LEARNER SPANISH ON COMPUTER. THE CAES 'CORPUS DE APRENDICES DE ESPAÑOL' PROJECT[1]

Guillermo Rojo

Ignacio M. Palacios

University of Santiago de Compostela

Abstract

This chapter aims to provide a general description of the 'Corpus de Aprendices de Español' (CAES). It starts by discussing the contributions of Corpus Linguistics to the study of language and it next explains the emergence of learner corpora by providing a general survey of the existing learner corpora for Spanish. The next section is concerned with the origin and development of the CAES corpus from the beginning to its current state and explains in detail the general design, compilation method, text coding and annotation plus the search tool accompanying the corpus. The last part presents the results of a study on false friends based on data extracted from the corpus. Several conclusions and reflections follow together with some suggestions for further research.

[1] The title selected for this work clearly resembles the first important book on English learner corpora, *Learner English on Computer* (Granger 1998) where a full account is provided of ICLE (*The International Corpus of Learner English*), a corpus which has been used as a model or reference for subsequent projects in learner corpus research. We believe the CAES project, as a computerised Spanish learner corpus, shares many of the general principles of corpus design and compilation discussed in this book and illustrated in ICLE.

We would like to express our gratitude to the Cervantes Institute for their financial support in the development of this project. We also want to express our thanks to all the members of the research team who made this corpus possible, as well as to all the learners and teachers from the different Cervantes Institutes and participating universities who directly or indirectly collaborated and helped us in its compilation.

Introduction

This chapter, organised into three main parts, aims to provide a general description of the CAES learner corpus, along with one main study that uses data extracted from it. Findings will be analysed and the pedagogical implications of this considered.

Part 1 includes a brief discussion of the contribution of Corpus Linguistics (CL) to the study of language, specifically in second language acquisition (SLA) research. Attention will be paid to the emergence of learner corpora and the application of research data derived from these. A general survey of existing learner corpora for Spanish will follow, as background for the description of CAES.

Part 2 focuses on the CAES project itself, looking at the following issues: the origin and development of the project up to its current state, general design and compilation, data collection methodology, text coding and annotation, plus its search tool and its different functions.

Part 3 discusses the results of one main study which uses data from CAES to explore issues of vocabulary in learner Spanish. It is intended as a simple example of the kind of research that can be conducted with material from this corpus. For reasons of space, we will not consider this in exhaustive detail, as it would merit a specific study of its own.

The chapter will conclude with some reflections on questions arising in previous sections, and with the identification of issues for further research. These may be of particular interest to teachers of Spanish as a second/foreign language, SLA researchers, language testers, teacher trainers, Spanish language teaching materials producers and developers, and any professional connected directly or indirectly with the teaching of Spanish.

Section 1: CL, general learner corpora and Spanish learner corpora

1.1. Brief overview of the importance of CL, the emergence of learner corpora and their applications

The emergence of CL has heralded a new approach to the study of language, one in which it is possible to work with real data and to describe the working of language in close detail. It has thus facilitated linguists the access to real examples of the language used in a given context (Adolphs 2008, Lüdeling and Kytö 2008, McEnery and Hardie 2012). According to Biber, Conrad and Reppen (1998: 4), the main characteristics of corpus-based analysis can be described as follows:

(i) It is empirical, in that the analysis and collection of data are required. Attention is paid to patterns of use in natural texts. In Leech's terms (1992: 105), Computer Corpus Linguistics (CCL) is focused on performance rather than on competence;

(ii) It is based on samples of text or a 'corpus', compiled with a particular aim in mind and conceived as representing a particular language;

(iii) Computers are mainly used for the analysis; both automatic and interactive techniques and tools may be used; and,

(iv) Qualitative and quantitative techniques may be applied to reach definite conclusions. Note that corpus data are generally characterised by their flexibility as they allow for multiple approaches and analyses.

Apart from these four features, Leech (1992: 105) also points out that CCL is more heavily focused on linguistic description than on language universals. All of the above can be applied to the acquisition or learning of a second language.[2] By doing so

[2] Although some scholars such as Krashen (1988) make a clear distinction between "acquisition" (more closely related to the first language (L1), being mainly a spontaneous and natural process) and "learning"

with learner corpora—that is, corpora compiled and created according to explicit design criteria for a particular SLA purpose, with samples of written and/or spoken language produced by the learners of a second or foreign language (Granger 1998, 2008)—we obtain information on how students learn the target language, and this is likely to be of practical relevance in language teaching. The starting assumption here is that it is not possible to know how learners learn a language unless we discuss and analyse data provided by them. It is true that learner corpora are not the only instruments available for obtaining data on SLA; Ellis (2004: 673-674) also mentions in this context metalinguistic judgments, that is, learners' judgments on the grammaticality of different L2 structures and patterns, and self-report data, which can be both spoken and written and which are generated by students themselves. However, learner corpora have a clear advantage over these two methods of data collection in being based on language in use, and thus are more direct and spontaneous, and less artificial.

Learner corpora provide data which may be analysed from different perspectives and approaches. Thus, learner corpora data can be used to carry out computer-aided error analysis, that is, by examining learner data we may obtain information on those areas of the target language which seem to be most difficult for students. Thus it is possible to get to know, for example, those grammar points learners of one level or of a particular L1 have most problems with. Although teachers and learners may have assumptions and intuitions about what causes learning difficulties, "this intuition needs to be borne out by empirical data from learner corpora", as Granger (2002: 23) notes.

(more directly connected with the second language (L2), where some kind of effort to learn is typically required), for the purposes of this study, the concepts "acquisition" and "learning" will be used interchangeably. The same will apply to the distinction between "second" versus "foreign" language, which will also be here used as synonyms. Notice, however, that in the case of the CAES project the learners were on the whole students of Spanish as a foreign language. The number of participants as second language learners was very limited indeed as this was restricted to the groups of students from the Spanish universities participating in the project (Santiago, Vigo, Alcalá, León and Vigo).

In some cases learner corpora include an error tag system which clearly facilitates the errors and types of mistakes made by the learners. In line with this, it may be useful to investigate the linguistic features in the target language which L2 learners use significantly more often ("overuse") or less frequently ("underuse") than native speakers. This is what Granger (1998: 12, 2008: 267) refers to as "Contrastive Interlanguage Analysis", usually abbreviated to CIA. Such an approach may involve two main types of comparisons: a) comparison of native language and interlanguage, for instance, native Spanish versus the interlanguage of Spanish produced by a group of Chinese learners with respect to a particular linguistic aspect, such as discourse markers, the use of verbal tenses, tags, prepositions (*para* versus *por*), *ser* versus *estar*, etc.; b) comparison of different types of learner languages, namely comparisons between students of Spanish from different language backgrounds; as an example, we might investigate the extent to which the difficulties which Arabic speaking students face with prepositions in a specific L2 are similar to those experienced by Portuguese students leaning the same language.[3]

Learner corpora studies may also have a wider range of applications (Braun, Kohn and Mukherjee 2006, Aijmer 2009, Lombardo 2009, Reppen 2010, Römer 2011). In this respect the distinction made by Römer (2011: 207) between direct and indirect applications can be here of use. Direct applications affect mainly learners and teachers and they focus on teacher corpus and learner corpus interactions, that is, they have more to do with the actual teaching methodology and pedagogical techniques. Indirect applications, in contrast, have effects on the teaching syllabus, reference works and

[3] For a selection of research studies using this kind of approach, see the learner corpus bibliography of the Centre for English Corpus Linguistics, Catholic University of Louvain, which can be freely accessed at <http://www.uclouvain.be/en-cecl-lcbiblio.html.>. It contains c. 1,100 references, updated on a regular basis. In September 2013 the Learner Corpus Association (LCA) was created whose website also provides interesting information on resources, events and forums on learner corpora research: <http://www.learnercorpusassociation.org/>.

teaching resources being material writers and researchers the agents here concerned. Some of the most important of these applications can be then summarised as follows. In line with the distinction made before, the first one could be regarded as more direct while all the rest would be more indirect.

(i) Language testing and classroom methodology. Learner corpora can provide useful information for both the design of language tests and for the statement of (reference) levels. Furthermore, several scholars (Seidlhofer 2002, Pérez Paredes and Cantos Gómez 2004, O' Keefe, McCarthy and Carter 2007) have made interesting proposals to integrate data derived from (learner) corpora into classroom techniques and activities.

(ii) L2 materials design. Data derived from learner corpora may assist authors and scholars in the production of pedagogical grammars, dictionaries, glossaries, textbooks, workbooks, videos and CDs, teaching guides, etc.[4] It is clear that L2 learners have special needs, and it is logical that publishers want to address their needs as effectively as possible. In spite of all this, all seems to indicate, as Römer (2011: 206) rightly notes that "there is still a lack of awareness of corpora and, in some cases, resistance toward corpora from students, teachers and material writers".

(iii) Computer tools that may help students in the learning of an L2, such as error recognition programs and hypertext on on-line grammars (Granger 2008).

(iv) Syllabus and course design. Learner corpora materials may help in the design of syllabuses and general language curricula, in that they can enhance the pedagogical and practical dimensions of these by yielding useful data for the selection, structuring and grading of teaching content (Granger 2002: 22).

[4] In English there are innumerable materials of this nature. The *Cambridge Learner's Dictionary*, The *Collins Cobuild Series* (grammar, dictionary, English guides), *Oxford Learner's Dictionaries*, *Oxford Learner's Grammar*, *Macmillan Dictionary for Advanced Learners*, *Longman Dictionary of English Online*, *Longman Advanced Learner's Grammar* are just a few. In Spanish fewer such materials are available, although among these we might cite *Gramática básica del estudiante de español* (Difusión), *Gramática práctica del español actual* (SGEL) and a wide range of textbooks.

(v) Planning and implementation of teacher training and teacher development modules. It is not unusual that learner corpora identify weaknesses in the language learning process that are closely related to the structure and contents of the teacher training programme followed by L2 instructors.

Although CL and learner corpora together, in other words, Learner Corpus Research (LCR) can make an important contribution to the study of language and to the language learning process in general, we should also be aware of some of its limitations, such as:

(i) The problem of representativity and the overgeneralisation of findings have always been controversial issues. A (learner) corpus, no matter how large and varied, can ultimately be representative only of its own data. The generalisation of findings to the whole language and to all the learners of different levels and backgrounds should be done with care.

(ii) Not everything can be studied with learner corpora; for instance, pragmatic features, the speaker's communicative intention, paralinguistic traits typical of spoken discourse, etc. are beyond the scope of most of the existing learner corpora (De Cock 1998) although it is true that in the last few years new multimodal learner corpora have been compiled (Adolphs and Carter 2013). That is the case of MULCE (Multimodal Learning and Teaching Corpora) and LETEC (Multimodal Learner Corpus Exchange), for example.[5]

(iii) It is not enough with the retrieval of examples or tokens and with a brief description of the data obtained. It is necessary to discuss and analyse that information in close detail and explore the reasons underpinning those findings. At a subsequent stage it will be important to examine the pedagogical implications that are derived from them.

[5] Further information can be found at the following website: < http://mulce-doc.univ-bpclermont.fr/?lang=fr>. See also the chapter in this volume on the *Spanish Proficiency Training Website* (Koike and Witte).

(iv) In spite of the high accuracy of automatic taggers such as CLAWS (the Constituent Likelihood Automatic Word Tagging System), used for version 2 of the British National Corpus, which are quite effective and serve to fulfill their main objective, corpus tagging (Lüdeling and Kytö 2009) is not always completely correct. On some occasions, it is necessary to revise the tagging provided by these automatic systems and disregard the irrelevant information because it is not totally accurate or it is not relevant to the study to be conducted.

(v) In the transcription of data, particularly spoken, problems often arise owing to the difficultly in achieving high quality recordings of speakers, especially in oral interactions. Being aware of this, recent oral learner corpora have tried hard to cater for this limitation.

(vi) Close attention needs to be paid in terms of how we apply linguistic findings to language teaching. This could be more a question of ethics rather than a limitation of LCR itself since it is derived from the application of the data. However, data should be carefully considered before any learner corpus-based changes are made in our teaching practices.

1.2. General review of the existing learner corpora in Spanish

There are now at least three other major ongoing corpora which can be regarded as similar in purpose to the CAES project. The first is the “Corpus para el análisis de errores de aprendices de E/LE”, that is, the *University of Alcalá Error Analysis Corpus*, containing data on Spanish L2 learners (Cestero et al. 2001). It was officially presented at the 2000 general conference of ASELE (Spanish Association of Teachers of Spanish as a Second Language). This corpus contains only written materials and has been specifically conceived to encode each of the errors found in the corpus, with the aim of exploiting the data for pedagogical purposes. The samples themselves were produced by foreign students of the University of Alcalá, based on controlled compositions and guided written essays. The database includes three main sources of information: the first reflecting participants' personal data (age, nationality, mother tongue, foreign language skills, studies in Spanish, proficiency level, etc.); the second contains the compositions written by participants; and the third lists the mistakes made by these students according to a coding system. The samples, collected in 2001, were from over 320 students of elementary, intermediate and advanced levels with different mother tongues, mainly Japanese, English, German, French, Swedish and Italian.

The second major project is also a written corpus of Spanish as an L2. The "Corpus Escrito del Español L2" (CEDEL2) is designed and compiled by Cristóbal Lozano from the University of Granada (Lozano 2009, Lozano and Mendikoetxea 2013). It is itself part of a larger project known as WOSLAC (Word Order in Second Language Acquisition Corpora), directed by Amaya Mendikoetxea from the Autonomous University of Madrid. CEDEL2 currently contains over 730,000 words from 1,750 English students of Spanish and also from 660 Spanish learners of English. Data collection was done online, after students had been classified into different levels of language proficiency according to the results of the University of Wisconsin's (1998) placement test. For the collection of the data, participants completed an essay on a topic they could select from a list of twelve. These included issues like the description of a famous person, a summary of what they had done over the weekend, their future plans, their opinions on the new Spanish anti-smoking law, the legalisation of marijuana, the problem of immigration, etc. This corpus is expected to reach one million words at the end of the project, and allows for contrasts between students of several levels of language proficiency and between native and non-native speakers, as well as including a subcorpus of native speakers of Spanish. The tagging of the data in XML format was done with the UAM Corpus Tool, developed by Mick O'Donnell (2008).

Whereas the previous two corpora focus exclusively on written language, the *Spanish Learner Language Oral Corpus* (SPLLOC) is an exclusively oral corpus, containing only spoken samples of English-speaking students of L2 Spanish, from beginners to advanced level. Currently this project brings together two related initiatives, SPLLOC1 and SPLLOC2, which began in April 2008 and was completed in January 2010 (Mitchell, Domínguez, Arche, Myles and Marsden 2008). In order to conduct contrastive studies, oral samples of speakers of Spanish as L1 were also compiled. The data collection instruments were basically stories told by the participants themselves, plus interviews and photograph descriptions. The final database contains samples of the oral production of Spanish students in different types of discourse genres, accompanied by written transcripts following the CHILDES format.

In addition to the above, there are some other Spanish learner corpora of a more limited size and representation. Among these we might mention: the “Corpus of Academic Texts”, containing the production of foreign university students and compiled by Álvarez López (2005), and consisting of 62 samples of 40 college students who were studying different courses at the Faculty of Philology of the University of Alcalá during the 2000-2001 academic year; the corpus of conversations of Spanish as a foreign language (García 2005), which includes the interactions of 11 students from 3 different levels of language proficiency; the corpus of texts by Italian university students of Spanish as foreign language (Gutiérrez Quintana 2005), involving 44 Italian informants who were completing the degree of Foreign Languages and Literatures at the University of Sassari; and, the corpus of written texts produced by Spanish Taiwanese college students (Tzu-Ju 2005), consisting of 185 essays completed by students of Spanish at Providence University in Taiwan whose L1 is Mandarin Chinese. Furthermore, the Spanish Learner Oral Corpus (*Corpus Oral de Español como Lengua Extranjera*) contains spoken samples of 40 learners of A2 and B1 levels up to a total of 50,000 words. It was compiled by Campillos Llanos (2014) as part of his doctoral thesis and it aims to improve the teaching of Spanish to foreign students by considering the errors and difficulties of learners with the same L1. Finally, within this group of corpora of a limited size we can include the longitudinal Spanish Corpus of Italian Learners (SCIL) which consists of 457 compositions written by a total of 43 informants, whose proficiency levels range from A1 to B2. It was developed by Bailini (2013) at the Università Cattolica del Sacro Cuore. To this list we can add the *Anglia Polytechnic University Learner English Corpus*, the *Aprescrilov* initiative, the *Díaz Corpus* based at the University Pompeu Fabra in Barcelona, the *Japanese Learner Corpus of Spanish* (University of Birmingham), the *Spanish Corpus Proficiency Level Training* (University of Texas) and the *Fono.Ele Corpus* (University of Alcalá).[6] The following table provides an overview of the most important features of the principal Spanish leaner corpora:

[6] See chapter in this volume for the Spanish Corpus Proficiencly Level Training. For further information about the Fono.Ele Corpus, visit : <http://www3.uah.es/fonoele/>

Table 1: Main Spanish Learner Corpora

Corpus name	**Compilers**	**Participants' native language**	**Compilation date**	**Size**	**Text types**	**Observations**
Corpus para el análisis de errores de aprendices de E/LE (CORANE)	U. of Alcalá (Cestero et al.)	English, German, French, Swedish, Portuguese, Japanese & Italian	2000		guided essays	Focussed on learner errors.
Corpus Escrito del Español L2 (CEDEL2)	U. of Granada (Lozano) Autonomous Univ. Madrid (Mendikoetxea) <https://www.uam.es/proyectosinv/ woslac/cedel2.htm>	English		730,000 words from over 1,700 students	essays	It includes a subcorpus of native speakers of Spanish. It is still in progress so those interested may even contribute to its final compilation.
Spanish Learner Language Oral Corpus 1 & 2 (SPLLOC)	U. of Southhampton, York & Newcastle (Marsden, Mitchell, Myles, Domínguez) <http://www.splloc.soton.ac.uk/>	English	2008-2010		spoken samples	It follows the CHILDES transcription model. It also includes samples of spoken Spanish produced by native speakers
Corpus de Textos Académicos	U. of Alcalá (Alvarez López)	English, French, Italian, Dutch	2000-2001	62 samples, 49,045 words	essays from exams	Academic writing
Corpus de Conversaciones del español como lengua extranjera	U. de Alcalá (García)	German, French, Serbian	2005		conversations	It tries to elicit spontaneous language
Corpus escrito de estudiantes italianos de EL/E	Gutiérrez Quintana (U. of Sassari)	Italian	2000-2001	10,000 words	essays	
Corpus of written texts of Taiwanese students of	Tzu-Ju Providence University (Taiwan)	Mandarin Chinese	1999-2001		written texts	Focussed on problematic issues for Chinese students

Spanish						
Corpus Oral de Español como Lengua Extranjera	Campillos Llanos (Autonomous U. of Madrid) <http://cartago.lllf.uam.es/corele/home_es.html>	nine different languages represented	2010-2012	more than 50,000 words	semi-spontaneous interviews, narrative and descriptive tasks	Focused mainly on error analysis of oral production.
Spanish Corpus of Italian Learners (SCIL)	Bailini Università Cattolica del Sacro Cuore	Italian	2012-2013	124,186 tokens		It allows both cross-sectional and longitudinal studies.
The Anglia Polytechnic University (APU) Learner Spanish Corpus	Anne Ife Anglia Ruskin University, UK	various		120,000 words	written	
Aprescrilov ("Aprender a Escribir en Lovaina")	Kris Buyse KU Leuven, Belgium	Dutch	2005-2011	c. 1 million words, 2,700 texts	written	Error-annotated
The Díaz Corpus	U. Pompeu Fabra (Díaz García)	German Swedish Icelandic Korean Chinese			spoken semi-spontaneous (structured interviews) & experimental (structured questionnaires)	
The Japanese Learner Corpus of Spanish	Yoshihito Kamakura, U. of Birmingham, UK	Japanese		83,400 words	written (student essays)	
The Spanish Corpus Proficiency Level Training (SPT)	U. of Texas (Dale Koike) <http://www.laits.utexas.edu/spt/>	English and Spanish heritage language learners	2010-2011		spoken (dialogues about a given set of topics)	Transcripts are provided for each of the videos. Conceived for multiple purposes: teacher training, research, self-improvement of Spanish and classroom assignments.

Fono.Ele Corpus	Mª Angeles Álvarez Martínez & Ana Blanco Canales, U. of Alcalá <www3.uah.es/fonoele>	6 diff. languages represented: German, Greek, Tawanese, Polish, Portuguese and Egyptian	2010 onwards	96 samples	spoken (short structured conversations, readings of texts, phrases and words	Focussed on the phonological component of Spanish as a foreign language.

SECTION 2: The CAES (*Corpus de Aprendices de Español*) Project

2.1. Origin and development of the project up to its current state

This project was wholly financed by the Cervantes Institute (CI) and carried out by a research team from the University of Santiago.[7] At the end of 2011 a proposal was submitted by the main researchers to the CI for the compilation and completion of the corpus, drawing attention to the importance this tool could have for the different sectors of the teaching of Spanish as a foreign language community. Once the proposal was approved, the first steps were taken for the design and creation of a computer program which could be used for entering the data by students themselves at CI centres across the world in a simple but reliable way. Thus the project would benefit from the CI international network of centres, and problems with the transcription of data would be avoided as the participants themselves were the ones who entered all primary data in the program, rendering all manner of intermediate agency unnecessary. This guaranteed that the data corresponded faithfully to the original, in that no subsequent interpretation or transcription took place. This is important in any corpus, but especially so in the case of learner corpora, where it is common to find samples with misspellings, inaccuracies and mistakes as the result of an incomplete command of the target language.

At this stage it was important to design a corpus which could be computerised, was representative of the language to be represented, that is, learner Spanish, and which was also well-organised, user-friendly and reflected participants' level of L2 and their

[7] The project members and their roles were as follows: Directors, Guillermo Rojo and Ignacio Palacios; computer programmer (collection and search programs), Mario Barcala; team members in charge of the manual disambiguation of the data, Marlén González González and Alba Fernández Sanmartín; team member responsible for the design and application of the tagging system, María Paula Santalla del Río; and, finally, Susana Sotelo Docío, team member responsible for the automatic annotation. The corpus can be freely accessed at the following website:
<http://www.cervantes.es/lengua_y_ensenanza/tecnologia_espanol/caes.htm>

L1. These two variables were particularly important because they would allow us to draw comparisons across levels of proficiency and according to learners' L1s. However, it also meant that a bespoke application had to be designed by an expert in CL technology.

The piloting of this application, created specifically for the collection of data, was conducted with three groups of students of different levels and language backgrounds from the Universities of Santiago de Compostela, Vigo (Spain) and do Minho (Portugal), that is, with groups of subjects with a similar profile to those of the final participants in the project. This preliminary process served to identify possible weaknesses in the procedures. Adjustments were made where necessary, such as tweaking the task instructions, which were at times not easy to understand or had not been reworded clearly enough. There were also some technological details that required attention. By September 2012 a broader, general data collection was conducted with the participation of over 28 CI centres and 8 universities from 15 different countries.[8] At a previous stage all the participating institutions had been contacted and briefed about the project. A data collection protocol was prepared with exact instructions to be followed at each stage. The teachers at each of the CI centres also had to fill in a report form detailing the number of students participating in the data collection as well as the number of samples obtained. This report form would serve as back-up information in case any technical or other issues arose during the reception of the samples.

Students of English, French, Arabic and Portuguese took part in this first part of the project. The second stage, which began one year later, incorporated participants of

[8] The whole list of CI centres and universities participating in the project is: Amman CI, Beirut CI, Brasilia CI, Brussels, CI, Bordeaux CI, Casablanca CI, Chicago CI, Curitiba CI, Damascus CI, Dublin CI, Cairo CI, Fez CI, Lyon CI, Marrakech CI, Moscow CI, New York CI, Oran CI, Paris CI, Beijing CI, Porto Alegre CI, Recife CI, Río de Janeiro CI, Salvador de Bahía CI, Sao Paulo CI, Sidney CI, Tétouan CI, Tangier CI, Tunisia CI, Univ. of Alcalá, Univ. of León, Univ. of Salamanca, Univ. of Santiago de Compostela, Univ. of Vigo, Univ. of Manchester, Univ. do Minho (Portugal) and Univ. of Washington (Seattle, USA).

two more L1s, Russian and Mandarin Chinese. The main objective was to expand and refine the samples already collected. All the data retrieved were stored on a server of the University of Santiago while the bespoke application capable of facilitating search and retrieval of the data according to different variables was being designed and tested (cf. section 2.5). This whole process, which involved a number of pilot sessions, also included the tagging, annotation and disambiguation of corpus samples.

2.2. General design and compilation

As mentioned above, CAES is a collection of written texts produced by students of Spanish as a foreign language of different levels, from A1 to C1, according to the Common European Framework of Reference (Council of Europe, 2001). Samples from C2 level were not included because, as also noted, students had to certify a particular level of the above when completing the tasks. For C2 students, since at the time of the general data collection they were still in the middle of their courses, the (very high) C2 level of proficiency had not yet been attained. Subjects of six native or L1 languages are represented: Arabic, Mandarin Chinese, French, English, Portuguese and Russian. In its current form the corpus contains a total of over 570,000 words, including data from participants of all levels and L1s. The original data had to be carefully filtered since there were samples of students with a different L1 to those considered, as well as other potential participants whose data were deemed invalid for a variety of reasons (incomplete or unclear tasks, difficulty in certifying level of proficiency, no understanding of the tasks to be done, etc.).[9] The current CAES version contains samples produced by 1,423 students of Spanish as a foreign language who wrote two or

[9] This was particularly so in the case of the universities since the groups of students were most often multilingual, hence making the control of the L1 variable difficult.

three texts in keeping with their level; this led to a total of 3,881 written tasks integrated in 1,423 samples. See Table 2 below. Further tables are also provided in appendix 1 with supplementary information regarding the participants’ general profile and the total number of sample units collected according to different variables.

Table 2: Main Features of the CAES Project

Compilers	**Participants' native language**		**Participants' gender**		**Participants' level**		**Participants' main countries represented**		**Participants' studies completed**		**Participants' age**		**Size**	**Text types**
University of Santiago de Compostela (Rojo, Palacios, et al.). See note 6.	Arabic	497	male	521	A1	526	Brazil	319	University	908	15-21	498	570,000 words	essays and guided writing tasks in keeping with the students' proficiency level
	Portuguese	361	female	902	A2	421	Morocco	312	Primary	205	22-30	466		
	English	227			B1	252	USA	139	Secondary	127	31-40	196		
	French	143			B2	162	China	127	Other	183	41-60	198		
	Mandarin Chinese	128			C1	62	France	92			+61	65		
	Russian	67					Siria	70						
							Russia	62						
							Afghanistan	52						
							Ireland	38						
							Algeria	32						
							Portugal	31						
							Lebanon	26						
							Jordan	21						
							Tunisia	16						

2.3. Tasks devised for each of the levels considered and description of the sample collection method

Participants had to complete a number of written tasks in keeping with their previously certified level of Spanish (cf section 2.2). These tasks were the same for all the students, independently of their country of origin and of the place where learners completed them. This guaranteed the comparability of the learner samples. The variable of level (language proficiency) was tightly controlled, since it was important to make sure that the students were classified correctly. These written tasks were designed according to the Common European Framework descriptors for each of the levels and following the guidelines provided by the CI regarding the DELE tests ("Diplomas de Español como Lengua Extranjera", General Certificate of Spanish as a Foreign Language) for each of the three levels (beginner, intermediate and advanced), as well as in accordance with the CI's General Curricular Document.[10]

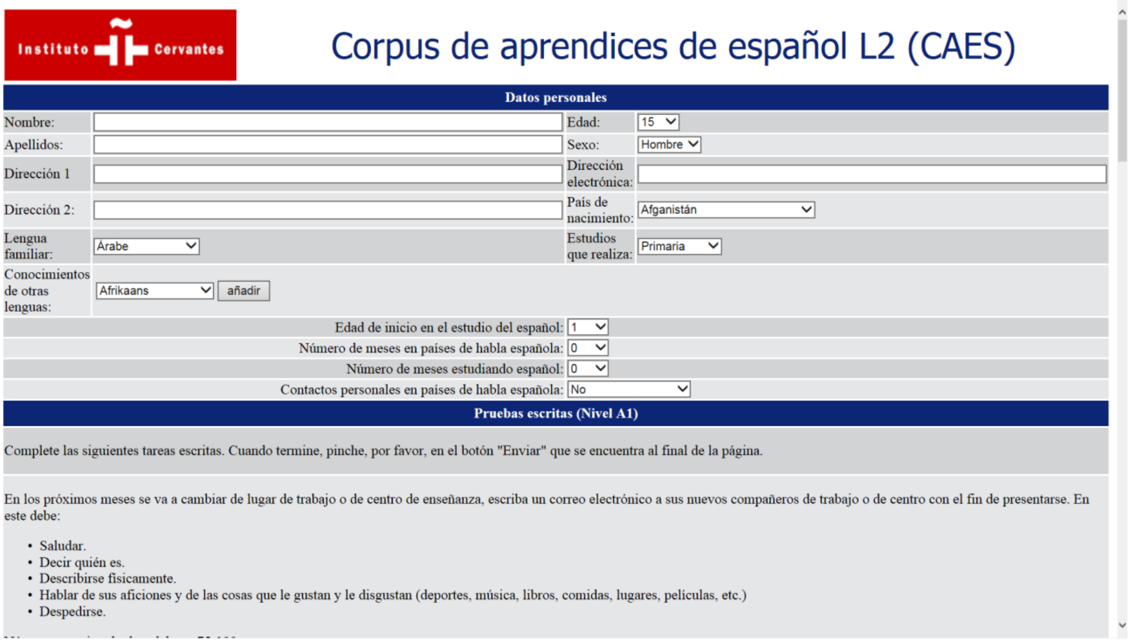
Clear instructions were provided for each of the tasks, indicating the number of words required, and with examples given when necessary. Thus, for instance, participants of A1 level were asked to write two 75-100 words emails, one introducing themselves to the group of students in their class or at work, and the second describing their family to a friend, and then to compose a brief note of 30-40 words addressed to the people they were living with saying they were going to be late for dinner. In line with their proficiency level, C1 learners had to write a critical review and an email, both of 400-500 words. An effort was made to make these writing activities resemble authentic or real life tasks as much as possible. Thus, as mentioned, tasks included writing emails to friends and relatives, applying for a job, composing notes and

[10] Further information can be found at the following website links: <http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/>, <http://diplomas.cervantes.es/> and <http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/default.htm>

messages, booking a hotel room, writing a postcard to friends, telling a funny story, making a complaint, filling in a form, writing a film review, writing an argumentative essay, etc. Participants did not have access to any reference materials during their writing and had one hour to complete the whole process.

Information on the project was provided to all the CI centres around the world encouraging them to participate (cf. section 2.2). Detailed information was then given by the corpus compilers to each of the teachers responsible for the different groups of students. As explained above, a computer tool was created so that participants could enter their personal details (age, sex, knowledge of foreign languages, stays in Spanish-speaking countries, L1, starting age for the study of Spanish) and complete the appropriate writing tasks for their level of Spanish. Immediately prior to this, they were asked to fill in a consent form giving their permission for the use of the data for research purposes.

Figure 1: CAES general interface for data collection



Corpus de aprendices de español L2 (CAES)

Datos personales

Nombre:		Edad:	15
Apellidos:		Sexo:	Hombre
Dirección 1		Dirección electrónica:	
Dirección 2:		País de nacimiento:	Afganistán
Lengua familiar:	Árabe	Estudios que realiza:	Primaria
Conocimientos de otras lenguas:	Afrikaans añadir		

Edad de inicio en el estudio del español:	1
Número de meses en países de habla española:	0
Número de meses estudiando español:	0
Contactos personales en países de habla española:	No

Pruebas escritas (Nivel A1)

Complete las siguientes tareas escritas. Cuando termine, pinche, por favor, en el botón "Enviar" que se encuentra al final de la página.

En los próximos meses se va a cambiar de lugar de trabajo o de centro de enseñanza, escriba un correo electrónico a sus nuevos compañeros de trabajo o de centro con el fin de presentarse. En este debe:

- Saludar.
- Decir quién es.
- Describirse fisicamente.
- Hablar de sus aficiones y de las cosas que le gustan y le disgustan (deportes, música, libros, comidas, lugares, películas, etc.)
- Despedirse.

Due to the design of the procedure, students' progress could be conveniently monitored, and the corpus team were able to deal with problems which arose during the whole process. Once all the data were entered in the computer, the participants themselves clicked on the screen command to send their materials. The information was then stored on a University of Santiago server.

As described above (cf. section 2.1), the process had been piloted beforehand with three groups of students to find out if the tasks proposed were suitable for each level and whether the computer programme actually worked effectively.

2.4. Text encoding and annotation

The texts integrated into CAES adopt the format of XML documents from the start. All the necessary data for the identification of the values in each of the tasks completed, and those data which correspond to the features considered for purposes of classification, are found in the header; the written text, however, occurs in the body of the document in each case. This means that all the documents can be processed and stored together in a database from which it is possible to extract tokens of a particular expression, applying filters according to one or more of the parameters considered (L1, proficiency level, speaker's gender, etc.). However, the design of the project was much more ambitious and also anticipated the annotation and lemmatisation of each of the forms contained in the corpus, as well as the construction of a search tool capable of retrieving considerably more refined data.

Automatic morphological annotation (and lemmatisation) is a complex and delicate process, and even among specialists there is sometimes a lack of agreement as to the appropriate description of a particular element. The first problem, of course, concerns the determining of the tagging system to be used. Here, a balance has to be kept between two opposite perspectives. On the one hand, there must be a general theoretical adequacy, so that it is not excessively biased towards a certain perspective and thus that it is suitable for different types of analysis. On the other hand, it should have a sufficient degree of detail and clarity so as to allow researchers to find the lexical elements and the grammatical phenomena that are of interest to them. The second problem concerns the reliability of the disambiguation process, which is especially difficult here due to the enormous number of homographs existing in Spanish. Finally, an issue arising when annotating any text, but which has added significance with these materials, is the lack of conformity to standard orthographic rules (those that are determined by the lexicon) and, more especially, the morphological and lexical features that are likely to occur in very large numbers in texts written by subjects with an incomplete command of the language.

The tagging system used in this project is an adapted version of the one generally employed in tasks of this nature by members of the Spanish Grammar Research Team at the University of Santiago de Compostela. In its final version, and for this first stage of the CAES project, it consists of 702 different tags.[11] This is a high figure, no doubt, but we believe there is a good reason for it. Considering that this is a general purpose corpus, we anticipated that a wide range of morphological and lexical features would potentially be present in the many different searches to be conducted, given the very different purposes and objectives of those using the corpus. The option of

[11] The whole list of categories and subcategories can be found at the CAES project homepage: <http://galvan.usc.es/caes>. The main ones are: *Abreviatura* (abbreviations i.e. etc.), Adj. (adjective), Adv. (adverb), *Número* (number), Conj. (conjunction), Det. (determiner), *Fecha* (date), *Fórmula* (formulae), *Hora* (time), Interj. (interjection), *Onomatopeya* (onomatopeia), Prep. (preposition), Pron. (Pronoun), *Símbolo* (symbol), Sust. (noun), verb. (verb), Punt. (punctuation mark), *Sigla* (acronyms, i.e. ONU). The main categories (noun, verb, adjective, adverb, pronoun and determiner) are, in their turn, subdivided into subcategories; thus, for instance, within the adverb group we find Tiem. (time adverb), Mod. (manner adverb), Quant. (quantity adverb), Int. (interrogative adverb), Rel. (relative adverb).

retrieving elements defined in close detail seems to be basic to us. Furthermore, we also kept in mind from the beginning that automatic annotation and disambiguation would resolve a limited number of elements and therefore most of the work would have to be done manually by specialists in the field, thus avoiding in great measure the problems found when using a very wide tagging system in the automatic disambiguation processes. Finally, the design of the research tool had already anticipated a hierarchised system going from the general to the particular in such a way that corpus users would not need to be acquainted with all the complexity of the tagging system and could arrive easily at the level they required.

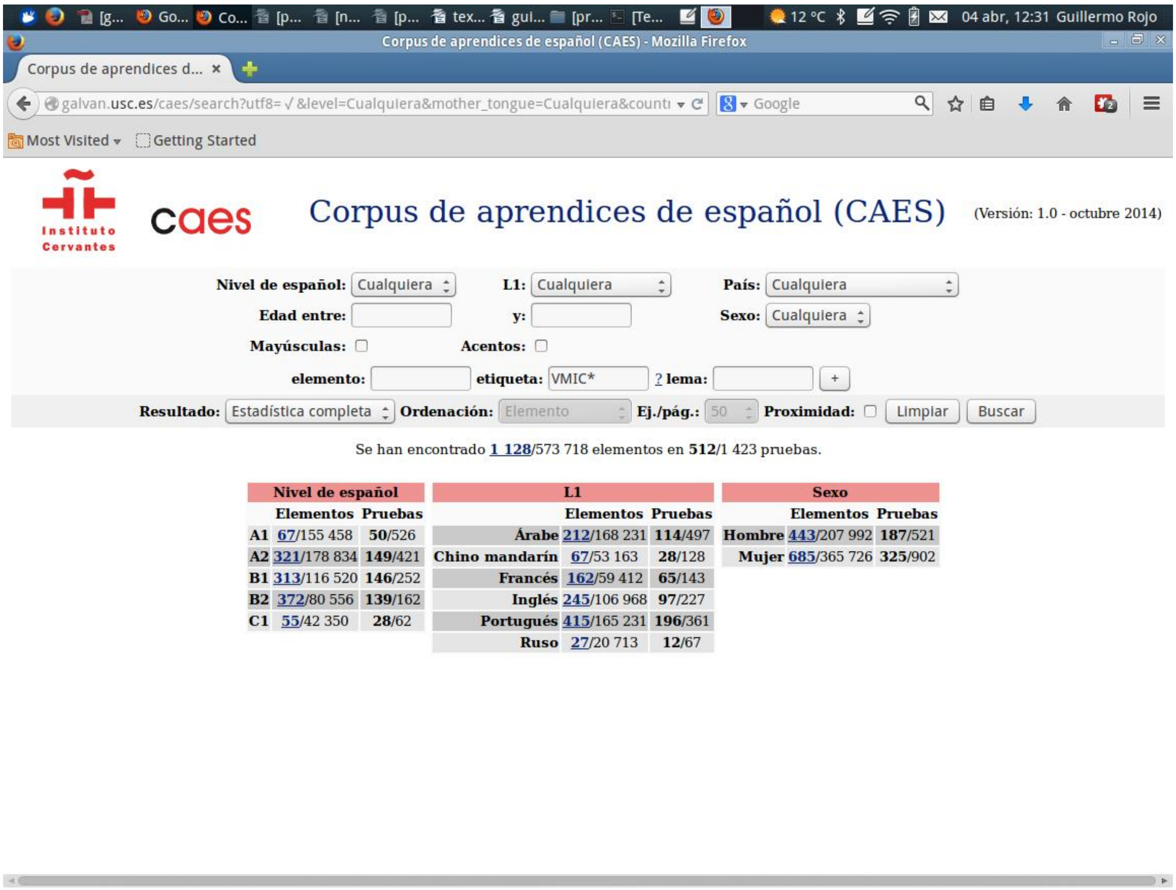
As the linguistic features of the CAES texts were quite different from those observed in native speakers of Spanish, and also differed greatly from one to another depending on the learner’s L1, it seemed to make little sense to spend a long time creating a training corpus, or perhaps as many training corpora as L1 involved and to extract from these the necessary statistical data to disambiguate automatically the rest of the texts. We therefore decided to use FreeLing (Padró & Stanilovsky 2012), an open source language analysis tool suite, and later on to make, through typical substitution routines, the necessary adjustments of the equivalences between the FreeLing automatic tagging system and the one our team intended to use. This obviously solved the problem of the conversion of tags in those cases in which one tag was equivalent to another individual tag, or when several tags were equivalent to a single one; however, this was not the case when one tag was equivalent to several of them. The existence of a large number of unknown elements was an additional problem here. As a result of all this, we created a program so that human experts could carry out the disambiguation process manually by associating every element to any of the tags attributed by FreeLing (not necessarily one selected by the program) or to any other tag not considered by the program. As expected, this was a long and tiring process, although the result was a corpus of almost 600,000 words properly annotated and controlled through several revision processes. This was undoubtedly the stage of the whole process which demanded the highest working load; however, it was worthwhile, not only in terms of the final product, the CAES project, but also because we now have a number of texts that we could use as pilot corpora for all the L1s present in the corpus.

2.5. The search tool

In keeping with the enormous effort made in the manual disambiguation process, the search tool created needed to be wide and flexible enough so that researchers could easily obtain the maximum amount of data from CAES. Overall, the tool developed allows researchers to retrieve statistical information and textual examples of elements, lemmas, word classes and grammatical categories with filters on the parameters that make up the corpus (basically, the learner's L1 and level of proficiency in Spanish, but also age, sex and country of origin). Furthermore, it gives us the possibility of distinguishing between lower and higher case words, accented or non-accented, as well as allowing searches based on the co-occurrence of several elements in specific relative positions.

The first line of the data retrieval is the statistical analysis. It is possible to obtain the overall frequency of any lemma, element or grammatical subcategory, that which corresponds to a number of parameters (a particular L1 or proficiency level), or all of them at the same time.

Figure 2: CAES screenshot providing information on the overall frequency of the postpreterite data


Corpus de aprendices de español (CAES) - Mozilla Firefox
Corpus de aprendices d...
galvan.usc.es/caes/search?utf8=√&level=Cualquiera&mother_tongue=Cualquiera&counti
Google
Most Visited
Getting Started
04 abr, 12:31 Guillermo Rojo
12 °C
Instituto Cervantes
caes

Corpus de aprendices de español (CAES) (Versión: 1.0 - octubre 2014)

Nivel de español: Cualquiera L1: Cualquiera País: Cualquiera
Edad entre: y: Sexo: Cualquiera
Mayúsculas: □ Acentos: □
elemento: etiqueta: VMIC* ? lema: +
Resultado: Estadística completa Ordenación: Elemento Ej./pág.: 50 Proximidad: □ Limpiar Buscar

Se han encontrado **1 128**/573 718 elementos en **512**/1 423 pruebas.

Nivel de español		
	Elementos	Pruebas
A1	67/155 458	50/526
A2	321/178 834	149/421
B1	313/116 520	146/252
B2	372/80 556	139/162
C1	55/42 350	28/62

L1		
	Elementos	Pruebas
Árabe	212/168 231	114/497
Chino mandarín	67/53 163	28/128
Francés	162/59 412	65/143
Inglés	245/106 968	97/227
Portugués	415/165 231	196/361
Ruso	27/20 713	12/67

Sexo		
	Elementos	Pruebas
Hombre	443/207 992	187/521
Mujer	685/365 726	325/902

As Figure 2 shows, we are provided with the number of tokens for each of the variables considered, together with the number of tasks (*pruebas*) where they are found. The total figures are also presented so that it is easy to find the normalised frequency of the element, lemma or grammatical category in question and compare it with others. Table 3 shows the figures according to the variables of proficiency level and L1.[12]

[12] From the data presented in Table 3, we gather that there is a clear increase in the use of these forms as the learner's proficiency level progresses. C1 is an exception to this general tendency which may be related to the types of texts learners had to write. As regards a possible correlation with the different L1s, two clear groups can be observed: the highest frequencies are found with students of L1 French, English and Portuguese while the lowest ones correspond to those students with L1 Arabic, Mandarin Chinese and Russian.

Table 3: General and normalised frequencies of the postpreterite according to the variables of learner's L1 and level of proficiency.

Source: CAES <http://galvan.usc.es/caes>

	tokens	**elements total**	**norm. freq.**
A1	67	155,458	430.98
A2	321	178,834	1794.96
B1	313	116,520	2686.23
B2	372	80,556	4617.91
C1	55	42,350	1298.70
Arabic	212	168,231	1260.17
Mand. Chinese	67	53,163	1260.28
French	162	59,412	2726.72
English	245	106,968	2290.40
Portuguese	415	165,231	2511.64
Russian	27	20,713	1303.53

The second line (see figure 3 below) provides the specific texts where a particular element, lemma or grammatical category is found. The sequences are presented in regular columns and also include information on the learner’s L1 and proficiency level. In addition, if we move the mouse cursor to the different areas of each line, we can obtain further information about each set of data. This basic information, which can be reorganised if necessary, together with the context provided by the search program, may be enough for most analyses. However, it is possible to retrieve more data if required. Thus if we click on the example number, we move to a second screen which provides relevant information on the leaner who wrote the text (sex, age, native language, country, educational level, proficiency level, number of years devoted to the study of Spanish, personal contacts for the learning of Spanish and, according to their own self-assessment, proficiency skills in other foreign languages) together with the following:

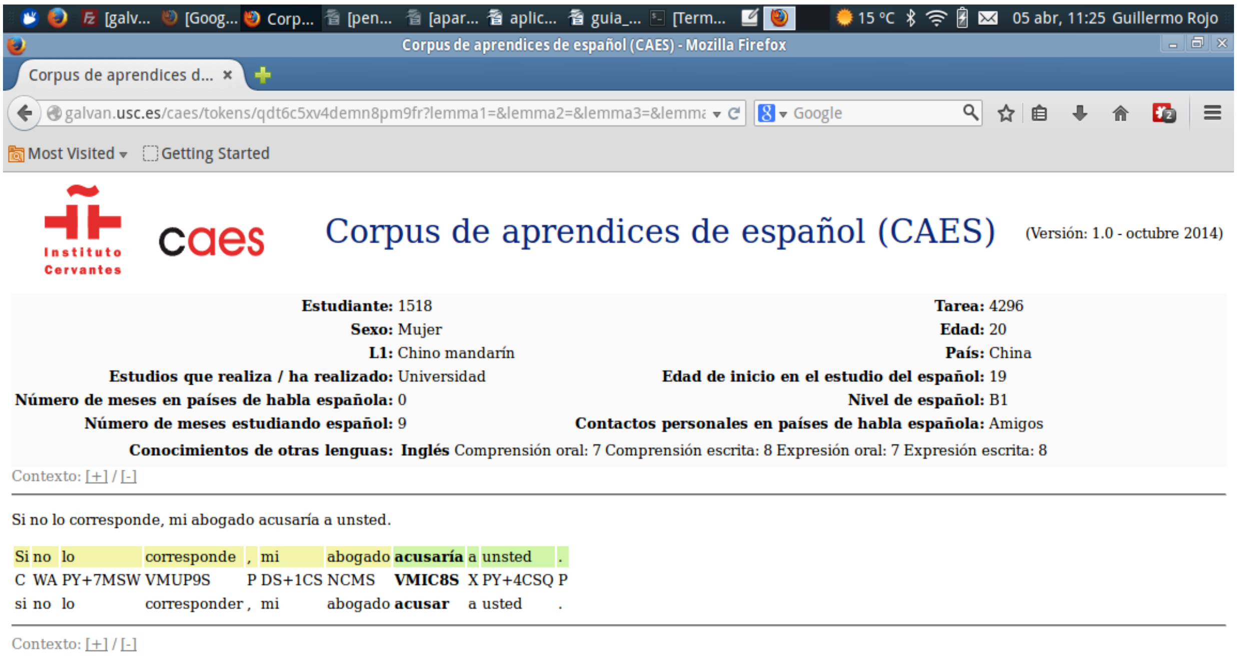
- full sentence where the retrieved form was found, as in the original, since no changes were made;

- lexical items present in the sentence;

- morphosyntactic tags corresponding to each of these elements, and, finally,

- lemmas to which they belong.

Figure 3: CAES screenshot with full information on one particular use of the postpreterite (conditional)


[galv... [Goog... Corp... [pen... [apar... aplic... guia_... [Term...
15 °C
05 abr, 11:25 Guillermo Rojo
Corpus de aprendices de español (CAES) - Mozilla Firefox
Corpus de aprendices d... ×
galvan.usc.es/caes/tokens/qdt6c5xv4demn8pm9fr?lemma1=&lemma2=&lemma3=&lemm
Google
Most Visited
Getting Started
Instituto Cervantes
caes

Corpus de aprendices de español (CAES) (Versión: 1.0 - octubre 2014)

Estudiante: 1518
Sexo: Mujer
L1: Chino mandarín
Estudios que realiza / ha realizado: Universidad
Número de meses en países de habla española: 0
Número de meses estudiando español: 9

Tarea: 4296
Edad: 20
País: China
Edad de inicio en el estudio del español: 19
Nivel de español: B1
Contactos personales en países de habla española: Amigos

Conocimientos de otras lenguas: **Inglés** Comprensión oral: 7 Comprensión escrita: 8 Expresión oral: 7 Expresión escrita: 8

Contexto: [+] / [-]

Si no lo corresponde, mi abogado acusaría a unsted.

Si	no	lo	corresponde	,	mi	abogado	**acusaría**	a	unsted	.
C	WA	PY+7MSW	VMUP9S	P	DS+1CS	NCMS	**VMIC8S**	X	PY+4CSQ	P
si	no	lo	corresponder	,	mi	abogado	**acusar**	a	usted	.

Contexto: [+] / [-]

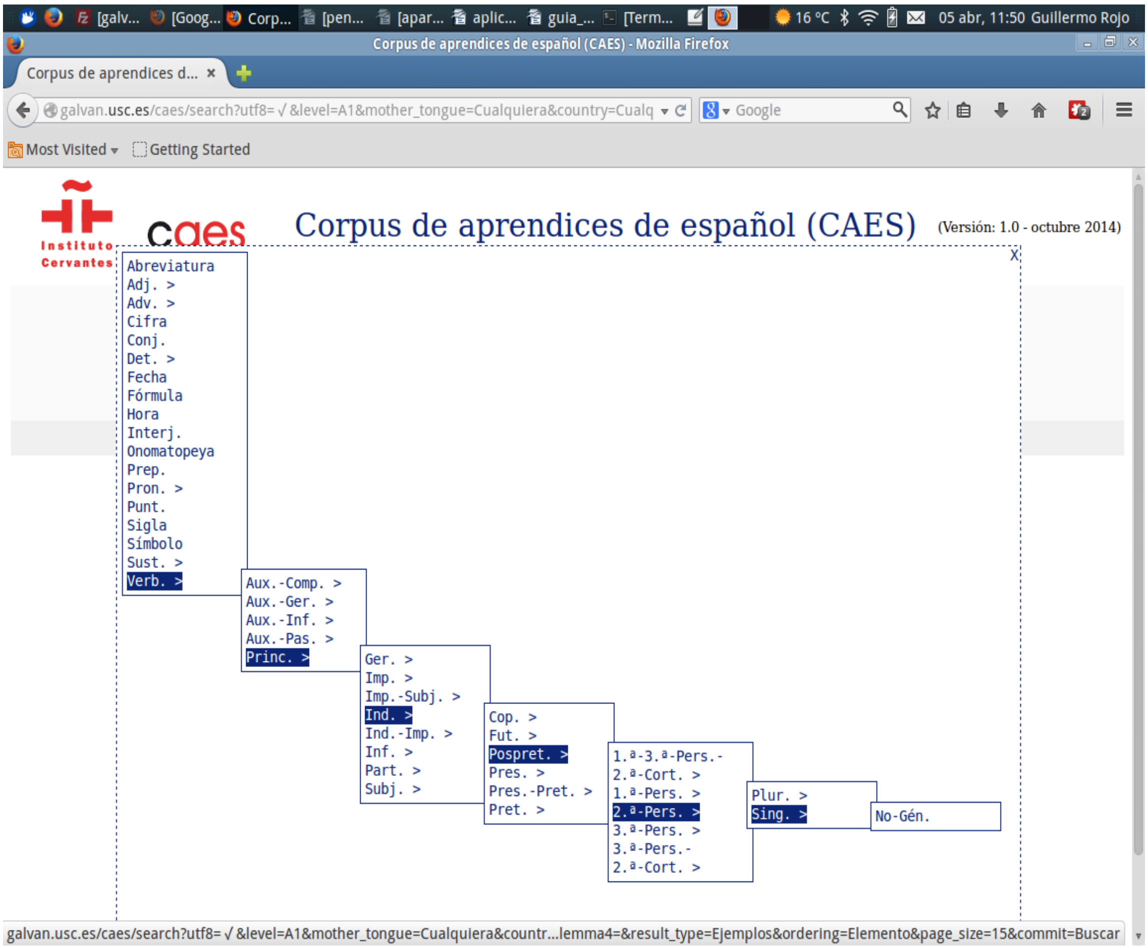
The retrieved sequence and the information associated with it correspond to the sentence where the element retrieved was found. If necessary, it is possible to expand the context before and after by clicking the windows with the '+' and '-' signs located at the top and bottom. All these searches can clearly be refined through the selection of the different options included in the general parameters; to continue with the same example, this would allow us to retrieve all cases of a postpreterite form corresponding to female B1 learners with L1 Mandarin Chinese.

It is also possible to retrieve fine-grained searches through the use of regular expressions that in combination with the grammatical properties associated with each of the elements may return significant results. Given that the corpus is lemmatised, the best way to retrieve all the uses of a particular verbal form belonging to a paradigm is not by using a regular expression (i.e. *lleg**) to simulate the corresponding morphological structure; it is faster and more efficient to select the lemma *llegar*. However, on other occasions the use of regular expressions may be more suitable. Thus, for example, it is possible to retrieve those cases of lemmas ending in *-ción* (singular or plural) by entering **ción* in lemma and noun in tag, which will return all nouns (masculine and feminine, singular and plural) that show this formal structure.

The manual disambiguation tasks carried out in the corpus allows us to retrieve, for example, all forms that a learner associates with a particular verb without any kind of limitations arising from their morphological or spelling features. Thus, for example, the search of the lemma *salir* gives the forms that correspond to the different lemmas that form part of the verb together with other forms used by the learners that are not connected either in terms of spelling or with the standard morphology of the element.

Although the tagging system is formally very complex, the search tool allows us to conduct the search in very simple terms: word classes and categories applied in each case are hierarchised so that the different features occur at the same time as the selection process. This is shown in the figure below:

Figure 4: CAES screenshot showing the gradual selection of features to construct a grammar search


[galv... [Goog... Corp... [pen... [apar... aplic... guia_... [Term...
16 °C 05 abr, 11:50 Guillermo Rojo
Corpus de aprendices de español (CAES) - Mozilla Firefox
Corpus de aprendices d... ×
galvan.usc.es/caes/search?utf8=√&level=A1&mother_tongue=Cualquiera&country=Cualq
Google
Most Visited
Getting Started
Instituto Cervantes
caes
Corpus de aprendices de español (CAES)
(Versión: 1.0 - octubre 2014)
X
Abreviatura
Adj. >
Adv. >
Cifra
Conj.
Det. >
Fecha
Fórmula
Hora
Interj.
Onomatopeya
Prep.
Pron. >
Punt.
Sigla
Símbolo
Sust. >
Verb. >
Aux.-Comp. >
Aux.-Ger. >
Aux.-Inf. >
Aux.-Pas. >
Princ. >
Ger. >
Imp. >
Imp.-Subj. >
Ind. >
Ind.-Imp. >
Inf. >
Part. >
Subj. >
Cop. >
Fut. >
Pospret. >
Pres. >
Pres.-Pret. >
Pret. >
1.ª-3.ª-Pers.-
2.ª-Cort. >
1.ª-Pers. >
2.ª-Pers. >
3.ª-Pers. >
3.ª-Pers.-
2.ª-Cort. >
Plur. >
Sing. >
No-Gén.
galvan.usc.es/caes/search?utf8=√&level=A1&mother_tongue=Cualquiera&countr...lemma4=&result_type=Ejemplos&ordering=Elemento&page_size=15&commit=Buscar

Finally, with this search tool it is possible to conduct combined searches of up to four elements, lemmas or tags. Thus, for example, if we select the lemma *llegar* followed by a preposition then followed by a proper name, the expected results are returned, including phrases such as *llegar en Madrid*. By using the right options we can retrieve examples of constructions such as *haber* + participle, *ir* + *a* + infinitive, *dejar* + *de* + infinitive, etc. If we tick on the noun tag window followed by adjective then a second adjective, we retrieve complex cases such as *vida española antigua*, *producción literaria latinoamericana*, *derecho civil ruso*, etc. The lemma *querer* followed by *que* and a verb in the indicative form will show cases of an incomplete knowledge of the arguments governed by this verb in that context, examples such as *quiero que vienes* instead of *quiero que vengas* etc. Apart from searches based on a particular position in the clause, the program also gives the option of using the specific place or situation in a particular context. Thus, for example, if we write *cerca* and *casa* in the two windows for lemma and select 4 as the distance, we obtain all the cases such as *cerca de mi casa*, *cerca de vuestra casa*, etc. This type of search is sensitive to the relative position in such a way that a search under the previous conditions, but with an inverted order of elements (that is, *casa* first and then *cerca*), would return examples such as *casa que está cerca*, *casa con playa cerca*, etc. instead of the ones already mentioned.

As is the case with all text corpora, this research tool is based on the retrieval of cases of a particular expression found in the corpus or indeed in any corpus that can be dynamically built. These searches cannot give us a general outline of the structure of the corpus or of the elements included in it. To fill this gap and to provide general information that could be of use in certain types of project, the CAES team prepared additional information on the corpus, this data presented in the section devoted to supplementary documents.

These documents provide general statistical data with overall information on the CAES elements but are also organised according to the learner's proficiency levels and L1. In another document we have included a list of the CAES lemmas indicating their general and partial frequency, the latter according to the learner's level and L1, as before. This, then, constitutes the general inventory of all the CAES lemmas. The information provided by this document is complemented by the list of elements and lemmas. In the latter, one can observe the connection of elements with lemmas and lemmas with elements, once again with an indication of their partial frequency according to each level and L1. Both are text documents presented in tsv format (tabseparated values) so that they can be entered in any database or spreadsheet. Since they are very large documents, they were compressed.

Section 3: Discussion of results obtained from the analysis of data gathered from CAES

In spite of its limited size, this corpus allows us to investigate different lexical and grammatical aspects which may be of interest to those scholars and professionals involved in the teaching of Spanish as L2. It is also possible to analyse the differences obtained according to the different proficiency levels and the subjects' native languages represented in the corpus. For limitations of space, we restrict ourselves here to an analysis of some of the most frequent false friends found in the different interlanguage samples. This will give us an idea of the problems students have in their learning of the Spanish vocabulary and also of the influences across languages in the learning of the target language.

3.1. False friends

In learning L2 vocabulary, false friends have always raised serious difficulties since they can be highly deceptive and confusing words. By "false friends" we understand L2 lexical items whose forms are identical or similar to words in the L1 but whose meanings are different (Ortiz, Trives and Heras 1998, Postigo 2007). False friends have been classified according to different criteria: orthographic, phonetic, semantic and contextual (Chacón Beltrán 2006). For the purpose of this study, we will mainly consider *total* versus *partial* false friends (Prado 2001: 9-14). In the case of the former, the two lexical items are very similar in form in the two languages but with two wholly different meanings. An example of this would be Spanish *librería* (*bookshop/bookstore* in English) versus English *library* (in Spanish, *biblioteca*). In contrast, we deal with *partial* false friends when we find two similar items in the two languages which share a number of denotations but not all of them, since contextual and other factors are here at play. That is the case with the English *circulation* and the Spanish word *circulación*. Both can be used to refer to the circulation of blood, water, money, ideas or the circulation of a newspaper, but while the Spanish *circulación* can also refer to the movement of cars, that is not the case in English, where we would perhaps say *road traffic* or simply *traffic*.

In this preliminary study we concentrated mainly on total false friends since they are the most distinctive and the ones that, especially at beginner levels, cause most problems for learners; however, references to partial and highly frequent false friends are also included in the survey since at times the distinction between total and partial false friends may be quite blurred. We intended, (i) to see the extent to which these lexical items were present in a learner corpus of this size, circa 600,000 words; secondly, (ii) to explore the question of whether they were really problematic or not, that is, if it is true that learners face difficulties and confusion with them; (iii) to investigate how they were actually used and what information we could gather from the corpus material; (iv) to study other phenomena that may be associated with false friends such as the use of a number of communication strategies learner may resort to in order to compensate for their deficiencies in their language system. These include, among other, word coinage and code mixing; finally (v) to examine how these lexical items varied from one L1 to another considering that although all the learners of the corpus share the same target language, that is, Spanish, they differ as regards L1s, given that the corpus contains samples of learners from six different language backgrounds.

For the purposes of this study, we restricted our analysis to three L1s, English, French and Portuguese and we considered as a starting-point a list of common false friends provided by different glossaries and dictionaries of these lexical items (Ortiz et al. 1998, Prado 2001, Postigo 2007). This means that our study should be regarded as corpus-based rather than as corpus-driven in line with the traditional distinction made by Tognini-Bonelli (2001) in this respect. Thus, the tables that follow present a list of false friends selected from the corpus for these three languages, although these lists are not intended to be totally exhaustive. The English/French/Portuguese terms are provided together with the target items in Spanish,[13] plus corpus example(s) as an illustration, and also an indication of the learner's proficiency level. Thus, for example, in the case of English we include a list of thirteen false friends, all of them quite common in the language and which certainly present problems for learners of Spanish. In the case of French and Portuguese a similar procedure was followed with a selection of ten and eleven false friends, respectively.

The findings confirm our initial assumption that false friends do cause difficulties for the learners of Spanish. Also, although students from the most basic levels (A1, A2) are the ones who tend to confuse them most often, as expected, they are present across all proficiency levels.

From the list of English terms, *move to* and *suburb* are the most frequent in the corpus. *Move to* in English shares with Spanish *mover* the meaning of movement but apart from that general sense it is also used when changing places or plans and even such as *mudarse*, *trasladarse*, *conmover*, *enternecer* are used for such meanings. Something similar happens with *suburb*. The two languages share the meaning of a place close or next to a large urban centre, yet whereas in English it is a neutral or even

[13] European Spanish is the variety of Spanish used as reference for this particular study.

a positive term, in Spanish it has negative connotations being equivalent to English *slum* or *slums*. In fact, these two lexical items would be partial false friends rather than full ones.

It is also curious to see how in some cases learners actually coin new words, taking as reference either a lexical item in the target language, such as *provienen*, probably from *provenir*, or from the native language, as with *accommodation*. At times learners make up new words by applying overgeneralisation processes; this is the case with *pilota del helicóptero* to refer to a woman helicopter pilot. This phenomenon of *word coinage* has been described in the literature as a type of communication strategy which learners use to overcome problems in their learning process. They are mainly associated with the spoken language although they can also be found in writing and are mostly of a lexical nature.[14] The examples of word coinages recorded in the corpus are numerous: *hermosidad* for *hermosura, contadora* for *contable, opinas* for *opiniones, excepcionarios* y *excepcionista* for *excepcional*, *inhibitó* for *habitaba*, *hicimos la decisión* for *tomamos la decisión*, *seriosa* for *seria*, *garantir* for *garantizar*, *reservación* for *reserva*, *ensolada* por *soleada*, *inexpectados* for *inesperados*, etc. Some of these items also reveal the highly creative nature of these learners in their use of the target language. Code-switching or code-mixing as a type of communication strategy, that is, the learner's use of the L1 and the L2 or any L3 in the construction of the same sentence, is also very common, more particularly among the learners of the lowest levels. Here are some examples: "Nosotros fuimos a la carnival de el Lago". (A2, English as L1), "Entonces fuimos a la Cloud Forest y hacemos el Zip-line y la Tarzan jump". (A2, English as L1), "Mi madre es un accountant y ella es muy buena en matemáticas". (A2, English as L1), "Me trabajo en un agency." (A1, Russian as L1), "a

[14] See Ellis (2004: 396-403) for a general overview of research in this area.

continuar su trabajo en el mundo tercera como un ambassador official de el UN". (A2, English as L1).

Table 4: Examples of English-Spanish false friends identified in the corpus

English	**Spanish**	**Corpus example**	**Students' level**
suburb	alrededores	Vivo con mi familia en la *suburbia* de Dublín.	A1
idiom	lengua, idioma	El habla cuatro *idioms* (corea, inglés, español y fortuges).	A1
firm	compañía, empresa	Trabaja en una *firma* derecha en la ciudad también.	A1
move	trasladarse	Lawrence nacio en Pincicolla, Florida en 1975 pero *movía* a Idaho cuando era muy joven.	A1
determined	decidido/a, resoluto/a	Yo la admito porque ella es *determinada,* chistosa, amable.	A2
involve	implicar	Sus deportes favoritos fueron los que *involve* la agua.	A2
large	grande	John y los otros hombres que eran en la ceremonia llevaron sombreros *largos*.	A2
realise	darse cuenta	La comé la comida misteria y *realicé* que era pollo!	B1
introduce	presentar	Estaba hablando con mi novio y decidimos ir a Mexico para *introducirlo* a la familia.	B1
conduct (an interview)	llevar a cabo	Me gustaría reunirnos en el próximo Viernes para *conducir* la entrevista.	B1
provide	proporcionar	¿Es posible todavía obtener un lugar en la resendencia universitaria o pudiese aconsejar me con unas agencias que *provienen* acomodación?	B2
accommodation	alojamiento	¿Es posible todavía obtener un lugar en la resendencia universitaria o pudiese aconsejar me con unas agencias que provienen *acomodación*?	B2
in addition	además	*En adición*, tuve que ir a la casa de mi hermano.	C1

In the case of speakers of L1 French, the words *campagne* and *se trouver* are most common. French *campagne* generally refers to the countryside or to a political/marketing campaign; the latter meaning, but not the former, is also present in Spanish. *Se trouver*, that is, *to find*/*be*, is frequently used to refer to two or more people meeting for the first time, while in Spanish we would use the verb *conocerse* for these situations. Note how on this occasion most of the examples recorded correspond to A2 learners although we also find examples at other levels even at the C1 level.

Table 5: Examples of French-Spanish false friends identified in the corpus

French	**Spanish**	**Corpus example**	**Students' level**
campagne	campiña, campo	Visitamos a Oxford, Dublin y la *campaña* irlandesa.	A2
se trouver	conocerse	*Encontramos* en 2001 cuando veni en Pariz por mis estudios.	A2
civilisation	cultura	Vivir en Buenos Aires me permitiría también de conocer su *civilización* y costumbres.	A2
cuisiner, faire la cusine	cocinar	A veces *hago la cocina* en casa.	A2
sentiment	impresión, intuición	antes de este viaje mama tenia un *sentimiento* que vaya a encontrar su marido alli en paris o en un sitio alli.	A2
concours	concurso	Cuando el solo tenía 16 años, fue en la *competición* de X Factor.	A2
période, saison	temporada	Espero que tiene ja habitaciones libres porque es la alta *perioda*.	A2
large	ancho/a	Mi maleta es muy *larga* y de plástica roja.	B1
succès	éxito	esperé sin *suceso* la salida de mi bolso a la llegada	B1
entendre	oir	Soy madame xxxx habia *entendido* buenas noticias de vuestra compañia ...	C1

With regard to Portuguese-Spanish false friends, we find quite a long list although our survey has reduced this to a small number; *romance* is clearly the most common in the corpus. It refers to a novel in Portuguese while in Spanish it is associated with a type of poetic composition or a love story.

Table 6: Examples of Portuguese-Spanish false friends identified in the corpus

Portuguese	**Spanish**	**Corpus example**	**Students' level**
romance	novela	los buenos libros, siendo mis preferidos, los *romances* y biografías.	A1
procurar	buscar	Después de estas vacaciones, tengo que repor el diñero que he gasto, por eso estoy *procurando* trabajo.	A1
aula	clase	Yo tengo *aula* de espanhol.	A1
brincar	bromear/jugar	Mi mamá no trabaja y le gusta mucho *brincar* y pasear con sus	A1

		nietos.	
combinar	quedar, concertar	No puedo llegar la hora *combinada*.	A1
		después encontrarme con mis padres en el lugar *combinado*.	A2
sucesso	éxito	Su marido hico muchas músicas de *suceso* en Brasil.	A2
balcâo	mostrador	Ya estuve muchas veces en el *balcón* de la compañía y no hay nada con mi nombre.	B1
		Hice una queja en el *balcón* de su compañía en el aeropuerto describiendo el equipaje.	B1
contestar	manifestarse, protestar	Escribo les para *contestar* sobre mi equipaje que no ha venido junto a mí en el viaje.	B1
lecionar	enseñar, impartir clase	Quantos professores *lecionan* en cada curso?	B2
histórico	historial	Me gradué periodista en la católica en 2010 y tuve un *histórico* universitario lleno de conquistas.	B2
passar	tener lugar, acontecer	pelicula esa *se pasa* en una barrio de Salvador de Bahía que nombra la película.	C1
		La historia *se pasa* en Brasil en 2012.	B1

From a pedagogical perspective, these findings reveal that false friends deserve special attention in the language learning and teaching processes since they may hinder communication and they may even lead to confusion and misunderstanding. Furthermore, they may be central in activities where translation and mediation processes and/or strategies are involved. Teachers should draw students' attention to the existence of such items, in particular those which seem to be the most common. The corpus provides useful information on how our learners process the language and also shows how they respond to learning difficulties. As mentioned at the beginning of this work, corpora data allow us to see what learners actually do with the language, how they deal with difficulties and their creativity. It would be almost impossible to obtain this kind of information without a resource such as CAES. Corpora examples could also be used as good illustrations and hence as starting-points in dealing with these issues in the classroom or in learning materials, since they are samples of language production which have not been adapted or simplified, although teachers could also resort to other pedagogical resources such as visualisations, language games, matching and self-discovery activities as effective techniques for the presentation and practice of these particularly troublesome lexical items (Roca Varela 2015). In conclusion, our findings confirm that when teaching vocabulary, second language teachers should pay attention not only to the meaning of the word but also to its spelling, correct pronunciation, collocations, register, context and actual use (Pérez Basanta 1999).

4. Final reflections and questions for further consideration

This chapter has described the CAES project from its origin to its current state. It has also given an account of the different steps and stages followed for its completion. Attention has also been paid to the problems and difficulties found not only in its design and compilation but also in its annotation and disambiguation, given that this itself might be of use to other scholars engaged in similar tasks. In its initial phase the CAES was conceived as an open corpus, that is, as a dataset that could grow in size, incorporating new samples from more learners and incorporating data from students from more L1s. It is within our plans to endow the corpus with an error tagging system which would allow teachers and researchers to focus on this area, thus offering a great deal of potential pedagogical uses. Also part of future developmental plans is the inclusion of spoken samples to complement the existing written ones, although we are aware of the complexities that this implies in terms of the collection and transcription of data.

The third part of this chapter has focused on applications of CAES, not only for linguistic research but also for the language teaching field. We believe there is still great scope for further development on these lines, since the corpus is not only of potential help to teachers in the planning of their lessons and in the search of materials, but might also constitute a rich source of material for those designing and implementing resources for the learning of Spanish as a foreign language. Without underestimating other similar Spanish learner corpora, we believe CAES has filled an important gap in learner corpus research in line with well-known international projects such as ICLE (International Corpus Learner English Corpus), developed at the Centre for English Corpus Linguistics of the Catholic University of Louvain.

References cited

Adolphs, Svenja. 2008. *Corpus and Context. Inestigating Pragmatic Functions in Spoken Discourse*. Amsterdam & Philadelphia: John Benjamins.

Adolphs, Svenja, Carter, Ronald A. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. New York: Routledge.

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

Alvarez López, Fátima. 2005. Corpus de textos académicos producidos por estudiantes universitarios extranjeros. *LINRED (Lingüística en la red)* 3. < http://www.linredes/informacion_pdf/informacion6_18072005.pdf>

Bailini, Sonia Lucia. 2013. SCIL: A Spanish corpus of Italian learners. *Procedia - Social and Behavioral Sciences* 8: 542-549.

Biber, Douglas, Conrad, Susan, Reppen, Randi. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: CUP.

Braun, Sabine, Kohn, Kurt, Mukherjee, Joybrato. (eds). 2006. *Corpus Technology and Language Pedagogy*. Frankfurt, Germany: Peter Lang.

Campillos Llanos, Leonardo. 2014. A Spanish oral learner corpus for computer-aided error analysis. *Corpora* 9 (2): 207-238.

Cestero Mancera, Ana, Penadés Martínez, Inmaculada, Blanco Canales, Ana, Camargo Fernández, Laura, Simón Granda, José. 2001. Corpus para el análisis de errores de aprendices de E/LE (CORANE). In *Actas del XII Congreso de ASELE tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE*, Ana Gimeno Sanz (coord). Centro Virtual Cervantes, 527-534.

Chacón Beltrán, Rubén. 2006. Towards a typological classification of false friends (Spanish-English). *Revista Española de Lingüística Aplicada* 19: 29-39.

Council of Europe. 2001. *Common European Framework of Reference for Languages. Learning, Teaching, Assessment*. Cambridge: CUP.

De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3(1): 59-80.

Ellis, Rod. 2004. *The Study of Second language Acquisition*. Oxford: OUP.

García, Marta. 2005. Corpus de Conversaciones en Español como Lengua Extranjera. *LINRED (Lingüística en la red)* 3. < http://www.linredes/informacion_pdf/informacion11_24012006.pdf>

Granger, Sylviane. 1998. *Learner English in Computer*. London & New York: Longman.

Granger, Sylviane. 2002. A Bird's-eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign language Teaching,* Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson (eds), 3-33. Amsterdam & Philadelphia: John Benjamins.

Granger, Sylviane. 2008. Learner corpora. In *Corpus Linguistics: An International Handbook,* Anke Lüdeling and Merja Kytö (eds), 259-274. Berlin, New York: Mouton de Gruyter.

Gutiérrez Quintana, Esther. 2005. Análisis de errores en la producción escrita de italianos aprendices de E/LE. *ELUA* 19: 223-242.

Krashen, Stephen. 1988. *Second Language Acquisition and Second Language Learning*. New York: Prentice Hall.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* 82, Stockholm 4-8 August 1991, Jan Svartvik (ed), 105-122. Berlin & New York: Mouton de Gruyter.

Lozano, Cristóbal. 2009. Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus. In *Representational Deficits in SLA*, Neal Snape, Yan-kit Ingrid Leung and Michael Sharwood-Smith (eds), 127-166. Amsterdam: John Benjamins.

Lozano, Cristóbal, Mendikoetxea, Amaya. 2013. Learner corpora and second language acquisition. The design and collection of CEDEL2. In *Automatic Treatment and Analysis of Learner Corpus Data,* Ana Díaz-Negrillo, Nicolas Ballier and Paul Thompson (eds), 65-100. Amsterdam: John Benjamins.

Lüdeling, Anke, Kytö, Merja. (eds). 2008. *Corpus Linguistics: An International Handbook*. Berlin, New York: Mouton de Gruyter.

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

Mitchell, Rosamond, Domínguez, Laura, Arche, María J., Myles, Florence, Marsden, Emma 2008. SPLLOC: A new database for Spanish second language acquisition research. *EUROSLA* 8: 287-304.

O’Donnell, Mick. The UAM CorpusTool: Software for corpus annotation and exploration. In *Applied Linguistics Now: Understanding Language and Mind/ La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*, Bretones Callegas, Carmen, Fernández, Francisco, Ibáñez, J. Ramón, García, Mª Elena, Cortés, Mª Enriqueta, Salaberri, Mª Sagrario, Cruz, Mª Soledad, Perdú, Nobel, Cantizano, Blasina (eds), 1433-1447. Almería: Universidad de Almería.

O'Keeffe, Anne, McCarthy, Michael, Carter, Ronald. 2007. *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: CUP.

Ortiz de Urbina, Cantera, Trives, J. Ramón, Heras Díez, Francisco. 1998. *Diccionario francés-español de falsos amigos*. Alicante: Universidad de Alicante.

Padró, Lluis, Stanilovsky, Evgeny. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA. Istambul, May 2012.

Pérez Basanta, Carmen. 1999. La enseñanza del vocabulario desde una perspectiva lingüística y pedagógica. In *Lingüística aplicada a la enseñanza de lenguas extranjeras*, Sagrario Salaberri Ramiro (ed), 262-306. Almería: Universidad de Almería.

Pérez Paredes, Pascual, Cantos-Gómez, Pascual. 2004. Some lessons students learn: self-directory and corpora. In *Corpora and Language Leaners*, Guy Aston and Stewart Dominic (eds), 247-257. Amsterdam & Philadelphia: John Benjamins.

Postigo Pinazo, Encarnación. 2007. *Diccionario de falsos amigos: inglés-español*. Madrid: Verba.

Prado, Marcial. 2001. *Diccionario de falsos amigos inglés-español*. Madrid: Gredos.

Reppen, Randi. 2010. *Using corpora in the language classroom*. Cambridge: CUP.

Roca Varela, María Luisa. 2015. *False Friends in Learner Corpora. A Corpus-Based Study of English False Friends in the Written and Spoken Production of Spanish Learners*. Linguistic Insights Series. Bern: Peter Lang.

Römer, Ute. 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics* 31: 205-225.

Seidlhofer, Barbara. 2002. Pedagogy and local learner corpora: Working with learning-driven data. In *Computer Learner Corpora, Second Language Acquisition and Foreign language Teaching*, Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson (eds), 213-234. Amsterdam & Philadelphia: John Benjamins.

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

Tracy-Ventura, Nicole. 2008. Spanish Learner Language Oral Corpus Project (SPLLOC1). A New Corpus of Oral L2 Spanish. Avaliable at <http://www.sploc.ac.uk> and at Talkbank <http://www.talkbank.org>

Tzu-Ju, Lin. 2005. Corpus de textos escritos por universitarios taiwaneses estudiantes de español. *LINRED (Lingüística en la red)* 3. < http://www.linredes/informacion_pdf/informacion7_13092005.pdf>

WOSLAC Research Group. 2007. Proyecto CEDEL2. Universidad Autónoma de Madrid. Available at <http://www.uam.es/woslac/cedel2.htm>

APPENDIX

Table 7: Participants' distribution according to their L1 and proficiency level

	Arabic	Chinese	French	English	Portuguese	Russian
A1	599	189	132	77	494	66
A2	364	100	88	344	257	58
B1	232	69	85	127	123	41
B2	99	15	48	41	99	11
C1	48	0	18	26	28	0

Table 8. Participants' distribution according to their country of origin

Countries	Elements	Sample units
Afghanistan	20 052	52
Algeria	10 029	32
Australia	3 343	6
Austria	627	1
Belgium	4 166	9
Belarus	446	1
Bolivia	587	1
Brazil	143 926	319
Burkina Faso	325	1
Canada	2 550	5
China	53 207	127
Colombia	194	1
Denmark	314	1
Egypt	4 601	10
France	39 317	92
Germany	896	2
Greece	416	1
Guinea	927	3
Indonesia	293	1
Irak	713	2
Ireland	18 680	38
Italy	420	1
Japan	257	1
Jordan	7 137	21
Kazakhstan	480	1
Kuwait	1 638	4
Lebanon	11 171	26
Morocco	97 425	312
Mauritania	444	1
Mexico	1 364	1
Moldova	278	1
Monaco	266	1
Pakistan	277	1
Philippines	316	1

Portugal	15 947	31
Russia	18 908	62
Saudi Arabia	454	1
Singapore	412	1
Syria	30 289	70
South Africa	673	1
South Korea	1 449	4
Spain	1 588	2
Switzerland	841	2
Taiwan	382	1
Tunisia	4 457	16
Turkey	148	1
Turkmenistan	332	1
Ukraine	575	2
United Arab Emirates	154	1
United Kingdom	3 978	9
United States	65 211	139
Venezuela	390	1
Other	448	1

Table 9. Participants’ distribution according to their proficiency level

Proficiency level	Elements	Sample units
A1	155 458	526
A2	178 834	421
B1	116 520	252
B2	80 556	162
C1	42 350	62

Table 10. Participants’ distribution according to their L1

L1	Elements	Sample units
Arabic	168 231	497
Mandarin Chinese	53 163	128
French	58 412	143
English	106 968	227
Portuguese	165 231	361
Russian	20 713	67

Table 11. Participants’ distribution according to their gender

Gender	Elements	Sample units
Male	207 992	521
Female	365 726	902

Table 12. Studies completed by participants

Studies completed	Elements	Sample units
Primary	72 961	205
Secondary	48 226	127
University	375 602	908
Other	76 929	183

Table 13. Participants’ contacts in Spanish-speaking countries

Contacts	Elements	Sample units
Friends	182 867	409
Friends & relatives	48 737	118
Relatives	33 389	96
No	285 592	742
Other	23 133	58

Table 14. Participants’ distribution according to age

Age	Elements	Sample units
>=15 - <=21	200 696	498
>=22 - <=30	187 311	466
>=31 - <=40	76 674	196
>=41 - <=60	83 750	198
>=61	25 287	65

Table 15. Participants’ starting age in the study of Spanish

Starting age	Elements	Sample units
<15	156 393	404
>=15 - <=21	178 064	417
>=22 - <=30	127 386	315
>=31 - <=40	51 828	133
>=41 - <=60	51 346	131
>=61	8 701	23

Table 16. Number of months participants have been engaged in the study of Spanish

Months	Elements	Sample units
<2	118 842	339
>=3 - <=6	104 203	300
>=7 - <=12	99 429	243
>=13 - <=24	124 875	277
>=25 - <=36	54 346	121
>=37	72 023	143

Table 17. Number of months participants have stayed in Spanish-speaking countries

Months	Elements	Sample units
0	347 288	911
>=1 - <=3	137 143	328
>=4 - <=6	42 193	91
>=7	47 094	93