Paper submitted to EUROSLA Yearbook 8 (2008):

SPLLOC: A new corpus for Spanish second language acquisition research

Rosamond Mitchell (University of Southampton)
Laura Domínguez (University of Southampton)
María Arche (University of Southampton)
Florence Myles (University of Newcastle)
Emma Marsden (University of York)

SPLLOC: A new corpus for Spanish second language acquisition research

Abstract

The contribution of Spanish to the field of SLA continues to grow (Lafford and Salaberry 2003, Montrul 2004), and the need for good L2 Spanish datasets is becoming increasingly evident. In this paper we introduce a newly created Spanish Learner Language Oral Corpus (SPLLOC), describing the rationale underlying the corpus design and methodology used for its construction.

This project applying CHILDES tools to L2 Spanish follows successful creation of a collection of French L2 oral corpora (Rule et al 2003), already available at www.flloc.soton.ac.uk . Creating a successful oral corpus is costly and available corpora are often built somewhat opportunistically from available material rather than designed in a balanced way to facilitate SLA research. The SPLLOC corpus has been designed to fill the existing gap in Spanish L2 resources and also to support a focussed research agenda investigating learner development with respect to the verb phrase, clitic pronouns, and word order, from an interface perspective.

Introduction: Origins of corpus research in SLA

The use of electronic learner corpora is making an increasing contribution to both L1A and SLA studies, though L1A research has had a considerable head start. Since the 1980s, the CHILDES project in particular, with its range of L1 and early bilingual acquisition corpora in various languages, has made a large amount of naturalistic data available electronically to the L1 acquisition research community, along with a transcription system (CHAT) and a range of analysis software (CLAN: MacWhinney 2000). Use of these datasets and procedures is routine in much L1A research, though tensions remain between the 'hypothesis-finding' descriptive orientation which is relatively common in corpus based research, and the 'hypothesis-testing' orientation of the generative linguistic tradition in particular (Barlow 2005: 344, Myles 2007: 380).

In SLA, there is a longstanding and continuing tradition of longitudinal case studies of individual learners, which have typically carried out analyses by hand of development over time, using small corpora of individual interviews either alone or alongside more formal tests (e.g. Huebner 1983, Sato 1990, .Lardière 1998, Herschensohn 2003: the first three of these studies relate to English as L2, the last to French as L2). A review and discussion of the earlier studies of this type can be found in Perdue 1993a (14-38); they include both more descriptive, hypothesis-finding studies (e.g. Huebner 1983), and more theoretical, hypothesis-testing studies (e.g. Lardière 1998).

There has also been a small number of large scale projects which have collected longitudinal naturalistic data from larger numbers of learners, and/ or involving several languages as both L1 and L2. The best known of these is the European Science Foundation project, titled *Second Language Acquisition by Adult Migrants*, which collected longitudinal data from 40 informants from diverse language backgrounds during the 1980s, as they learned a range of European languages as target L2s (Dutch English, French, German, Swedish: Perdue 1993a, 1993b). The informants were audiorecorded

at regular intervals for a period of more than two years, undertaking oral tasks such as story retelling, personal life history interviews, and role plays of service encounters. This project used computers as a means for housing, sharing and archiving data (i.e. transcriptions of learner talk), while early analyses to trace the morphosyntactic development of the so-called "Basic Variety" were conducted by hand (Feldweg 1993). As suitable commercial software became available, a range of programs were used for text searching, creation of frequency tables and concordances (Feldweg 1993: 117). However the main reports produced from this early L2 corpus work have been qualitative in nature (see e.g. papers in Klein and Perdue 1992; Perdue 1993b).

By the 1990s, mainstream corpus linguistics was demonstrating the power of bottom up computerized analyses of large scale collections of electronic texts to reveal new insights into lexical and syntactic patterning of natural languages, through projects such as the British National Corpus or Cobuild/ Bank of English (see Hunston 2002, Conrad 2005 for overviews). Concordancing software for English such as Wordsmith Tools (Scott 1997) and Part of Speech (POS) taggers such as CLAWS were showing that a range of repetitive data analysis tasks could be automated; active experiments were being conducted with parsers which could address specific areas of natural language grammar (see discussion in Conrad 2005 :395).

These developments in corpus linguistics attracted the attention of some SLA researchers, and attempts were made during the early 1990s to produce and disseminate software which would address the special problems of comprehensively tagging and analyzing L2 data. The two best known programs of the time were COALA (Pienemann 1992) and COMOLA (Jagtman and Bongaerts 1994). However these programs did not succeed in attracting significant numbers of users from the SLA community, and have now been discontinued. In the same period an increasing number of L2 corpora were also being created but were generally not publicly accessible (e.g. early versions of the Progression corpus

of beginner spoken L2 French created at the University of Southampton, UK: Mitchell and Dickson 1997; or the InterFra corpus of advanced spoken L2 French created at the University of Stockholm, Sweden: Bartning 1997).

The publication of the International Corpus of Learner English (ICLE: Granger Dagneaux and Meunier 2002) attracted considerable attention in the early 2000s, and gave fresh impetus to corpus based L2 studies, much of it of a bottom-up, hypothesis-finding character. This corpus contains c2.5 million words of advanced written L2 English, mainly argumentative academic writing produced by university students from 11 different L1 backgrounds. The ICLE team experimented with POS tagging (Granger 2002: 17-18; Barlow 2005: 339), but the published version of the corpus is not POS tagged. The team has however produced a semi-automated routine for tagging learner errors, which has been used in a number of error analysis studies using the corpus (e.g. Dagneaux Denness and Granger 1998). Other researchers who have carried out error tagging on learner corpora include Milton and Chowdhury (1994), who annotated a corpus of written L2 English produced by Hong Kong learners; and the researchers responsible for a transcribed corpus of spoken L2 English created in Japan (Tono 2001; Izumi et al. 2004). This Standard Speaking Test (SST) corpus has been created by recording Japanese L1 learners undertaking a 15 minute oral interview test including storytelling, role play, picture description and informal conversation (Tono 2001). This is one of the few large learner corpora which prioritise spoken language; like the smaller EVA corpus at the University of Bergen (Hasselgren 2002), it is based on learner performances during oral examinations. Lexical and grammatical errors have been hand tagged using a 45-item error tagset (Izumi et al.2004: 35-37); in addition, the researchers have been experimenting with an automated tagging procedure, though this cannot yet be used reliably to tag a good proportion of learner errors (ibid 37-45).

Thus ICLE and similar L2 corpora, both spoken and written, have started to popularise the idea of corpus based SLA research, and also the idea of comparative work comparing parallel L2 and L1 corpora. (For recent reviews see Granger 2002, Barlow 2005, Myles 2005a, 2007). Researchers with theoretical interests in frequency based approaches to SLA have also shown increasing interest in the analysis of learner corpora (e.g. Ellis 2002, Ellis and Ferreira 2007; Crawford et al 2007). However the analysis tools available have been fairly limited, and much work using these corpora has relied on combinations of lexical searches, frequency counts, concordancing and hand tagging.

For example Hasselgren (2002) has searched the EVA spoken L2 English corpus for "small words" (a range of fillers, vague language and discourse markers) which she has interpreted as markers of oral fluency. The study compared the use of small words by native speakers and by learners at two different proficiency levels, and Hasselgren argues that this work allows her to produce sharper descriptions of fluency with potential for use in language testing. Aijmer (2002) has analysed modality markers in several subsections of ICLE, again using a combination of lexical searching, frequency counting and hand tagging. The L2 data is compared with L1 data from a similar written English corpus, and conclusions drawn about mother tongue influences and L1 transfer in explaining patterns of 'overuse' of modal auxiliaries etc. Altenberg and Granger (2001) similarly analysed the uses of various forms of the verb 'make' in two different subsections of ICLE using Wordsmith Tools, explaining patterns of over- and underuse by comparison with an L1 parallel corpus, in terms of interlingual influences. They describe their methodology as "a combination of fully automatic analysis and minute manual investigation" (176). Nesselhauf (2003) has studied the difficulties of advanced learners with verb-noun collocations in a selection of L2 English essays drawn from the German L1 subsection of ICLE. In order to do this she "manually extracted all verb-object-noun combinations" from the dataset, before classifying the degree of restriction of the collocations and evaluating their acceptability in English

(again through comparisons with L1 corpus evidence). Two papers in Aston et al. (eds) (2004) however include a POS tagging element to analyse learner data, combined with other analysis techniques.

CHILDES-based corpora for SLA research

As can be seen from the foregoing brief account of the emergence of corpus based SLA research, there is a growing international awareness of the potential of learner corpora to address a range of issues in SLA, and interest/ willingness to use them. However while there has been awareness of the great increase in usefulness of POS tagged or error tagged corpora, and a range of experiments with tagging programmes for SLA data, manual or semi-automated tagging has predominated where morphosyntactic analysis has been undertaken. Also, early L2 corpora tended to be privately held by the research teams that created them, and thus the heavy upfront investment made by these teams in data collection, editing and/or transcription (in the case of oral learner corpora) did not benefit a potential community of secondary users.

Rutherford and Thomas (2001) argued that instead of attempting to undertake independent software development, the best way forward for the SLA research community was to re-examine the procedures and tools of the L1A CHILDES project, and explore their potential for SLA corpus development and analysis. The CHILDES project already possessed a robust set of transcription conventions (CHAT), and a range of analysis software, including programs to calculate frequencies and create concordances. Above all, the CLAN suite included POS taggers for a range of languages, which while not created to handle SLA data, could potentially be adapted to take account of interlanguage features such as neologisms, indeterminate forms, and loanwords. The CHILDES software has the great merit of being freeware, not proprietary software, and the only condition of using it is willingness to make datasets created using CHAT and CLAN available freely to the research community via either the CHILDES or

Talkbank websites. (Talkbank is a depository for corpora other than child L1A datasets, and already houses a number of L2 datasets including e.g. the ESF dataset.)

A number of researchers have started to use aspects of CHILDES in this way, including Malvern and Richards (2002), who conducted lexical analysis of a CHAT-transcribed corpus of spoken L2 French, and Housen (2002) who used a range of CLAN programs to support the analysis of verb morphology in a large cross-sectional corpus of spoken L2 English. Researchers in this tradition have also shown greater interest in a hypothesis-testing orientation, e.g. Housen's work testing the claims of the Aspect Hypothesis (Andersen and Shirai 1996). At the University of Southampton UK, the team who had in the 1990s created the longitudinal Progression corpus of spoken L2 French, had been conducting a range of hand-done analyses with this corpus on e.g. the role of chunks in early SLA, and the emergence of verb morphosyntax (Myles et al. 1998, 1999). From 2001 onwards, this team has collected further spoken French L2 corpora from intermediate and advanced learners, some longitudinal some cross-sectional, and the entire collection (including the converted Progression corpus) has been made available in CHAT to the research community via the French Learner Language Oral Corpora website www.flloc.soton.ac.uk, and also via TalkBank. (These initiatives are described in Rule et al. 2003; Rule 2004; Myles and Mitchell 2005; Myles 2007). Other researchers have contributed further corpora to the FLLOC collection, which now comprises c2,000,000 words of spoken L2 French. Much of the collection has been POS tagged using a French version of the CLAN MOR programme (Parisse and Le Normand 1997, 2000), and new lexical and morphosyntactic analyses of the datasets are appearing which exploit this enhancement of the datasets to address more theoretical issues relating to interface between syntax, morphology and lexis, and the acquisition of functional categories (e.g. Myles 2005b, Rule and Marsden 2006, David 2007). The use of CHAT and CLAN are allowing analyses to be undertaken on a larger scale, comparing learners at different stages of L2 development, comparing L1 and L2 speakers, and linking different types of morphosyntactic and

lexical analysis. Overall this experience has led the authors to the view that well planned oral corpora with learners undertaking a good variety of speaking tasks, can make a distinctive empirical contribution to the testing of specific claims about acquisition processes and thus to the advancement of language learning theory.

SPLLOC: A new Spanish L2 oral corpus

In following sections of this paper we describe the design and creation of a new corpus of spoken L2 Spanish, the Spanish Learner Language Oral Corpus (SPLLOC). This corpus is one major outcome of a two-year project (2006-2008) on the acquisition of L2 Spanish by L1 English learners, which is being undertaken as a collaboration between the universities of Southampton, Newcastle and York, UK, and funded by the UK Economic and Social Research Council (Award no. RES-00023-1609).

Spanish is widely learned as an L2, and there is an actively developing research literature on its acquisition; the language offers insights into a range of theoretically interesting issues in SLA (see recent reviews in Lafford and Salaberry 2003, Montrul 2004), especially given a number of parametric contrasts between Spanish and English, the L1 of the SPLLOC project participants. Spanish exhibits a specific cluster of properties, such as null subjects, subject-verb inversion and rich inflectional morphology not found in English. This makes English and Spanish a valuable language pair for studying acquisition from a linguistic perspective. Publicly accessible electronic databases of child language already exist for L1 Spanish (see the CHILDES collection); an unpublished survey by Myles et al. (2004) showed that Spanish SLA researchers are aware of the potential of information technology and corpus based methodologies for studying learner Spanish. However, as yet there is no generally available resource of this type for L2 Spanish.

The new SPLLOC corpus will complement existing corpora of learner English and learner French, and provide a resource for L2 Spanish which will be freely available for use by the SLA research community. As well as corpus creation, the research team are undertaking a short programme of substantive research aiming to contribute to current theoretical debates about interfaces between syntax, morphology and pragmatics, and their role in second language acquisition. This work involves investigating the acquisition by English L1 learners of Spanish word order, the acquisition of Spanish clitic pronouns, development of the verb phrase, and lexical development. It will not be described in fuller detail here but other outputs are already available introducing aspects of this programme (Dominguez and Arche 2007a, 2007b, 2007c; Marsden and David forthcoming).

Designing the SPLLOC corpus

The design of the new SPLLOC corpus has been guided in general terms by a) experience in building corpora/ corpus collections to support SLA research for other languages; b) a commitment to creating an open resource which would be maximally useful both to its creators, but also to a public of secondary users, for purposes not completely pre-determined; c) the need to deliver quality within the relatively limited resources of a two-year project. These general considerations led to adoption of a number of key principles which underpin the SPLLOC design.

Principle 1: Focus on speech

It was decided the corpus would prioritise collection of semi-naturalistic L2 speech data, rather than written data, for the reason that spontaneous speech produced in face to face interaction is likely to provide more direct evidence about the state of the L2 learner's underlying interlanguage system. In producing written data, L2 learners may reflect to a considerable extent on their performance and undertake self-correction, using metalinguistic knowledge including explicit 'rule' knowledge. In

producing L2 speech under the pressure of real time face to face communication, this type of monitoring and self-correction is minimized.

Principle 2: Variety of genres

Rather than collecting L2 speech data of a single type (e.g. from oral proficiency interviews), it was decided that project participants would undertake a range of semi-naturalistic oral activities in different genres (narrative, interview and picture description, peer discussion). Learner speech is well known to contain considerable variability, e.g. in use of target language morphology, and learners are also prone to avoid in speech production areas of the target language system where they feel insecure or dysfluent. To some extent these are inescapable features of oral production data, but they create difficulties in estimating and interpreting what learners really know (Chaudron 2003: 767). In designing the SPLLOC corpus an attempt was made to minimize these problems both by eliciting a substantial speech sample from individual participants (40-60 minutes per learner), and by using open-ended tasks in a range of different speech genres (interview, narrative, discussion), and with varying interlocutors (research team members and fellow L2 learners).

Principle 3: Balance of open ended and focused tasks

In addition to the more open ended tasks described above, it was decided to have the same learners complete a small number of more focused tasks, relevant to the substantive research agenda of the project. Activities prompting learners' production and/or interpretation are widely used in linguistically oriented SLA (see e.g. Gass and Mackey 2007: 71-107 for a recent overview). They address problems of learner avoidance of particular target structures of interest to researchers, and also allow researchers to infer "not only what learners know is correct in the second language, but also what learners know is not possible" (Gass and Mackey 2007: 73). Inclusion of focused elicitation tasks alongside more open ended tasks being undertaken by the same L2 participants also creates the possibility for triangulation

across different data types, when investigating particular morphosyntactic areas. However the number of focused tasks which can be administered to participants has to be strictly limited on practical grounds. In the SPLLOC case, three focused tasks were designed, relevant to the team's theoretical interests in the acquisition of Spanish clitic pronouns, and in word order issues relevant to the syntax/ pragmatics interface.

Principle 4: variety of learner levels

In order to maximize the usefulness of the corpus to study development in L2 Spanish, it was necessary to include learners at different proficiency levels, plus small numbers of age-matched native speakers of Spanish who would undertake the same elicitation tasks. Because of the short timescale of the project, however, it was necessary to adopt a cross sectional rather than a longitudinal design. All learner participants were L1 English speakers, undergoing formal instruction in L2 Spanish, and 20 learners were located at each of 3 levels. The levels were differentiated by age and number of years of instruction, rather than by any formal independent language test. While there is variability among the learners at each level defined in this way, in terms of their L2 Spanish proficiency, it is not sufficient to jeopardize the overall design. In addition, five age-matched native speakers of Spanish undertook the different sets of tasks at each level. Details of the learner participants and their educational background are shown in Table 1.

Table 1: SPLLOC Project Participants (L2 learners)

L2 Spanish level	Typical age	Approx no hours of Spanish instruction	Educational level (English system)	Approx position on Common European Framework
Beginners N = 20	13-14 years	c 180 hours	Lower secondary school (Year 9)	A2
Intermediate N = 20	17-18	c 750 hours	Sixth form college (Year 13)	B1-B2
Advanced N =	21-22	C 895 hours + year	University (Year 4)	C1

20		abroad		

Principle 5: Use of CHILDES procedures

It was decided that the learners' spoken L2 output would be captured as digital audio files, and these would be transcribed using CHAT conventions, to facilitate subsequent analyses using the CLAN suite of analysis programs. The resulting transcripts would also be POS tagged using the CHILDES MOR program, which would be adapted as necessary to take account of interlanguage features.

Principle 6: Accessibility

The complete dataset (digital audio files, CHAT transcripts, POS tagged files) would be made available to the research community through a specially created database and web interface www.splloc.soton.ac.uk, and also eventually through Talkbank deposit.

Task development

During the early months of the project the set of tasks to be used with participants at different levels was developed, piloted and evaluated. A number of the tasks were adapted from previous SLA studies, while some were specially developed for the project. The tasks eventually selected for SPLLOC data collection are briefly described below.

Narrative task: "A Monster Mistake"

This task was based on a sequence of pictures taken with permission from Hunt (2003), which tell the story of a family on holiday by the shores of Loch Ness, who create a fake monster and deceive the public. A member of the research team narrated the story to individual participants, following an

agreed script; the learners then re-told the story with the support of the set of pictures. This picture sequence had previously been used to collect L2 French narratives (in several components of the FLLOC project: the Progression, Linguistic Development and Newcastle corpora), and was known to be suitable for learners at beginner and intermediate level.

Narrative task: "Modern Times"

Sequences from the Charlie Chaplin film "Modern Times" have been used successfully by SLA researchers to elicit narrative data from adult learners in a number of research projects including the ESF project (Perdue 1993a) and the FLLOC project (Newcastle corpus). A sequence from the film was trialled for use in SPLLOC data collection, because of concern that the "Monster Mistake" narrative might be subject to a ceiling effect (i.e. fail to show the full narrative abilities of advanced learners). However piloting showed that the "Modern Times" narrative did not work well with the relatively young beginners used in the study, and consequently it was used only with the advanced group of SPLLOC participants. The procedure followed was to show participants a short (5 minute) sequence from the film (the bread-stealing sequence), and ask them to re-tell the story, with support from a set of still images and vocabulary list.

Picture description and interview task

All participants undertook a version of this task individually, with a member of the research team as their interlocutor. The interview for intermediate and advanced learners was in three parts. In the first part, the learners were shown a series of six stimulus photographs (of young British people on holiday in Mexico) and asked to describe the various scenes and activities. In the second part they were asked to find out as much additional information as they could about the characters shown in the pictures, by asking questions. In the third part, the researcher asked the learner a range of questions about their current interests, their past activities, and their plans for the future. The interview for beginners was

very similar but the stimulus pictures showed British adolescents undertaking leisure activities at home/on holiday in Europe.

Discussion task

This task was developed in two slightly different versions for use with intermediate and advanced learners, and took the form of a pair discussion between two learners. The task was included to elicit expressions of opinion and preferences, and also evidence of learners' turntaking, initiation and repair skills when talking to an interlocutor of similar L2 level. The pairs were offered a choice of four discussion cards, setting out a series of claims/ arguments on different current topics. The discussion cards were modeled on a set first developed by Dippold (2007), for her study of argumentation in L2 German, and were adapted with permission. Having chosen one topic, the pair were asked to discuss these arguments and rank them in order of importance. (Two examples are shown as Figure 1.)

INSERT FIGURE 1 ABOUT HERE

Clitic interpretation task

This first focussed task was designed to investigate learners' knowledge of clitic object pronouns. The idea for the task was adapted with permission from the research of Franceschina (2003). It comprised 32 short multiple choice items created using Macromedia Authorware, and was administered on a laptop computer. Each item presented the learner with a stimulus Spanish sentence including a clitic object pronoun marked for gender and number, systematically sampling the following options:

- Canonical feminine: -*a* ending (e.g. calculadora 'calculator')
- Canonical masculine: -o ending (e.g. teléfono 'phone)
- Non canonical: no –a/-o ending (e.g. lápiz)

Collocation: Proclitic (as in conjugated verbs) vs. enclitic (as in infinitives).

Each sentence was presented both orally and in writing. Four nouns of varying number/gender were offered as possible responses and the learner was asked to select the one which matched the stimulus.

This task was administered to all beginner, intermediate and advanced learners. Sample screen shots are included as Figure 2.

INSERT FIGURE 2 ABOUT HERE

Clitic production task

This focused elicitation task was designed to 'push' learners to produce orally a range of Spanish clitic pronouns, and like the interpretation task, consisted of 32 items presented on a laptop computer. For each item, the learners saw a stimulus sequence of two pictures, and heard and read a questionabout the activities shown. The items were designed to elicit a range of object pronouns varying by number and gender (canonical and non canonical), and in different collocational contexts. This task is adapted with permission from one devised by Cadierno (1993); sample items are included as Figure 3.

INSERT FIGURE 3 ABOUT HERE

Word order task

The final focused task was a 28 item acceptability judgement task which was specially developed for the SPLLOC project. The object of the task was to document learners' knowledge of word order variation at the syntax/ pragmatics interface. It was administered as a pencil and paper multiple choice task. Learners were offered a situation (described in English) plus a question in Spanish and three alternative responses. They had to select the correct response. A sample item is included as Figure 4.

INSERT FIGURE 4 ABOUT HERE

Data collection

Suitable learner participants were identified for the project through schools, colleges and universities accessible to the research team. All participants were volunteers; because of the nature of the UK language learning population, it was necessary to visit several institutions to locate sufficient numbers of Spanish L2 learners, and it was not possible to create a gender balanced sample (most were female). The project followed the ethical procedures recommended by Talkbank and by the British Association for Applied Linguistics; informed consent was obtained from participants who signed an appropriate data release form. In the case of the beginner subjects aged 13-14, parental consent was obtained via the collaborating schools. Data was collected on site in the various collaborating institutions, by trained members of the research team. The computer based tasks were administered on standalone project laptops, and all speech was audiorecorded using portable digital equipment.

Suitable age matched native speaker participants were identified either in the UK (visiting Erasmus students) or in Spain through schools known to research team members, and tasks were run on school sites. The locations available for data collection in schools and colleges in the UK and in Spain varied in the degree of privacy/ soundproofing, but soundfiles of acceptable quality were obtained in all cases.

Data preparation

The soundfiles were transcribed and checked by members of the research team, following a specially produced, detailed guide on the use of the CHAT transcription system with Spanish L2 data (Arche 2007a). This process is appropriately described by Chaudron (2003: 767) as “highly labour intensive”,

requiring up to 10 hours' transcription and checking time for each hour of audio data. A sample transcription extract is included as Figure 5; summary information on the transcription procedures follows.

INSERT FIGURE 5 ABOUT HERE

In line with CHAT conventions, the SPLLOC transcriptions were created using conventional Spanish orthography to facilitate later analysis with the CLAN program suite. At times therefore the transcription is somewhat deviant from the actual phonological shape of the words produced by learners. The CHAT error tier ("%err") has been used to indicate such cases, e.g.:

```
@Begin
@Languages: es
@Participants: S02 Subject, MJA Investigator
@ID: es|splloc|S02||female|Year9||Subject||
@ID: es|splloc|MJA||female|||Investigator||
@Date: 27-MAR-2007
@Location: K
@Situation: Picture Sequence
@Coder: CSP
@Time Duration: <0:06:56>
*MJA: [^ eng: student number two picture sequence task] .
*MJA: qué hace el estudiante con el bolígrafo ?
*S02: estudiante usar [*] el bolígrafo estudiante guardar el bolígrafo .
%err: iusar = usar
*MJA: qué hace el chico con las sillas ?
*S02: el chico tirar las sillas el chico recoger [*] las sillas .
%err: recoguer = recoger
```

Researchers interested more specifically in e.g. L2 Spanish phonology can of course refer in future to the actual soundfiles and add their own level of coding to the transcripts provided.

The standard CHAT header set was used when starting and concluding all transcriptions, and utterances were segmented at the level of main clauses, with coordinating conjunctions and adverbials (

such as *y, pero, entonces, luego, o, puesto que, ya que, sin embargo, no obstante*…) being used as guides to segmentation. Normal CHAT conventions were also followed regarding the representation of speech, e.g. on the use of punctuation, and markup of pauses, retracings, incomplete utterances, overlaps, direct speech etc. Specialist guidance and codes were produced for specific issues arising in the transcription of SLA data:

- Codeswitching into L1 English at word or phrase level is marked by adding "@s:" followed by a different code corresponding to different part of speech categories (e.g. noun (d), verb (v) etc.):

  *P63: y cómo se dice scuba@s:d diving@s:v ?

- Complete codeswitched utterances are marked between square brackets starting with the code "^eng:".

  *P04: [^ eng: I don't know what that means].

- Direct learner imitations of investigator utterances in Spanish are marked with "@g" at the end of the imitated word.

  *P51: no están en el sol están en shade@s:d.
  *MJA: la sombra.
  *P51: la@g sombra@g.

- Use by learners of indeterminate forms and idiosyncratic neologisms is marked with "@n" at the end of the word.

  *P54: um ehm detrás de lo eh pictura@n eh hay [/] hay un número de turistas .

(For full details of these specialist conventions see Arche 2007a: 16-19.)

Once transcribed and checked, according to CHAT conventions, the soundfiles and transcriptions have been fully anonymised preparatory to inclusion in the project database and public dissemination via the web. A second stage of transcript preparation has then been undertaken, involving the part of speech tagging of the CHAT transcripts using the MOR and POST programs for Spanish available from CHILDES. These programs require some adaptation for use in SLA research with adults (e.g. additional vocabulary needs to be added to the existing word lists within MOR). While the programs tag much of the data automatically, final checks, disambiguations and corrections have to be conducted by hand. Again, a specialist guide has been produced (Arche 2007b), and is being followed by team members in this final stage of data preparation, which at the time of writing is not yet complete. A sample POS tagged transcript extract is included as Figure 6.

INSERT FIGURE 6 ABOUT HERE

Finally a database to house all soundfiles, CHAT transcriptions, POS tagged files, and XML versions of the CHAT transcriptions is being created at the University of Southampton using the ORACLE database program. The database will be accessible through a web interface and website viewable at www.splloc.soton.ac.uk. All bona fide researchers willing to follow the CHILDES protocols regarding ethics and data-sharing will be able to access and download the material for teaching and research purposes.

Conclusion

This paper has briefly reviewed the origins and development of corpus based research in second language acquisition. We have noted the emergence of different types of learner corpus (spoken vs

written) and of different approaches to computer aided data analysis. We have described the overall design of a new corpus of spoken L2 Spanish, which is being created using CHILDES tools, and which will be made freely available in early 2008 for use by the SLA research community. The corpus has been designed so as to facilitate hypothesis-testing research, exploring a number of claims regarding the role of interfaces in second language acquisition; the team will be reporting separately elsewhere on the outcomes of this programme of substantive research using the corpus. Here we invite the research community to access the corpus and assess its usefulness for their own diverse research purposes, both hypothesis-building and hypothesis-testing. We look forward to receiving feedback from other users which will help us to improve and further develop this shared resource, so that it can make a lasting contribution to the ongoing development of Spanish SLA research.

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 and S. Petch-Tyson (eds). 55-76. Amsterdam: John Benjamins.

Altenberg, B. and Granger, S. 2001. "The grammatical and lexical patterning of MAKE in native and non-native student writing". *Applied Linguistics* 22(2): 173-195.

Andersen, R. and Shirai, Y. 1996. "The primacy of aspect in first and second language acquisition: The pidgin-creole connection". In *Handbook of Second Language Acquisition*. W. Ritchie and T. Bhatia (eds). 527-570. London: Academic Press.

Arche, M. J. 2007a. *SPLLOC Transcription Guidelines*. Unpublished manuscript, University of Southampton. On-line access www.splloc.soton.ac.uk.doc/SPLLOCTranscriptionGuidelines.doc .

Arche, M. J. 2007b. *MOR Guidelines for SPLLOC*. Unpublished manuscript, University of Southampton.

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

Barlow, M. 2005. "Computer-based analysis of learner language". In *Analysing Learner Language*. R. Ellis and G. Barkhuizen (eds). 335-354. Oxford : Oxford University Press.

Bartning, I. 1997. "L'apprenant dit avancé et son acquisition d'une langue étrangère. Tour d'horizon et esquisse d'une caractérisation de la variété avancée". *Aile* 9 : 9-50.

Cadierno, T. 1993. *Explicit Instruction in Grammar : A Comparison of Input Based and Output Based Instruction in Second Language Acquisition*. PhD Thesis, University of Illinois.

Chaudron, C. 2003. "Data collection in SLA research". *In Handbook of Second Language Acquisition*. C. J. Doughty and M. H Long (eds). 762-828. Oxford: Blackwell

CHILDES, Child Language Data Exchange System. Pittsburgh: Carnegie Mellon University. On-line access http://childes.psy.cmu.edu, consulted December 2007.

CLAWS Part of Speech Tagger for English. Lancaster: University of Lancaster. On-line access http://ucrel.lancs.ac.uk/claws/, consulted December 2007.

Crawford, B., Becker, T. and Nekrasova, T. 2007. "Lexical bundles in L2 writing: frequency of input". Paper presented at 30th Annual Second Language Research Forum, University of Illinois, October 2007.

Dagneaux, E. Denness, S. and Granger, S. 1998. "Computer-aided error analysis". *System* 26 (2):163-174.

David, A. 2007. "Vocabulary and morphosyntactic complexity in L2 learners". Paper presented at EuroSLA 17, Newcastle upon Tyne, September 2007.

Dippold, D. 2007**.** *Faces, Roles and Identities in Argumentative Discourse : The Development of Facework Strategies by L2 Learners of German*. PhD Thesis, University of Southampton.

Domínguez, L., Arche, M.J. 2007a. "Deviant optional forms in L2 Spanish: the case of word order variation". Poster presentation at GALA, Barcelona, 6-8 September.

Domínguez, L., and Arche, M. J. 2007b. "The L2 Acquisition of SV/VS contrast in Spanish". Paper presentation at the Hispanic Linguistic Symposium, Texas, 1-4 November.

Ellis, N. C. and Ferreira, F. 2007. "Form, function and frequency in the learning of L2 constructions". Paper presented at 30th Annual Second Language Research Forum, University of Illinois, October 2007.

Feldweg, H. 1993. "Transcription, storage and retrieval of data". *Adult Language Acquisition: Crosslinguistic perspectives*. Volume I: Field methods. C. Perdue (ed.). 108-130. Cambridge: Cambridge University Press.

FLLOC, French Learner Languae Oral Corpora. Southampton: University of Southampton. On-line access at www.flloc.soton.ac.uk, consulted December 2007.

Franceschina, F. 2003. *The Nature of Grammatical Representations in Mature L2 Grammars: The Case of Spanish Grammatical Gender*. PhD Thesis, University of Essex.

Gass, S.M. and Mackey, A. 2007. *Data Elicitation for Second and Foreign Language Research.* Mahwah, NJ: Lawrence Erlbaum.

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 and S. Petch-Tyson (eds). 3-36. Amsterdam: John Benjamins

Granger, S., Dagneaux, E. and Meunier, F. (eds). 2002. *International Corpus of Learner English*. Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics, 2002.

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

Hasselgren, A. 2002. "Learner corpora and language testing: smallwords as markers of learner fluency". In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. S. Granger, J. Hung and S. Petch-Tyson (eds). 143-174. Amsterdam: John Benjamins.

Herschensohn, J. 2003. "Verbs and rules: two profiles of French morphology acquisition". *Journal of French Language Studies* 13(1): 23-45.

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 and S. Petch-Tyson (eds). 77-118. Amsterdam: John Benjamins.

Huebner, T. 1983. *The Acquisition of English*. Ann Arbor: Karoma.

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

Izumi, E., Uchimoto, K. and Isahara, H. 2004. "SST speech corpus of Japanese learners' English and automatic detection of learners' errors". *ICAME Journal* 28: 31-48.

Jagtman, M. and Bongaerts, T. 1994. "Report-COMOLA: a computer system for the analysis of interlanguage data". *Second Language Research* 10: 49-83.

Klein, W. and Perdue, C. (eds) 1992. *Utterance Structure: Developing Grammars Again*. Amsterdam: John Benjamins.

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

Lardière, D. 1998. "Case and tense in the fossilized steady state". *Second Language Research* 14(1): 1-26.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition. Volume 1: Transcription Format and Programs.*. Mahwah NJ: Lawrence Erlbaum Associates.

Malvern, D. and Richards, B. 2002. "Investigating accommodation in language proficiency interviews using a new measure of lexical diversity". *Language Testing* 19(1): 85-104.

Marsden, E. and David, A. forthcoming. "Comparison of lexical development in L2 French and L2 Spanish". To appear in *Language Learning Journal*.

Milton, J. and Chowdhury, N. 1994. "Tagging the interlanguage of Chinese learners of English". In *Entering Text*. L. Flowerdew and K.K. Tong (eds). 127-143. Hong Kong: Hong Kong University of Science and Technology.

Mitchell, R. and Dickson, P. 1997. *Progression in Foreign Language Learning. CLE Occasional Paper 45*. University of Southampton: Centre for Language in Education.

Montrul, S. 2004. *The Acquisition of Spanish*: *Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and in 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. 2005. "The emergence of morphosyntactic structure in French L2". In *Focus on French as a Foreign Language: Multidisciplinary approaches.* J.-M. Dewaele (ed). Xx-xx. Clevedon, Avon: Multilingual Matters.

Myles, F. 2007. "Using electronic corpora in SLA research". In *French Applied Linguistics*, D. Ayoun (ed). 377-400.

Myles, F., Hooper, J. and Mitchell, R. 1998. "Rote or rule? Exploring the role of formulaic language in classroom foreign language learning". *LanguageLearning* 48: 323–363.

Myles, F. and Mitchell, R. 2005. "Using information technology to support empirical SLA research". *Journal of Applied Linguistics* 1: 69–95.

Myles, F., Mitchell, R. and Hooper, J. 1999. "Interrogative chunks in French L2: a basis for creative construction?" *Studies in Second Language Acquisition* 21: 49–80.

Nesselhauf, N. 2003. "The use of collocations by advanced learners of English and some implications for teaching". *Applied Linguistics* 24(2): 223-242.

Parisse, C. and Le Normand, M. T. 1997. « Etude des catégories lexicales chez le jeune enfant à partir de deux ans à l'aide d'un traitement automatique de la morphosyntaxe ». *Bulletin d'Audiophonologie* 13(6): 305-328.

Parisse, C. and Le Normand, M. T. 2000. "How children build their morphosyntax: The case of French". *Journal of Child Language* 27(2): 267-292.

Perdue, C. (ed.) 1993a. *Adult Language Acquisition: Cross-Linguistic Perspectives. Volume 1: Field Methods.* Cambridge: Cambridge University Press.

Perdue, C. (ed.) 1993b. *Adult Language Acquisition: Cross-Linguistic Perspectives. Volume 2: The Results.* Cambridge: Cambridge University Press.

Pienemann, M. 1992. "COALA – a computational system for interlanguage analysis". *Second Language Research* 8: 59-92.

Rule, S. 2004. "French interlanguage corpora: recent developments". *Journal of French Language Studies* 14: 343–356.

Rule, S., Marsden, E., Myles, F. and Mitchell, R. 2003. "Constructing a database of French interlanguage oral corpora". In *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL Technical Papers16, D. Archer, P. Rayson, E. Wilson and T. McEnery (eds), 669-677. University of Lancaster.

Rule, S. and Marsden, E. 2006. "The acquisition of functional categories in early French second language grammars: the use of finite and non-finite verbs in negative contexts". *Second Language Research* 22(2): 188-218.

Rutherford, W. and Thomas, M. 2001. "The Child Language Data Exchange System in research on second language acquisition". *Second Language Research* 17 (2): .

Sato, C. 1990. *The Syntax of Conversation in Interlanguage Development*. Tübingen: Gunter Narr.

Scott, M. 1997. *Wordsmith Tools*. Oxford: Oxford University Press. For on-line access see http://www.lexically.net/wordsmith/, consulted December 2007.

Talkbank. Pittsburgh: Carnegie Mellon University On-line access http://talkbank.org/ , consulted December 2007.

Tono, Y. 2001. "The Standard Speaking Test (SST) Corpus: A 1 million word spoken corpus of Japanese learners of English and its implications for L2 lexicography". In *Proceedings of ASIALEX 2001*: 257–262.

Figure 1: Discussion Task, Sample Items

Sample task card (advanced learners):

¿Cómo podríamos mantener y mejorar los derechos de los animales?

_____ prohibiendo todos los experimentos científicos con animales

_____ prohibiendo la venta de abrigos de piel

_____ no permitiendo la caza de animale

_____ cerrando todas las granjas de animales

_____

Puntúa estas actividades otorgando la mayor puntuación a la opción que creas ser la mejor (1), y la menor puntuación a la que creas ser la peor (5) Por favor incluye también una sugerencia propia.

Después comparte tu puntuación con tu compañero. El objetivo es que os pongáis de acuerdo y creéis una lista común. Es importante que te asegures de que tu opinión sea escuchada y siempre explica el por qué de tu elección.

Sample task card (intermediate learners):

¿Cómo podríamos los ciudadanos ayudar con la preservación del medio ambiente?

_____ reciclando todo lo que podamos (vidrio, cartón, plástico, etc.)

_____ instalando paneles solares en las casas

_____ recogiendo el agua de lluvia y reusándola en el hogar

_____ utilizando el autobús siempre que podamos

_____

Please rank the suggested measures from what you think is the most acceptable / helpful (1) to the least acceptable /helpful (5), according to you. Add a further suggestion of your choice.

Then discuss the above question with your partner. Your task is to find the best compromise and agree on a rating. Offer the pros and cons of each argument and make sure your opinion is heard, and always give reasons for your choices!

Citizens	ciudadanos	Environment	medio ambiente
Recycle	reciclar	Glass	vidrio/cristal
Home	hogar	To collect	recoger
The rain	agua de lluvia		

Figure 2: Clitic Interpretation task, Sample Items

Figure 3: Clitic Production Task, Sample Items

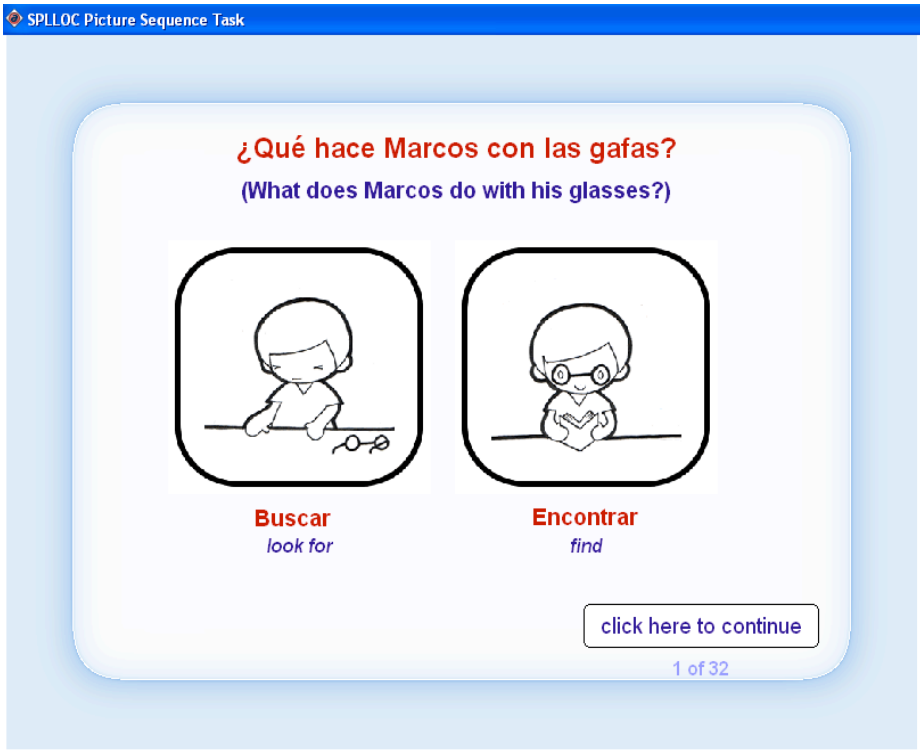

Figure 4: Word Order Task, Sample Item

You are in the cinema watching a film with some friends. One of your friends, you don't know who, sneezes very loudly so you ask Andrés: "¿Quién ha estornudado?" (Who has sneezed?)

What could Andrés say?

a. Ha estornudado Juan b. Juan ha estornudado c. Both sentences

Figure 5: Sample CHAT Transcription

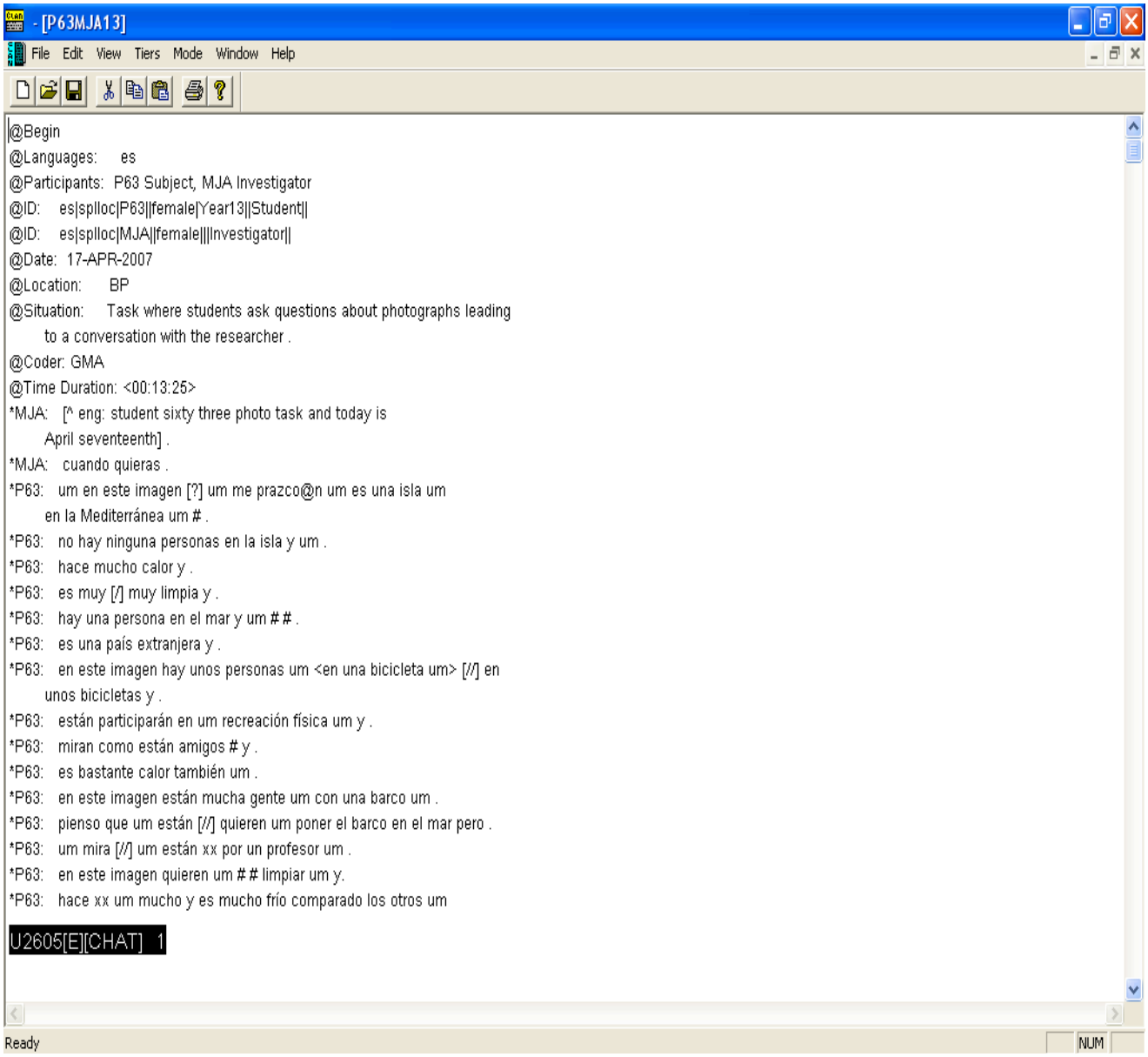

Figure 6: Sample POS Tagged File