Research Report:

Linguistic Development in L2 Spanish: Creation and analysis of a learner corpus (ESRC award RES-000-23-1609, April 2006-March 2008)

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

1. Background

Electronic learner corpora are making an increasing contribution to second language acquisition (SLA) research (Granger 2002, Barlow 2005, Myles 2005a, 2007, Mitchell et al 2008). At the University of Southampton, a database has been developed titled French Learner Language Oral Corpora (FLLOC), comprising c 3,000,000 words of learner French at different levels (see www.flloc.soton.ac.uk for details). The FLLOC research programme has adopted the procedures of the well known CHILDES project in first language acquisition (MacWhinney 2000), including the transcription system known as CHAT and analysis software known as CLAN. Much of the collection has been tagged for parts of speech using a French version of the CLAN MOR program (Parisse and Le Normand 1997, 2000), and current analyses of the datasets address issues relating to interfaces between syntax, morphology and lexis, and the acquisition of functional categories (e.g. Myles 2005b, Rule and Marsden 2006, David 2007).

Spanish is a global language with increasing numbers of first language users and of L2 learners. Correspondingly there is a developing international community of SLA researchers specializing in Spanish, but as yet there are no generally available learner corpora for L2 Spanish. Recruitment at Southampton of a Spanish linguistics specialist (Domínguez) plus involvement as co-investigator of a second Spanish specialist with experience of learner corpora (Marsden, University of York) made it possible to propose creation of a parallel database for learner Spanish (Spanish Learner Language Oral Corpus: SPLLOC), and to conduct an initial programme of substantive research in Spanish SLA using the new dataset. Building on FLLOC experience, it was decided to adopt CHAT and CLAN for the new SPLLOC project. An attractive feature was the availability of a part of speech tagger for Spanish (Spanish MOR), among other linguistic analysis programs.

The principal Southampton investigator (Mitchell) and the Newcastle co-investigator (Myles) had collaborated in the creation and development of the FLLOC programme. Two part time Research Associates (Domínguez and Rule) were initially appointed to the new SPLLOC project, which started in April 2006 with task development and piloting. However both were appointed to linguistics lectureships in September 2006, though Domínguez remained very active in the project in the role of Assistant Director and Rule gave ongoing technical advice on CHAT and CLAN. Project fieldwork began in autumn 2006 with a temporary research assistant; a new full time Research Associate (Arche) was appointed from January 2007, who managed the remaining fieldwork and directed all aspects of data preparation including training transcribers, coding data and preparing files for inclusion in the public database. Within the team, Domínguez, Arche and Marsden have led on different aspects of the substantive research programme; Mitchell and Myles had overall responsibility for project management and strategic direction, and Mitchell oversaw local database development by the project IT Development Officer (Xiang).

2. Objectives

The original objectives were:

1. To establish a small scale high quality database of spoken learner Spanish, and promote its use among the Spanish SLA research community.
2. To undertake a short programme of substantive research into L2 Spanish (development of verb inflections, clitic pronouns, word order);
3. To clarify interfaces between morphology, syntax and pragmatics in the development of L2 Spanish.

The objective of creating a database of spoken learner Spanish comprising digital audio files and accompanying transcriptions has been fully achieved. The design and creation of the database are described in the "Methods" section below. The database has already been made fully available to the research community through the project website www.splloc.soton.ac.uk, described in Section 6.1. Promotional activity comprised a series of conference presentations described in Section 6.3, and a successful one-day seminar organized in Southampton in January 2008 (see Section 5).

The only difficulties encountered in database creation arose because the Spanish MOR part of speech turned out to be less developed and less accurate than anticipated. This meant that additional effort had to be invested in developing the Spanish lexicon required for use with MOR, and an unexpectedly large amount of hand correction had to be undertaken on MOR-tagged output files. Having taken expert advice we concluded that this problem could not be solved within the scope of this project, and consequently we have so far tagged only part of the publicly available corpus (70 of 290 files). Tagging was focused on one particular task (picture description + interview), to support the lexical analysis being undertaken in the substantive research programme. During coming months we will be addressing this software problem and completing the POS tagging of the corpus.

Objectives 2 and 3 relate to our substantive research programme in Spanish SLA. This agenda was altered to some extent, partly as a result of personnel changes (loss of Rule in autumn 2006) and the difficulties experienced with the MOR programme, and partly because of evolving collaboration with the parallel FLLOC research programme. Work on the acquisition of Spanish clitic pronouns and on word order was undertaken as planned, led within the SPLLOC team by Domínguez and Arche. These investigations have allowed us to address the broader 'interface' issues identified in Objective 3. However in place of the proposed work on verb inflections also mentioned under Objective 2, a programme of comparative work has been undertaken on lexical development in both L2 Spanish and L2 French, led for SPLLOC by Marsden and for FLLOC by Dr Annabelle David; this work also has 'interface' implications. The main findings of these investigations are presented in Section 4.

3. Methods

3.1 Design principles

Database design has been guided overall by a number of key principles:

Principle 1: Focus on speech

It was decided the database would prioritise collection of semi-naturalistic L2 speech data, rather than written data, since speech produced under real time communicative pressure can provide more direct evidence about the state of the L2 learner's underlying interlanguage system.

Principle 2: Variety of genres

Learner speech is characterized by variability, e.g. in use of target language morphology, and learners also tend to avoid 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). 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 different speech genres, and with varying interlocutors. This semi-naturalistic part of the database can be defined as a learner corpus in a strict sense (Nesselhauf 2004, 2005).

Principle 3: Balance of open ended and focused tasks
In addition to the more open ended elicitation tasks, the same learners completed selected focused tasks, relevant to the substantive research agenda. Activities prompting learners' production and/or interpretation can address problems of learner avoidance of particular target structures, 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). Three focused tasks were designed, relevant to the team's theoretical interests in the acquisition of Spanish clitic pronouns (Tasks 5 and 6), and in word order issues relevant to the syntax/ pragmatics interface (Task 7).

Principle 4: variety of learner levels
Because of the short project timescale, a cross sectional design was adopted. All learner participants were L1 English speakers, learning L2 Spanish in classroom contexts. 20 learners were located at each of 3 levels, differentiated by age and number of years of instruction. In addition, comparison groups of native speakers of Spanish undertook all tasks, and were age-matched where appropriate. Details of the learner participants 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 = 20	21-22	C 895 hours + year abroad	University (Year 4)	C1

Principle 5: Use of CHILDES procedures
Learners' spoken L2 output would be captured as digital audio files and transcribed using CHAT conventions, adapted as necessary to take account of interlanguage features, to facilitate subsequent analyses using the CLAN software suite.

Principle 6: Accessibility
Anonymised data arising from the oral tasks (soundfiles + transcripts) would be made available to the research community through a specially created website and deposited in the Talkbank database www.talkbank.org .

3.2 Task development

In summer 2006 a number of candidate tasks were developed, piloted and evaluated. Some were adapted from previous SLA studies, while others were specially developed. The tasks eventually selected are briefly described below. Fuller details are provided at http://www.splloc.soton.ac.uk/taskdesc.html

Task 1: "A Monster Mistake"
This storytelling task was based on a picture sequence taken with permission from Hunt (2003), which tells a 'Loch Ness Monster' story. This task had previously been used in the FLLOC programme, and was known to be suitable for learners at beginner and intermediate level.

Task 2: "Modern Times"
Sequences from the Charlie Chaplin film "Modern Times" have been used with adult learners in other SLA research including the ESF project (Perdue 1993) and the FLLOC programme. A sequence from the film was trialled for SPLLOC, because of concern that "Monster Mistake" might fail to show the full narrative abilities of advanced learners. However piloting showed that the "Modern Times" narrative did not work well with younger learners, and consequently it was used only with the advanced participants. These participants watched a 5 minute sequence from the film and then re-told it, with support from a set of still images and vocabulary list.

Task 3: Picture description and interview
All participants undertook a version of this three-part interview individually, with a member of the research team as interlocutor. In Part 1, learners were shown an age-appropriate set of stimulus photos and asked to describe the various scenes and activities. In Part 2 they were asked to find out as much additional information as they could about the characters shown. In Part 3, the researcher asked the participant a range of questions about their current interests, past activities, and future plans.

Task 4: Discussion
This task was adapted with permission from Dippold (2007). It took the form of a pair discussion between two learners (intermediate and advanced only). The task aimed to elicit expressions of opinion and preferences, and also evidence of learners' turntaking, initiation and repair skills. Pairs were offered a choice of discussion cards, setting out a series of claims/ arguments on current topics, e.g. 'global warming'. Having chosen a topic, the pair were asked to discuss and rank the given arguments.

Task 5: Clitic production
This focussed elicitation task was designed to 'push' learners to produce orally a range of Spanish clitic object pronouns. Computer-based, the task consisted of 32 items. For each item, the learners viewed a two-picture sequence, heard and read a stimulus question, and answered it orally. The aim was to elicit a range of clitics varying in number and gender (both canonical and non canonical), and in different collocational contexts. This task is adapted with permission from Cadierno (1993); a sample item follows.

Qué hace la chica con su pelo?

(What does the girl do to her hair?)

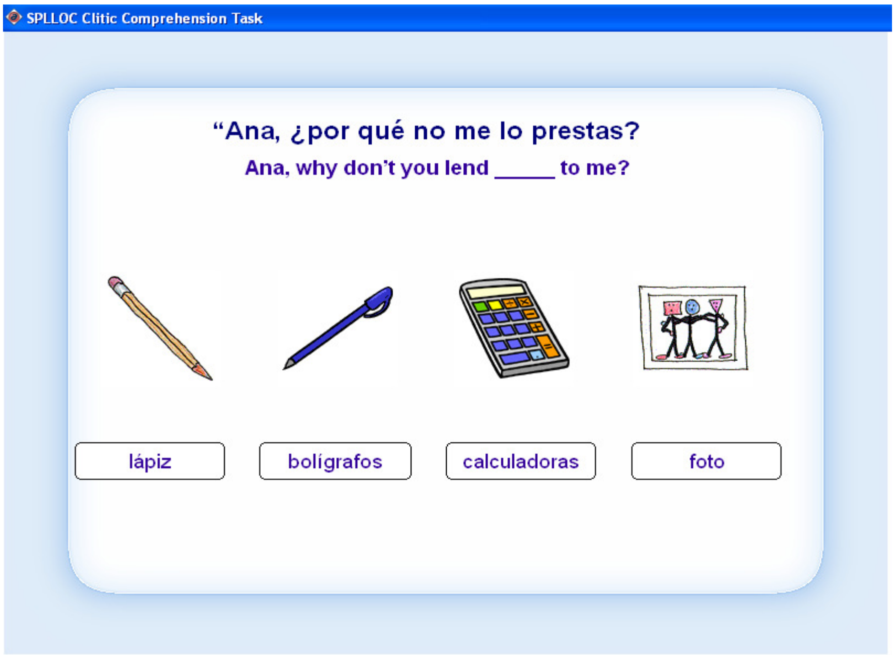
Lavar
Wash

Secar
dry

Typical native speaker response: primero se lo lava y luego se lo seca.

Task 6: Clitic interpretation

This focussed interpretation task also investigated learners' knowledge of clitic object pronouns, and was adapted with permission from Franceschina (2003). Computer-based, it comprised 32 multiple choice items. Each item presented the participant with a stimulus Spanish sentence both orally and in writing. Each sentence included a clitic object pronoun marked for gender and number. Four nouns of varying number/gender were offered as possible responses and the participant was asked to select the one which matched the stimulus. A sample item follows.



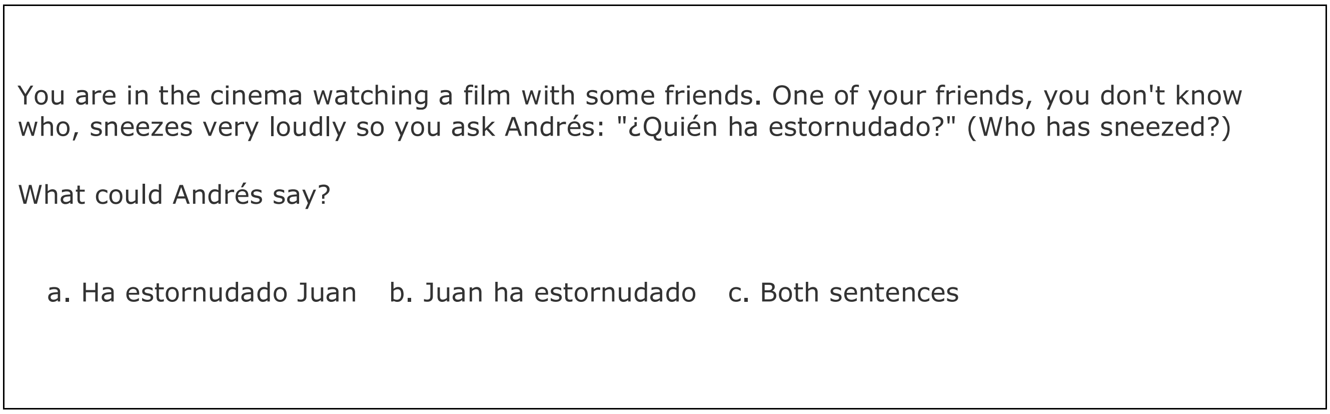
SPLLOC Clitic Comprehension Task

"Ana, ¿por qué no me lo prestas?

Ana, why don't you lend _____ to me?

Task 7: Word order

The final focused task was a 28 item paper based context-dependent preference test based on Hertel (2003) and Domínguez (2007). The object was to document participants' knowledge of word order variation at the syntax/ pragmatics interface. Participants were offered a situation (described in English) plus a question in Spanish and three alternative responses, and selected the response(s) they judged appropriate. A sample item follows.



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

3.3 Data collection

Participants were identified for the project through visits to schools, colleges and universities. All 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 (see http://www.baal.org.uk/about_goodpractice_full.pdf); informed consent was obtained from participants who signed an appropriate data release form. For the participants aged 13-14, parental consent was also obtained. Data was collected by trained members of the research team following uniform protocols for each task. The computer based tasks were administered on standalone 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. The locations available for data collection in schools and colleges varied in the degree of privacy/ soundproofing, but soundfiles of acceptable quality were obtained in all cases.

An overview of the numbers of learners and native speakers completing each task is shown in Table 2.

Table 2: Number of participants per task

Task	Year 9 learners	Year 13 learners	Undergraduate learners	Native speakers
“Monster Mistake” narrative	20	20	20	15
“Modern Times” narrative	-	-	20	5
Picture description + interview	20	20	20	10
Discussion	-	20	20	10
Clitics production	20	20	20	10
Clitics interpretation	20	20	20	10
Word order	20	20	20	20

3.4 Data preparation

The soundfiles generated by tasks 1-5 were transcribed and checked by research team members plus additional trained transcribers, following a specially written guide (Arche 2007a). This process requires up to 10 hours' transcription and checking time for each hour of audio data. A sample transcription extract follows:

```
@Begin
@Languages:  es
@Participants: P63 Subject, MJA Investigator
@ID:   es|splloc|P63||female|Year13||Student||
@ID:   es|splloc|MJA||female|||Investigator||
@Date:         17-APR-2007
@Location:     BP
@Situation:    Task where students ask questions about photographs leading
         to a conversation with the researcher .
@Coder:        GMA
@Time Duration:      <00:13:25>
*MJA: [^ eng: student sixty three photo task and today is April seventeenth] .
*MJA: cuando quieras .
*P63:   um en este imagen [?] um me prazco@n um es una isla um
        en la Mediterránea um # .
*P63:   no hay ninguna personas en la isla y um .
*P63:   hace mucho calor y .
*P63:   es muy [/] muy limpia y .
*P63:   hay una persona en el mar y um # # .
*P63:   es una país extranjera y .
*P63:   en este imagen hay unos personas um <en una bicicleta um> [//] en
        unos bicicletas y .
*P63:   están participarán en um recreación@n física um y .
*P63:   miran como están amigos # y .
*P63:   es bastante calor también um .
```

In line with CHAT conventions, the SPLLOC transcriptions used 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 learner utterances. The CHAT error tier ("%err") has been used to indicate such cases; future transcript users interested more specifically in e.g. L2 Spanish phonology can access the soundfiles and add their own levels of coding.

The standard CHAT header set was used when starting and ending 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…*) 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 was developed for specific issues arising in the transcription of SLA data, including codeswitching into L1 English; imitations of investigator utterances; and use by learners of indeterminate forms and neologisms (see Arche 2007a for details).

Once checked, the soundfiles and transcripts from Tasks 1-5 were fully anonymised preparatory to public dissemination. A second stage of transcript preparation was then undertaken, involving part of speech (POS) tagging of selected CHAT transcripts using the Spanish MOR and POST programs. The original intention had been to POS tag all transcripts. However it quickly became apparent that these existing programs require some adaptation for use in SLA research with adults (e.g. vocabulary needs to be added to the existing word lists within MOR). More importantly, while the programs tag much of the data automatically, quite extensive final checks, disambiguations and corrections have to be conducted by hand. Again, a specialist guide was produced (Arche 2007b), and POS tagged versions produced for all Task 3 transcripts (N = 70). A sample POS tagged transcript excerpt follows.

```
@Begin
@Languages:  es
@Participants:P63 Subject, MJA Investigator
@ID:   es|splloc|P63||female|Year13||Student||
@ID:   es|splloc|MJA||female|||Investigator||
@Date:          17-APR-2007
@Location:     BP
@Situation:    Task where students ask questions about photographs leading
        to a conversation with the researcher .
@Coder:         GMA
@Time Duration:        <00:13:25>
*MJA: [^ eng: student sixty three photo task and today is April seventeenth] .
*MJA: cuando quieras .
*P63:   um en este imagen [?] um me prazco@n um es una isla um
        en la Mediterránea um # .
%mor: co|um prep|en=in det:dem|este=this n|imagen&FEM=image co|um pro:per|me=me
        neo|prazco co|um vpres|se-3S&PRES=be det:art|un-FEM=one n|isla&FEM=island
        co|um prep|en=in det:art|la&FEM&SG=the n:prop|Mediterránea
        co|um .
*P63:   no hay ninguna personas en la isla y um .
%mor: adv|no=no vpres|habe-3S&PRES&SPEC=have det:indef|ninguno-FEM=none
        n|persona-PL&FEM=person prep|en=in det:art|la&FEM&SG=the n|isla&FEM=island
        conj|y=and co|um .
*P63:   hace mucho calor y .
%mor: vpres|hace-3S&PRES=do det:indef|mucho-MASC=much n|calor&MASC=heat conj|y=and
        .
*P63:   es muy [/] muy limpia y .
%mor: vpres|se-3S&PRES=be adv|muy=very adj|limpio-FEM=clean conj|y=and .
```

Finally a database to house all soundfiles, CHAT transcriptions, POS tagged files, and XML versions of the CHAT transcripts has been created and is described in Section 6.

4. Results of substantive research programme

To date substantive research has been carried out in three main areas: the acquisition of clitic object pronouns; the acquisition of Spanish word order; and the development of L2 Spanish vocabulary. A summary of the findings of each investigation follows.

4.1 **Acquisition of clitic object pronouns** (for full account see Arche and Domínguez forthcoming).

The investigation into clitic object pronouns relates to our theoretical interests in the morphology-syntax interface in SLA. Clitics can be characterized as bundles of morphological features (person, number, gender) agreeing with the DP they refer to (Sportiche 1996). In minimalist theory (Chomsky 1995, 2000, 2001), this agreement process is established within the syntactic derivation.

It has been observed that L2 learners produce very few clitic pronouns, and fail to interpret their morphological features in a consistent way. Two different hypotheses can account for this: either L2 learners do not have the appropriate syntactic representation (Impaired Representation Hypothesis - Tsimpli and Roussou 1991, Meisel 1997, Hawkins and Chan 1997, Franceschina 2005) or they fail to map unimpaired syntactic representations to their appropriate morphological forms (Missing Surface Inflexion Hypothesis - Haznedar & Schwartz 1997, Lardière 1998, Prévost & White 2000, Bruhn de Garavito 2003).

In this investigation these two hypotheses were compared using data from Tasks 5 and 6 (clitics production and interpretation). The two hypotheses make different predictions about task modality. The IRH predicts no different outcomes from different kinds of test, since, if learner's representation is impaired, this should be attested regardless of the task. However, the MSIH predicts better results from comprehension than production tests, where additional factors can affect the construction of an appropriate representation.

The results show that target-like responses in both tasks correlate with the level of proficiency of the learners. Beginners perform poorly in both the use of clitics (4%) and the comprehension task (40%). Advanced learners show very high rates of accurate responses in both production (75%) and comprehension (90%), which shows that acquisition of clitics is possible, though they contain features absent in learners' L1, such as gender. Intermediate learners show a low use of clitics (20%), but high comprehension rates (70%), which suggests that the syntactic representation is not impaired. This gives empirical evidence in favor of the MSIH and not the IRH, according to which general inaccuracy is expected. Interestingly, although the intermediate group's production is low, 90% of the clitics produced are targetlike. This supports the idea that the appropriate morphosyntactic features are represented in the learners' interlanguage and strongly suggests that lack of production may be due to factors such as processing limitations or communication pressures.

4.2 Acquisition of Spanish word order (for full account see Domínguez and Arche 2008).

The appearance of optional constructions is quite a pervasive phenomenon in L2 near-native grammars (Papp 2000, Prévost and White 2000, Sorace 2000) and their source still remains unclear. In Spanish, word order variation is ruled by both syntactic and pragmatic constraints: whether subjects appear preverbally (SV) or postverbally (VS) depends on both the syntax of the verb (unergative (1a) vs. unaccusative (1b)) and the type of information encoded in the sentence (broad (2) vs. narrow focus (3)). Thus, these structures present an ideal scenario for investigating optionality in non-native grammars which has been previously explained as a result of deficits in the syntax-pragmatics interface (Hertel 2003, De Miguel 1993, Lozano 2006).

The data collected using SPLLOC Task 7 (word order) were analysed to test alternative explanation for the optionality of SV/VS structures in learner Spanish. The results show that the acceptability of VS orders is in strict correlation with learner proficiency levels. Subject-verb inversion (an option not allowed in their L1) is not selected by learners in the lower two groups, but is correctly preferred by the advanced group. More importantly, there is a sharp contrast in the advanced group between their consistent preference for the inverted VS order in clitic left dislocations (4) on the one hand, and sentences with intransitive verbs on the other, where both inverted and non-inverted forms were allowed by the native controls. The syntax of the verb (i.e. unergative or unaccusative) barely affects the answers of the advanced group. This is relevant since the acceptability of both SV and VS clause types in sentences with unaccusative verbs weakens previous hypotheses that the syntactic constraints ruling inversion are properly acquired from early on and, consequently, mismatches between native and non-native forms have to be analysed as the result of a pragmatic deficit. If this was the case, inversion involving unaccusatives (only affected by syntactic constraints) would be allowed more consistently than inversions with unergatives (affected also by pragmatic constraints), but this was not attested in the data. Moreover, if a pragmatic deficit was the source of problems in the acquisition of focus-driven constructions, the acceptability of VS in clitic left dislocations would be unexpected as they are subject to pragmatic constraints also.

These results suggest that the availability of optional forms in L2 developing grammars are the result of an overgeneralisation of one of the options in the target language to contexts where neither syntactic nor pragmatic rules would allow them. Consequently, the optionality shown by advanced learners should be understood as an intermediate stage showing grammar restructuring, rather than a case of pragmatic deficit.

Examples

(1) a. Juan ha estornudado (unergative) SV
'Juan has sneezed'

b. Ha llegado Juan (unaccusative) VS
has arrived Juan
'Juan has arrived'

(2) *What happened?* (broad focus)

a. [$_{F}$ Juan ha estornudado] SV
'Juan has sneezed'

b. [$_{F}$ Ha llegado Juan] VS
has arrived Juan
'Juan has arrived'

(3) *Who has sneezed?* (narrow focus)

a. Ha estornudado [$_{F}$ **Juan**] V**S**
has sneezed Juan
'Juan has sneezed'

Who has arrived?

b. Ha llegado [$_{\mathbf{F}}$ **Juan**] V**S**
has arrived Juan
'Juan has arrived'

(4) *Who brought the dog?* (narrow focus)

a. El perro, lo trajo [$_F$ Juan] O#, Cl-V- **S**
The dog, it brought Juan
'The dog, John brought it'

b. *El perro, Juan lo trajo O#, **S**-Cl-V

4.3 Development of L2 vocabulary (for full account see Marsden and David 2008).
This comparative investigation has provided several key insights into lexical progression amongst school learners of French and Spanish. The analysis was based on the MOR tagged files for SPLLOC Task 7 (picture description + interview) for the Year 9 and Year 13 learners, plus parallel data for learner French. Using CLAN software, a range of analyses were carried out including counts of lemma types and tokens, and frequency counts for different parts of speech. Lexical and inflectional diversity were analysed using a range of measures (TTR, D and LRD: McKee et al 2000, Malvern et al 2004).

When the measures of diversity were calculated using D, the results suggest that the year 13 learners used significantly more lexically and inflectionally diverse language than the Year 9 learners.

Few statistically significant differences were found between the French and Spanish learners in terms of the amount of language produced in the task. The data did suggest some differences however. For example, the analysis suggested that Spanish learners had more diverse verbs and nouns than the French learners, regardless of level of proficiency, and that they had a higher ratio of verb types to noun types (LRD) than the French year 13. (Overall, however, all the learners' nouns were more diverse than their verbs.) It seemed that the Spanish learners had slightly greater overall diversity of lemmas, in both years, though the statistical significance of this was borderline, and was not supported when *D* was calculated on words. Another difference between learners of French and Spanish was that Spanish learners produced more tokens and types of nouns than the French learners in both years. They also produced proportionally more nouns (and adjectives). Any differences that were found between the languages were constant across the different proficiency levels. This could suggest that the differences were due to the nature of the language (e.g. Spanish may, intrinsically, have more diverse lemmas, particularly nouns, than French) and/or to the teaching (e.g. Spanish instruction may facilitate development of the lexicon to a greater extent than French instruction), rather than quicker or easier learning in Spanish than in French.

The findings suggested that year 9 learners produced a greater proportion of nouns (out of their total production) than year 13, regardless of language, and that year 13 produced a greater proportion of verbs than year 9. The proportion of verbs was also the same in French as it was in Spanish amongst learners in the same year of study (despite the previously noted differences for nouns and adjectives). This finding could support the idea that the increase in proportion of verbs, specifically, is an indicator of progression, in line with the claim of Broeder, Extra and van Hout (1993) that more proficient learners use proportionally more verbs. We also found that, in both languages, verb types (out of all types) increased between year 9 and year 13, yet noun types (out of all types) decreased.With further corroboration, this could have important implications for assessment descriptors (e.g. should we be looking out for more 'verby' productions as an indication of higher proficiency?) and for teaching (should we be emphasising the learning of verbs?).

5. Activities

The SPLLOC project has benefited throughout from ongoing collaboration with the FLLOC research team, especially with regard to technical issues concerning transcription of learner language, database development, and the use of CLAN software. This has led to comparative research being undertaken by members of the two teams (Marsden and David).

In January 2008 the SPLLOC project team organized a one day seminar to promote the new corpus among the Spanish SLA research community. There were 22 participants, from universities in the UK, Spain, Portugal and the United States. SPLLOC team members presented short papers on the development of the corpus and preliminary research results, with an international expert in SLA corpora Dr Amaya Mendikoetxea from the Universidad Autonoma de Madrid as discussant. A hands-on workshop gave participants practical experience in accessing and analysing SPLLOC corpus data. Details of the programme and summaries of the papers presented are available at http://www.splloc.soton.ac.uk/event.html .

6. Outputs

6.1 SPLLOC database and website

The database of spoken learner Spanish created in the course of the project comprises 290 digital audiofiles, in two formats (.wav and .mp3). These are accompanied by full transcriptions in CHAT format. All files from the picture description + interview task have also been tagged for parts of speech (n = 70), and 115 files have to date been converted to XML format (Loch Ness task and clitics production task). Tagged and XML versions of the remaining files will be added to the database in coming months.

The database is already available to the research community through the project website www.splloc.soton.ac.uk. The website includes descriptions of all data elicitation tasks plus the project's transcription conventions and listings of project outputs. Data files can be located through a search facility and are freely downloadable. The SPLLOC project adheres generally to the researcher and user ground rules which have been developed by the international CHILDES project, available at http://talkbank.org/share . CLAN software is needed to read and analyse the files and is downloadable from the CHILDES website. The dataset has been offered to the ESRC Economic and Social Data Service, and will also be offered in due course to Talkbank.

6.2 SPLLOC publications to date

Arche, M.J. 2008a. *SPLLOC Transcription Conventions*. Available at http://www.splloc.soton.ac.uk/trancon.html

Arche, M.J. 2008b. *MOR Guidelines for SPLLOC*. Internal document.

Domínguez, L., Arche, M.J. 2008. "Optionality in L2 Grammars: the Acquisition of SV/VS Contrast in Spanish." In *BUCLD 32 Proceedings* , H. Chan, H. Jacob and E. Kapia (eds.), 96-107. Somerville , MA : Cascadilla Press.

Mitchell, R., Domínguez, L., Arche, M. J., Myles, F. and Marsden, E. 2008. "SPLLOC: A new database for Spanish second language acquisition research." To appear in *EuroSLA Yearbook 8*.

Marsden, E. and David, A.2008. “Vocabulary use during conversation: a cross-sectional study of development from year 9 to year 13 amongst learners of Spanish and French.” To appear in *Language Learning Journal*.

Arche, M. J. and Domínguez, L. forthcoming. “Morphology and syntax interaction in SLA : a study of clitic acquisition in Spanish.” To appear in Galani, A., Hicks, G. and Tsoulas, G (eds) *Morphology and its Interfaces* . John Benjamins.

6.3 External conference presentations

Refereed presentations have been given/ accepted at the following events:

- BAAL LLT SIG Conference “Towards a Researched Pedagogy”, Lancaster, July 2007.
- GALA, Barcelona, September 2007.
- EuroSLA 17. Newcastle, September 2007.
- Second Language Research Forum, Urbana-Champaign, October 2007.
- Hispanic Linguistics Symposium, Texas, November 2007 (2 papers).
- Boston University Conference on Language Development (BUCLD), Boston, November 2007.
- SLA workshop, Free University, Brussels, March 2007.
- BAAL/CUP seminar “Conceptualising learning in Applied Linguistics”. Newcastle, June 2008.
- Barcelona Linguistics Institute, August 2008.
- AILA World Congress, Essen, August 2008.
- EuroSLA 18, Aix-en-Provence, September 2008.

For full details see http://www.splloc.soton.ac.uk/publication.html

7. Impacts

The main impact of this research is anticipated to lie within the field of Spanish SLA, where a rich new resource has been made available for both research and teaching. In addition, this research has potential to influence curriculum design and pedagogy for the teaching of Spanish as a second/ foreign language. For example, Marsden and David (forthcoming) draw preliminary conclusions from their work on lexical development e.g. regarding the need for pedagogy to prioritise development of knowledge of verbs over nouns. They have prioritized publication of their work in a professional journal and the team is committed to ongoing dissemination through professional as well as research networks.

8. Future research priorities

A follow-up project titled “Emergence and development of the tense-aspect system in L2 Spanish” has been funded by ESRC award no RES-062-23-1075 to commence in August 2008. During this project research will be conducted on the development of the verb system using both the existing dataset and additional data to be added to the collection. Solutions to the problems of POS tagging SLA data will be further explored through further work on Spanish MOR and/or the evaluation and adoption of alternative programs. Research will also continue on the development of L2 vocabulary and related theoretical issues including the interface between lexical and morphosyntactic development.