to tag any linguistic phenomena they wish); (iv) to make freely available the full version of the corpus via a dedicated webpage, where the full corpus will be available in plain text format and also in a tagged format.

5. Learner corpora: The way forward

The review of learner corpora and learner corpus research presented in this paper suggests that if corpus-based research is going to make a significant contribution to the field of SLA, new, well-designed corpora need to be made available to the research community. It has been also argued that there is a need for:

1. corpora of L2s other than L2 English
2. corpora of spoken language
3. longitudinal corpora (to address the developmental dimension of L2 learning), and
4. cross-sectional corpora, with learners at different levels of proficiency.

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* or *ad hoc* corpora and that they are based upon formal measurements of proficiency. A further requirement is that they 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. This paper has (i) furthered work for points 1 and 4 and (ii) focused on these requirements and criteria for the creation of CEDEL2 (Lozano 2009a), a corpus compiled for and by L2 researchers (see also Rollinson & Mendikoetxea 2010 for WriCLE).

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, and methodologies have to be developed to combine corpus data with experimental data in the search for converging evidence and to test aspects which cannot be adequately tested with corpus data (see Gilquin & Gries 2009). 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),

but this will only be possible if corpus design and methodologies are useful for SLA purposes.

6. Conclusion

This paper has addressed the need for well-constructed large-scale learner corpora in SLA research. For learner corpora to be useful for L2 researchers and practitioners certain design principles have to be followed. We have illustrated this by focusing on the main design principles of CEDEL2 (*Corpus Escrito del Español como L2*) (Lozano 2009a) is a large L1 English – L2 Spanish written corpus. The corpus already consists of 750,000 words coming from over 2,500 participants (both learners of Spanish and Spanish native speakers for comparative purposes). Unlike other learner corpora, it has been designed according to ten standard corpus design principles, so it is hoped that it can be beneficial to users of L2 Spanish (researchers, practitioners and students alike) as a reliable source of naturalistic data.

References

Aijmer, K. 2002. Modality in advanced Swedish learners' written interlanguage. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 55–76. Amsterdam: John Benjamins.

Aijmer, K. 2009. *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Alonso Ramos, M., Wanner, L., Vázquez Veiga, N., Vincze, O., Mosqueira Suárez, E. & Prieto González, S. 2010a. Tagging collocations for learners. In *eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of ELEX2009, Cahiers du CENTAL 7*, S. Granger & M. Paquot (eds), 375–380. Louvain-la-Neuve: Presses universitaires de Louvain.

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor del Bosque, G., Vázquez Veiga, N., Mosqueira Suárez, E. & Prieto González, S. 2010b. 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)*, N. Calzolari (ed.), 3209–3214. Valetta: Language Resources Evaluation.

Altenberg, B. 2002. Using bilingual corpus evidence in learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 37–54. Amsterdam: John Benjamins.

Asención-Delaney, Y. & Collentine, J. 2011. A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics* 32(3): 299–322.

Aston, G., Bernardini, S. & Stewart, D. 2004. *Corpora and Language Learners*. Amsterdam: John Benjamins.

Baker, P., Hardi, A. & McEnery, T. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Barlow, M. 2005. Computer-based analysis of learner language. In *Analysing Learner Language*, R. Ellis & G.P. Barkhuizen, 335–357. Oxford: Oxford University Press.

Biber, D., Conrad, S. & Reppen, R. (eds). 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brown, J.D. & Rodgers, T.S. 2002. *Doing Second Language Research*. Oxford: Oxford University Press.

Burnard, L. & McEnery, T. (eds). 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.

Cestero Mancera, A., Penadés Martínez, I., Blanco Canales, A., Camargo Fernández, L. & Simón Granda, J. 2001. Corpus para el análisis de errores de aprendices de E/LE *(CORANE).* In A.M. Gimeno Sanz (ed.), *Actas de ASELE XII.* http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/asele_xii.htm [Accessed 1.8.2012]

Coffin, C., Hewings, A. & O'Halloran, K. (eds). 2004. *Applying English Grammar: Corpus and Functional Approaches*. London: Hodder Education.

Connor, U. & Upton, T.A. (eds). 2004. *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi.

Chaudron, J. 2003. Data collection in SLA research. In *The Handbook of Second Language Acquisition*, C. Doughty & M. Long (eds), 717–761. Oxford: Blackwell.

CHILDES. 2010. *CHILDES – Child Language Data Exchange System*. http://childes.psy.cmu.edu/ [Accessed 1.8.2012]

Chocano, G., Jiménez, R., Lozano, C., Mendikoetxea, A., Murcia, S., O'Donnell, M., Rollinson, P. & Teomiro, I. 2007. An exploration of word order in learner corpora: The WOSLAC project. In *Proceedings of Corpus Linguistics 2007*, M. Davies, P. Rayson, S. Hunston & P. Danielsson (eds), article #113. http://ucrel.lancs.ac.uk/publications/CL2007/ [Accessed 1.8.2012]

Davies, M. 2010. corpus.byu.edu. http://corpus.byu.edu/. [Accessed 1.8.2012]

Dörnyei, Z. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Ellis, R. 2008. *The Study of Second Language Acquisition. (2nd edition).* Oxford: Oxford University Press.

Fitzpatrick, E. (ed.). 2007. *Corpus Linguistics Beyond the Word*. Amsterdam: Rodopi.

Gilquin, G. 2001. The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3(1): 95–125.

Gilquin, G. & Gries, S. T. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1–26.

Gilquin, G., Papp, S. & Díez-Bedmar, M.B. (eds). 2008. *Linking up Contrastive and Learner Corpus Research.* Amsterdam: Rodopi.

Granger, S. 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*, K. Aijmer, B. Altenberg & M. Johansson (eds), 37–51. Lund: Lund University Press.

Granger, S. 1998. The computerized learner corpus: A versatile new source of data for SLA research. In *Learning English on Computer*, S. Granger (ed.), 3–14. London: Addison Wesley Longman.

Granger, S. 2002. A bird's eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 3–33. Amsterdam: John Benjamins.

Granger, S. 2004. Computer learner corpus research: Current status and future prospects. In *Applied corpus Linguistics: A Multidimensional Perspective*, U. Connor & T.A. Upton (eds), 123–145. Amsterdam: Rodopi.

Granger, S. 2008. Learner corpora. In *Corpus Linguistics: An International Handbook*, A. Lüdeling & M. Kytö (eds), 259–275. Berlin: Mouton de Gruyter.

Granger, S., Dagneaux, E. & Meunier, F. (eds). 2002a. *International Corpus of Learner English. Handbook and CD-ROM. Version 1.1*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., Hung, J. & Petch-Tyson, S. (eds). 2002b. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain.

Hawkins, R. 2001. *Second Language Syntax: A Generative Introduction*. Oxford: Blackwell.

Hawkins, R. 2007. The nativist perspective on second language acquisition. *Lingua* 118(4): 465–477.

Hasselgard, H. 1999. Review of S. Granger (ed.) 'Learner English on Computer'. *ICAME Journal* 23: 148–152.

Hidalgo, E., Quereda, L. & Santana, J. (eds). 2007. *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.

Housen, A. 2002. A corpus-based study of the L2 acquisition of the English verb system. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 77–118. Amsterdam: John Benjamins.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & Francis, G. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Lafford, B.A. & Salaberry, R. (eds). 2003. *Spanish Second Language Acquisition: State of the Science*. Washington DC: Georgetown University Press.

Lardiere, D. 1998. Dissociating syntax from morphology in a divergent L2 end-state grammar. *Second Language Research* 14(4): 359–375.

Leech, G., Rayson, P. & Wilson, A. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Pearson Education Limited.

Liceras, J. M., Zobl, H. & Goodluck, H. (eds). 2008. *The Role of Formal Features in Second Language Acquisition.* Mahwaw NJ: Lawrence Erlbaum Associates.

Lozano, C. 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*, C.M. Bretones et al. (eds), 197–212. Almería: Universidad de Almería.

Lozano, C. 2009b. Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In *Representational Deficits in Second Language Acquisition*, Y. Leung, N. Snape & M. Sharwood-Smith (eds), 127–166. Amsterdam: John Benjamins.

Lozano, C. & Mendikoetxea, A. 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*, G. Gilquin, S. Papp & M.B. Díez-Bedmar (eds), 85–125. Amsterdam: Rodopi.

Lozano, C. & Mendikoetxea, A. 2010. Postverbal subjects in L2 English: A corpus-based study. *Bilingualism: Language and Cognition* 13(4): 475–497.

Lüdeling, A. & Kytö, M. (eds). 2008. *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter.

Mackey, A. & Gass, S. M. 2005. *Second Language Research: Methodology and Design*. Mahwah NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analysing Language. (3rd edition).* Mahwah NJ: Lawrence Erlbaum Associates. http://childes.psy.cmu.edu [Accessed 1.8.2012]

McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Mitchell, R., Domínguez, L., Arche, M., Myles, F. & Marsden, E. 2008. SPLLOC: A new corpus for Spanish second language acquisition research. In *EUROSLA Yearbook 8,* L. Roberts, F. Myles & A. David (eds), 287–304. Amsterdam: John Benjamins.

Mitchell, R. & Myles, F. 2004. *Second Language Learning Theories*. *(2nd edition)*. London: Hodder Arnold.

Montrul, S. 2004. *The Acquisition of Spanish: Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and Adult L2 Acquisition*. Amsterdam: John Benjamins.

Myles, F. 2005. Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4): 373–391.

Myles, F. 2007a. Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In *The Longitudinal Study of Advanced L2 Capacities*, L. Ortega & H. Byrnes (eds), 58–72. Hillsdale NJ: Lawrence Erlbaum.

Myles, F. 2007b. Using electronic corpora in SLA research. In *Handbook of French Applied Linguistics*, D. Ayoun (ed.), 377–400. Amsterdam: John Benjamins.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

O'Donnell, M. 2009. The UAM CorpusTool: Software for corpus annotation and exploration. In *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente*, C.M. Bretones & et al. (eds), 1433–1447. Almería: Universidad de Almería.

Oshita, H. 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, H. 2004. Is there anything there when there is not there? Null expletives and second language data. *Second Language Research* 20(2): 95–130.

Perdue, C. 1993. *Adult Language Acquisition. Vol. 1: Field Methods*. Cambridge: Cambridge University Press.

Pérez-Leroux, A.T. & Liceras, J. (eds). 2002. *The Acquisition of Spanish Morphosyntax: The L1/L2 Connection*. Dordrecht: Kluwer Academic Press.

Real Academia Española. 2010a. *Corpus de Referencia del Español Actual (CREA)*. http://www.rae.es [Accessed 1.8.2012]

Real Academia Española. 2010b. *Corpus Diacrónico del Español (CORDE)*. http://www.rae.es [Accessed 1.8.2012]

Renouf, A. & Kehoe, A. (eds). 2006. *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.

Reppen, R. 2006. Corpus studies: Second language. In *Encyclopedia of Language & Linguistics*, K. Brown (ed.), 248–250. Oxford: Elsevier.

Reppen, R. & Simpson, R. 2002. Corpus linguistics. In *An Introduction to Applied Linguistics*, N. Schmitt (ed.), 93–111. London: Arnold.

Rollinson, P. & Mendikoetxea, A. 2010. Learner corpora and second language acquisition: Introducing WriCLE. In *Analizar Datos > Describir Variación/Analysing Data > Describing Variation*, J.L. Bueno Alonso, D. Gonzáliz Álvarez, U. Kirsten Torrado, A.E. Martínez Insua, J. Pérez-Guerra, E. Rama Martínez & R. Rodríguez Vazquez (eds), 1–12. Vigo: Universidade de Vigo (Servizo de Publicacións).

Rutherford, W. & Thomas, M. 2001. The Child Language Data Exchange System in research on Second Language Acquisition. *Second Language Research* 17(2): 195–212.

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

Sinclair, J. 2005. How to build a corpus. In *Developing Linguistic Corpora: A Guide to Good Practice*, M. Wynne (ed.), 79–83. Oxford: Oxbow books.

Slabakova, R., Montrul, S.A. & Prévost, P. (eds). 2006. *Inquiries in Linguistic Development: In Honor of Lydia White*. Amsterdam: John Benjamins.

Sorace, A. 2000. Syntactic optionality in non-native grammars. *Second Language Research* 16(2): 93–102.

Sorace, A. 2004. Native language attrition and developmental instability at the syntax-discourse interface: Data, interpretations and methods. *Bilingualism: Language and Cognition* 7(2): 143–145.

Sorace, A. 2005. Selective optionality in language development. In *Syntax and variation: Reconciling the Biological and the Social*, L. Cornips & K.P. Corrigan (eds), 55–80. Amsterdam: John Benjamins.

Sorace, A. & Serratrice, L. 2009. Internal and external interfaces in bilingual language development: Beyond structural overlap. *International Journal of Bilingualism* 13(2): 195–210.

Tono, Y. 2003. Learner corpora: Design, development and applications. In *Proceedings of the 2003 Corpus Linguistics Conference*, D. Archer, P. Rayson, A. Wilson & T. McEnery (eds), 800–809. UCREL: Lancaster University.

Tono, Y. 2004. Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In *Corpora and Language Learners*, G. Aston, S. Bernardini & D. Stewart (eds), 45–66. Amsterdam: John Benjamins.

Tono, Y. 2005. Computer-based SLA research: State of the art of learner corpus studies. In *Studies in Language Sciences (4): Papers from the Fourth Annual Conference of the Japanese Society for Language Sciences*, M. Minami, H. Kobayashi, M. Nakayama & H. Sirai (eds), 45–77. Tokyo: Kurosio Publishers.

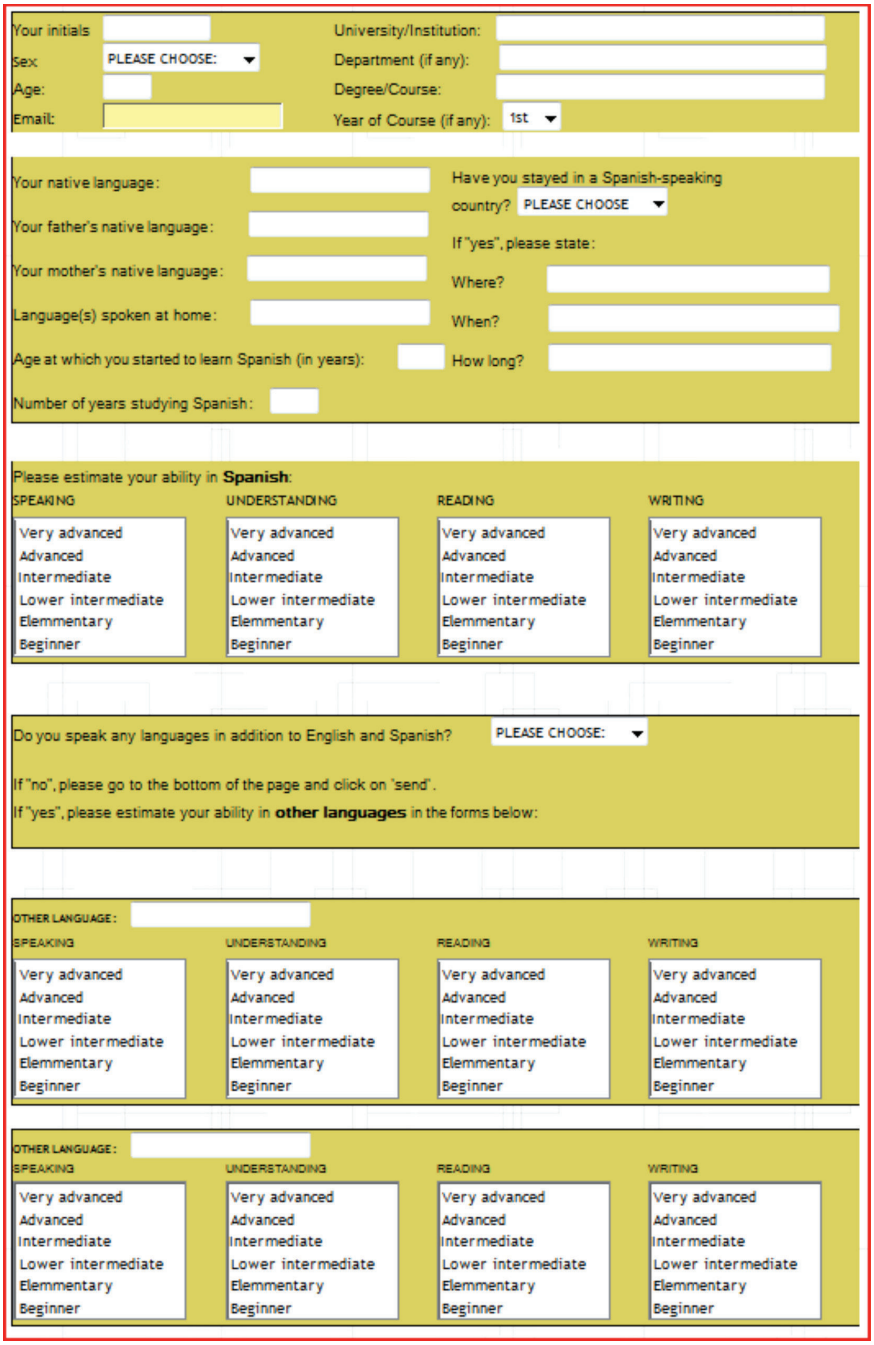
University of Wisconsin. 1998. *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. Madison WI: University of Wisconsin Press.

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

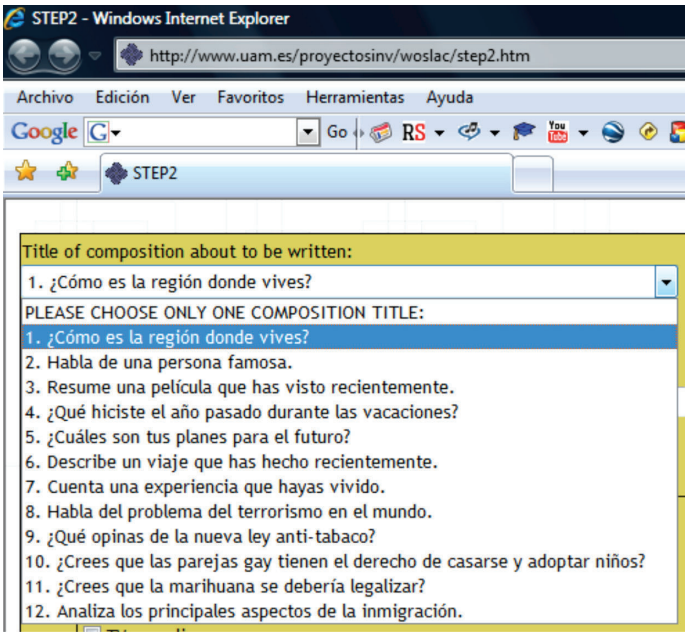
White, L. 2009. Grammatical theory: Interfaces and L2 knowledge. In *The New Handbook of Second Language Acquisition*, W.C. Ritchie & T.K. Bhatia (eds), 49–68. Bingley: Emerald.

Wynne, M. (ed.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books.

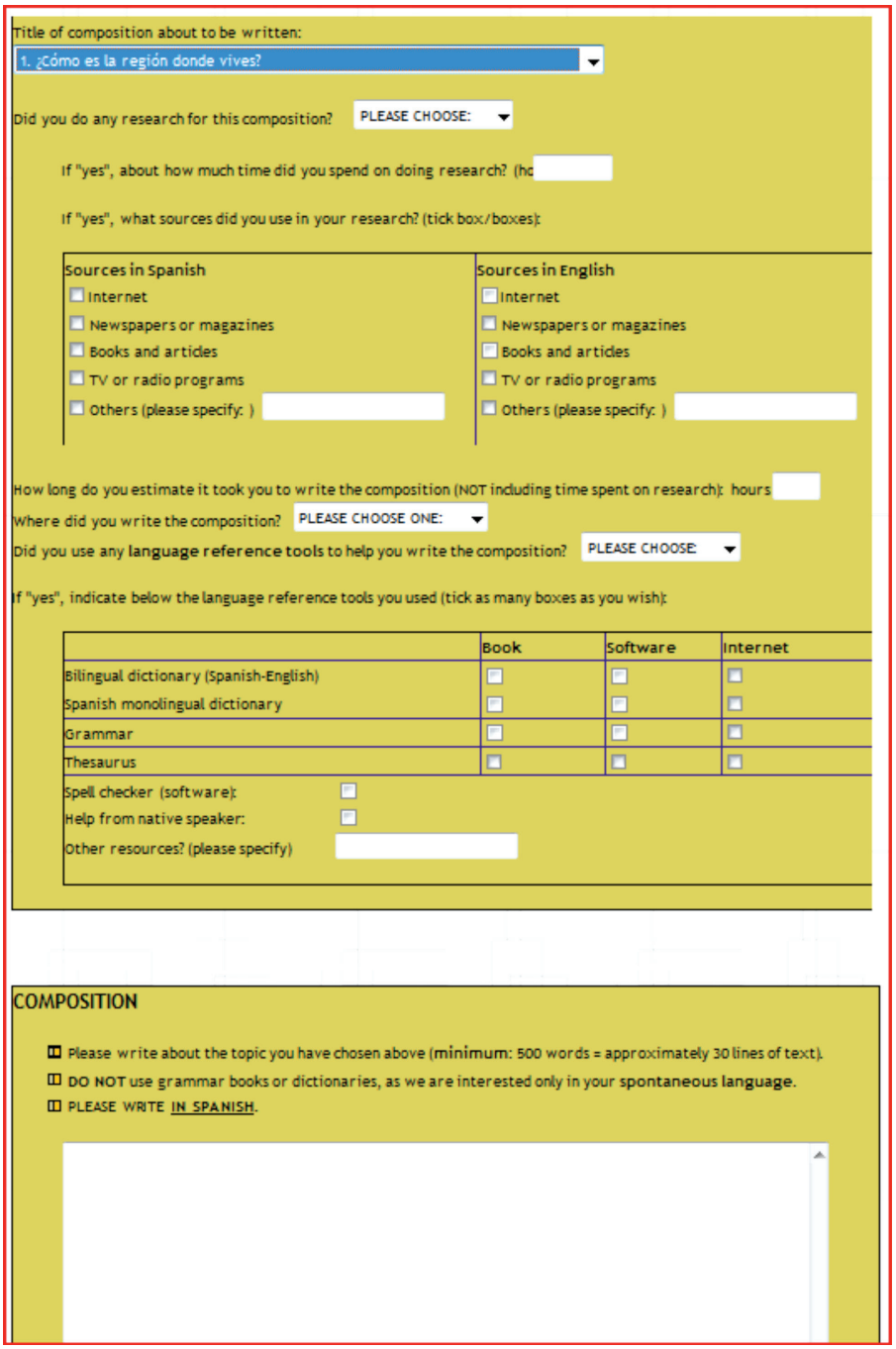
Appendices



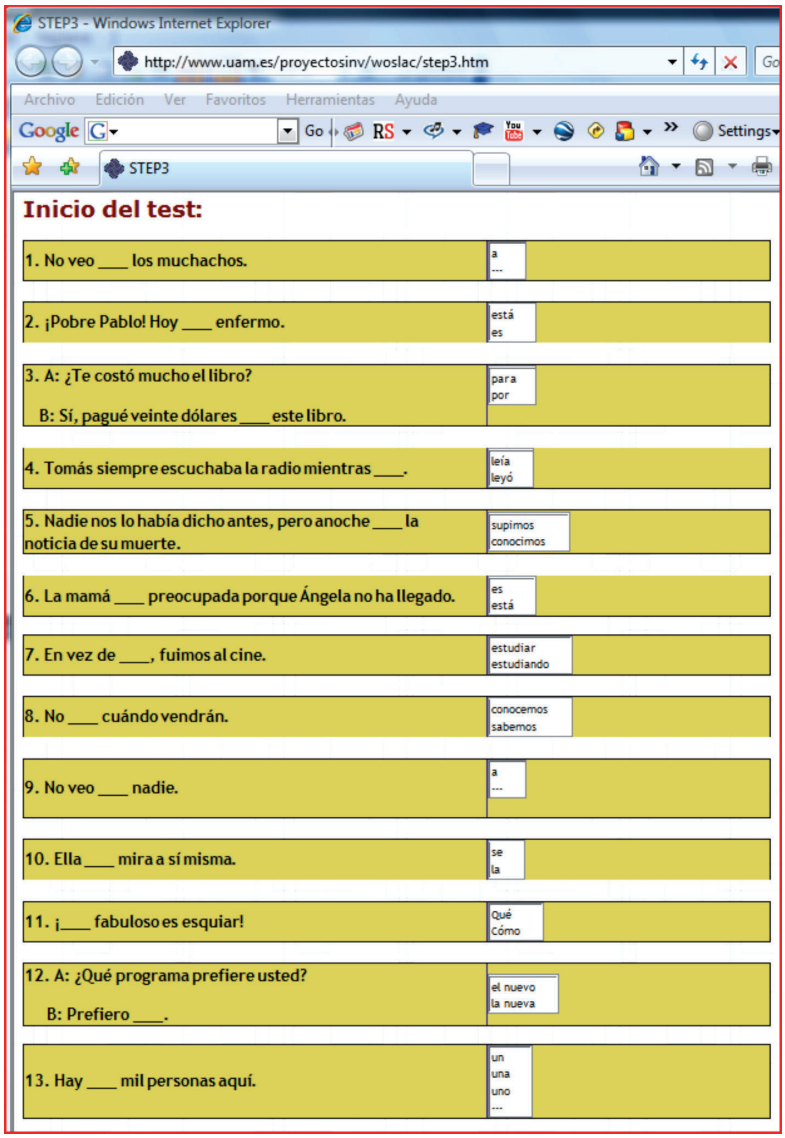
Appendix 1. Learning background form in CEDEL2



Appendix 2. Composition titles in CEDEL2



Appendix 3. Composition form in CEDEL2



Appendix 4. Sample questions from the Spanish placement test form (University of Wisconsin 1998)

Appendix 5. Source of data in CEDEL2 (USA, UK and Spain)

USA	N	UK	N	SPAIN	N
Georgia State University	409	Open University	20	CCCS (Centre for Cross Cultural Study Seville)	27
		Queen Mary	11		
University of Florida	202	University of London		*Universidad de Cantabria*	25
Pennsylvania State University	78	King's College, University of London	8	*Escuela oficial de idiomas Madrid*	7
John F. Kennedy High School	65	Essex University	7	Middlebury College	5
		St.Paul's School	5	Other universities	16
Central Catholic High School	53	University of Leeds	5		
		Other universities	53		
Syracuse University	48				
Franklin High School	48				
Saint Louis University–Madrid Campus	31				
University of Illinois	30				
Southern Methodist University	26				
Bob Jones University	19				
Illinois Wesleyan University	16				
Messiah College, Pennsylvania	15				
Zionsville Community High School	12				
Grand Valley High School	11				
Other universities/schools	269				
Total USA	**1332**	**Total UK**	**109**	**Total Spain**	**80**

Appendix 6. Source of data in CEDEL2 (New Zealand, Australia Canada and other countries)

Other countries	N	Unknown source	N
New Zealand	19	[These are learners who did not specify their University/School]	
Australia	8		
Canada	9		
Other countries	10		
Total	46	Total	163

Appendix 7. Calls for participation in CEDEL2

Date and distribution list	Date and distribution list (cont'd)
– May 2006 Portal del Hispanismo (Instituto Cervantes) – May 2006 TodoELE.net – May 2006 INFOLING – May 2006 AEDEAN (Asociación Española de Estudios Anglo-Norteamericanos) – May 2006 WordPress.com – May 2006 Centro Virtual Cervantes (Tablón del foro didáctico) – May 2006 FORMESPA – June 2006 OESI (Oficina de Español en la Sociedad de la Información, Instituto Cervantes) – June 2006 Corpora List – June 2006 Linguist List – June 2006 Corpus4you [Japanese webpage] – June 2006 AltaTECH – June 2006 International Speech – Communication Association (ISCA) – Oct 2006 AESLA – Oct 2006 Linguist List – Oct 2006 Infoling – Oct 2006 DeEstranjis blogspot – Oct 2006 FORMESPA – Mar 2007 AEDEAN – Mar 2007 AESLA	– Mar 2007 Corpora List – Mar 2007 Linguist List – Apr 2007 Infoling – May 2007 Democratic Underground.com – May 2007 ELE.inicios.es – Oct 2007 Linguist List – Oct 2007 FORMESPA – Oct 2007 AESLA – Oct 2007 AEDEAN – Oct 2007 Infoling – Oct 2007 Corpora List – Nov 2007 AATPS (American Association of Teachers of Spanish and Portuguese) – March 2008 Linguist List – May 2008 Linguist List – Sept 2008 Linguist List – Nov 2008 Linguist List – Feb 2010 AATSP (American Association of Teachers of Spanish and Portuguese) – Feb 2010 INFOLING – Feb 2010 Linguist List – Feb 2010 Comunidad TodoELE – Feb 2010 Corpora List – June 2010 Linguist List – Feb 2011 Linguist List