Contribution of learner corpora to SLA

SPLLOC: A NEW CORPUS OF LEARNER SPANISH

UNIVERSITY OF SOUTHAMPTON

25/1/2008

Amaya Mendikoetxea,
Universidad Autónoma de Madrid

Aims of the discussion:

- To evaluate the impact of learner corpus research in recent SLA theory.
- To evaluate the contribution of the SPLLOC project to this area of inquiry.
- To point out future challenges for corpus-based SLA research.

Gathering learner data (1)

Much current SLA research favours experimental, metalinguistic and introspective data, and tends to be dismissive of natural language use data.

[Granger 2002]

- Aim of SLA research ► to build models of:
  - The underlying mental representations of learners at a particular stage in the process of L2 learning.
  - The developmental processes which shape and constrain L2 production.

"The language produced by learners, whether spontaneously or through various elicitation procedures remains a central source of evidence for these mental processes, and the success of SLA research therefore relies on having access to good quality data." [Myles 2005: 374]

Gathering learner data (2)

- Reasons why elicitation techniques are favoured in SLA research [from Mackey & Gass 2005]:

1. The particular structure you want to investigate may not occur in natural production: it may be absent or there may not be enough instances.
2. To answer your research question you may need to know what learners rule out as a possible L2 sentence.
   - Presence of a particular structure/feature in the learners´ natural output does not necessarily indicate that the learners "know" the structure.
   - Absence of a particular structure/feature in natural language use data does not necessarily indicate that learners do not know the structure.
   - In addition, if learners do not use a form at all, we cannot assume that they do not use that form unless they consistently do not use it in a required context.

Gathering learner data (3)

3. It is difficult to control the variables that affect learner output in a non-experimental setting (Granger 2002: 6).

→ Consequence

"As it is difficult to subject a large number of informants to experimentation, SLA research tends to be based on a relatively narrow empirical base, focusing on the language of a very limited number of subjects, which consequently raises questions about the generalizibiity of the results." (Granger 2002: 6)

Gathering learner data (4)

► Why do we need corpora in SLA research?

- To test current hypotheses on larger and better constructed datasets in order to see if SLA findings can be generalized.
- To find sets of data not normally found in small studies: structures which are crucial to inform current debates. [large datasets]
- To discover patterns.
- For quantitative studies (e.g. frequency).

Computer learner corpora (CLC) (1)

- Rough definition of CLC:
  Electronic collections of learner data.
  - This definition is too broad and fuzzy: it leads to the term being used for data types which are in effect not corpora
- Granger´s (2002: 7) definition of CLC (based on Sinclair´s 1996 definition of corpora):

> Computer learner corpora are electronic collections of authentic FL/SL textual data according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and are documented as to their origin of provenance.

Computer learner corpora (CLC) (2)

- CLC types:

Commercial:
Longman Learners´Corpus (10 million words).
Cambridge Learner Corpus (16 million words).

Academic:
The Hong-Kong University of Science and Technology Learner Corpus (25 million words)
International Corpus of Learner English (ICLE) (2.5 million words).
French Language Learner Oral Corpus (FLLOC) (2 million words)
Spanish Language Learner Oral Corpus (SPLLOC)
[WriCLE & CEDEL2]

Monolingual ⇔ Bilingual (parallel)
General ⇔ Technical (ESP)
Synchronic ⇔ diachronic (Longitudinal)
Written ⇔ Spoken

CLC in SLA research (1)
Types of studies with CLC data

1) Hypothesis-driven/corpus-based

Using CLC data to test specific hypotheses/ research questions about the nature of IL generated through introspection, SLA theories, or as a result of the analysis of experimental or other non-corpus based sources of data.

2) Hypothesis-finding/corpus-driven

Investigating CLC data in a more exploratory way and initiating analyses that yield patterns of data, which can then be inspected for unusual features. Such features may then be used to generate hypotheses about learner language.

[from Barlow 2005]

CLC in SLA research (2)
Types of studies with CLC data

- **Alternberg, B. (2002)** 'Using bilingual corpus evidence in learner corpus research' in Granger *et al.* (eds).
  - It studies one type of causative construction *(make somebody happy)*, which Swedish learners overuse, in a parallel corpus.
  - It wants to explore questions like: how central are these structures in the two corpora?; to what extent are these constructions retained in translation?; what are the main causative types and how are they used?
  - Its aim is to see how contrastive data can use to explain Swedish overuse of the construction. Transfer is explained in terms of prototypicality.
- **Aijmer, K. (2002)** 'Modality in advance Swedish learner interlanguage' in Granger *et al.* (eds.)
  - Aijmer uses CLC data to compare the range and frequency of some key modal words in native English writing and L2 English writing of advanced level university students: L1 Swedish and some L1 French and L1 German from ICLE.
  - The author is interested in the overuse and underuse of some forms and suggests possible pedagogical implications.

CLC in SLA research (3)
Types of studies with CLC data

Housen, A. (2002) 'A corpus-based study of the L2-acquisition of the English verb system' in S. Granger *et al* (eds.)

It aims to investigate the Aspect Hypothesis put forward in the L2 literature: the emergence, early use and subsequent development of verb morphology in L2 is strongly influenced by the inherent semantic properties of the lexical verb which the learner selects to refer to a particular event.

Tono, Y. (2004) 'Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English', in G. Aston *et al* (eds.)

It investigates the acquisition of verb subcategorization frame patterns by Japanese learners of English by examining the relative influence of factors such as : The effect of L1, the amount of exposure to L2 input, and the properties of inherent verb semantics on the use and misuse of verb subcategorization patterns.

CLC in SLA research (4)
Evaluation

- The majority of CLC studies are hypothesis-finding studies.
- There are biases in practice (Barlow 2005):
  - The experimental/generative tradition favours hypothesis-driven/corpus-based studies.
  - Corpus linguists have a preference for a hypothesis-finding/corpus-driven methodology.
- On the whole, the contribution of CLC research so far has been much more substantial in description than interpretation of SLA data (Granger 2004, Myles 2005). Two reasons (Granger 2004: 134-135):
  - Learner corpus research has been mainly conducted by corpus linguists, rather than SLA specialists (Hasselgard 1999).
  - The type of IL CLC researchers have been most interested in (intermediate to advanced) was so poorly described in the literature that they felt the need to establish the facts before launching into theoretical generalizations

CLC in SLA research (5)
Evaluation

- Researchers use almost exclusively written L2 corpora. Very little use has been made of oral corpora.
- Overwhelming focus on advanced learners (but no formal measure of proficiency provided).
- Most L2 corpora are untagged and when they are tagged they use very specific schemes, which makes it difficult to share data.
- Most of the studies using corpora make little use of software other than concordances.
- Analysis tools are fairly limited: lexical searches, frequency counts, concordances and manual annotation.
- The developmental dimension is almost lacking.
- Most work is rather descriptive: documenting differences between native and non-native English, rather than explaining.
- Corpus-based (or corpus-driven) L2 studies are also not sufficiently informed about SLA theory: Little or no reference to current debates and hypotheses in the SLA literature.
- Strong pedagogical bias: researchers tend to assume that finding out differences in use between learners and L1 speakers will have direct pedagogical implications, which is not always the case.

CLC in SLA research (6)
Evaluation

Such research is useful nonetheless, as we need to have good descriptions of learner language in order to inform our understanding of what shapes its development, but it is now time that corpus linguists and SLA specialists work more closely together in order to advance both their agendas (Myles 2005: 381).

SPLLOC:
A spoken corpus of L2 Spanish (1)

"[..] 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." [Mitchell et al. 2008]

Key principles:

Principle 1: Focus on speech:

Semi-naturalistic L2 speech data vs. written data: "spontaneous speech produced in face to face interaction is likely to provide more direct evidence about the state of the L2 learners' underlying interlanguage system." [Mitchell et al. 2008]

SPLLOC:
A spoken corpus of L2 Spanish (2)

- There are likely to be fewer performance errors in the written language and the errors found are those that escape monitoring, indicating grammatical or lexical gaps in the learners' mental grammar.
- Learners tend to use more complex structures when they are writing, which could be more revealing in terms of their linguistic competence than the simplified language often found in oral language.
- Written corpora are often used to study of native grammars and are considered to be a good reflection of language competence.
- Written corpora are particularly suitable to study the features of the interlanguage of advanced learners, especially in comparison with similar corpora of native speakers:
  - Learner corpus research in the ICLE tradition shows that advanced learner texts are a valuable source of data to study aspects such as modality, degree adverbs, tenses, collocations, phraseology, the expression of causativity, information structure, clefts, anaphora, etc.
  - Written corpora can also be used in hypothesis-testing studies: passivised structures and expletives (Oshita 2000, 2004), the study of subject inversion in L2 English (Lozano & Mendikoetxea, in press).

SPLLOC:
A spoken corpus of L2 Spanish (3)

Principle 2: Variety of genres

Learners undertake a range of semi-naturalistic oral activities in different genres: narrative, interview and picture description, and peer discussion.

► Purpose: to minimise typical problems of oral production data: avoidance in speech production of structures or features of the target language where learners feel insecure or dysfluent.

- What is authentic data in L2 corpora? (Granger 2002)
  - Learner data is rarely fully natural, especially in the case of EFL learners, who learn English in a classroom.
  - Scale of naturalness (Nesselhauf 2004:128)
    fully natural – product of teaching process – controlled task – scripted

► In general, the more intervention by the researcher, the further away we are from 'authentic' data.

SPLLOC:
A spoken corpus of L2 Spanish (4)

- Authentic' learner data in a classroom environment = "data resulting from authentic classroom activity" (Granger 2002: 8).
- In a foreign language environment, what comes closest to naturally occurring texts are:
  - Texts that are produced for pedagogical reasons.
  - Texts that are produced for the corpus but that use procedures exerting very little control.

(Nesselhauf 2004: 128)

- ✓ Free compositions produced for a certain course.
- ✓ Free compositions produced for a corpus.
- ✓ A text read aloud in class
- ✓ Oral interview for a corpus.

SPLLOC:
A spoken corpus of L2 Spanish (5)

- What is NOT authentic data in L2 corpora?
  - x Composition guided by pictures.
  - x Students´ translations
  - x Typical experimental data resulting from elicitation techniques.

- Can these be part of learner corpora?
- Nesselhauf (2004: 128): "Since the distinction between more or less controlled is , naturally, not clear-cut, such collections might be considered peripheral parts of learner corpora"

- Sinclair (1996): Data collected though major intervention by the linguist, or the creation of special scenarios, form "experimental" corpora. Speech corpora are often experimental.

SPLLOC:
A spoken corpus of L2 Spanish (6)

Principle 3: Balance of open ended and focused tasks

- To address the problem of learner avoidance
- To get an insight into what learners know is correct and what they know is not possible in the L2 grammar.
- To allow for triangulation across different data types.

- Given the limitations of exclusively corpus-based approaches, corpus linguists are arguing for the combination of corpus and experimental data (Gilquin 2007, de Mönnink 1997, 2000) BUT:
  - Combining corpus data with experimental data is not simply a question of integrating two statistical outputs (see de Mönnink 1997): There are problems involved in combining two approaches to gathering data which are very different in essence: e.g. how to interpret the occurrence of constructions in a corpus which are not expected on the basis of experimental data?
  - The use of both types of data should not be a linear, uni-directional process (see de Mönnink 2000´s multi-method approach).

SPLLOC:
A spoken corpus of L2 Spanish (7)

Principle 4: Variety of learner levels

- The corpus includes learners at three different proficiency levels to maximize its usefulness to study development in L2 Spanish.

- No formal independent proficiency measure is provided: the levels are differentiated by age and years of instruction.

"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" [Mitchell *et al.* 2008]

SPLLOC:
A spoken corpus of L2 Spanish (8)

Principle 5: Use of CHILDES procedures

- It is associated with a robust set of transcription conventions (CHAT) and a range of analysis software: programs to calculate frequency, concordancers.
- You can use the CLAN suit; the POS tagger can be adapted to take into account IL features.
- It facilitates data sharing.

An annotated learner corpus should ideally be based on standardised annotation software in order to ensure comparability of annotated learner corpora with annotated native corpora. (Granger 2002: 10)

- Most learner data are ´raw´ because of the difficulty of annotating learner language.
- Researchers tend to develop their own annotation schemes & software.

- Any limitations of CHILDES? user-friendliness, suitability for other types of data (written, complex structure), etc.?

SPLLOC:
A spoken corpus of L2 Spanish (8)

Principle 6: Accessibility

The complete dataset will be made available to the research community through the SPLLOC webpage.

SPLLOC-based research

- Focused tasks (Domínguez & Arche):
  - SV/VS order in L2 Spanish
  - Acquisition of clitics in L2 Spanish

- Open-ended task (Marsden & David)
  - Vocabulary use during conversation

SPLLOC-based research (1)

- Focused tasks (Domínguez & Arche):
  - ► Aim:

    To test the hypothesis that grammatical structures with features the syntax-discourse interface are more prone to instability (more vulnerable) than features in narrow syntax (Sorace 2000, 2004, 1005, Tsimpli et al. 2004).

    Learners:

    3 groups of L2 learners (beginner, intermediate, advanced) and a control group of native speakers.

    Task:

    SV/VS order in L2 Spanish: acceptability test

    Clitics: production and comprehension task

SPLLOC-based research (2)

► Results:

- SV/VS order:
  - Only advanced learners show native-like behaviour.
  - Beginners and intermediate learners show divergent grammars, preferring SV structures independently of discourse (syntactic deficit).
- Clitics:
  - -Accuracy in performance correlates with the level of proficiency.
  - -Intermediate level learners score very high in the comprehension task (higher than 80%) but show very low usage of clitics (around 20%).

► Interpretation:

- SV/VS order:

  Ambiguity and lack of robustness in the input forces indeterminacy, even at advanced level of proficiency, and that is independent of learners´ knowledge of pragmatic rules.
- Clitics

  The acquisition of the morphosyntactic properties of the clitics takes place earlier than the pragmatic ones.

SPLLOC-based research (3)

► Main challenge:

How to combine these results with corpus-based studies using SPLLOC?

- For VS/SV order a corpus analysis should look at both VS and SV structures – the latter are particularly interesting for unaccusative verbs.
- According to what is found in the corpus, the acceptability test may have to be modified (e.g. to include gradiency).
- A corpus study of clitics may throw more light on the production/ comprehension discrepancy found for intermediate learners.
- It may be interesting to look at written corpora in search for more evidence regarding the nature of deficits.

SPLLOC-based research (4)

Open-ended task: Marsden & David

- It shows some of the features of the types of studies corpora are most suited for, but contributes some new features:
  - Comparative analysis of learners acquiring different L2

    [most corpus-based studies compare subgroups of learners of different L1 backgrounds and same L2]
  - Comparative analysis of learners at different proficiency levels.

    [most corpus-based studies focus on advanced learners]

  - It incorporates more sophisticated measures of lexical variety, diversity and richness.

CLC in SLA Research: The way forward

- Existing corpora need to be made available to the research community.
- Corpora of L2´s other than English have to be created
- There is also a need for spoken corpora and for longitudinal corpora to address the developmental dimension of L2 learning, as well as for cross-sectional corpora, with learners at different levels of proficiency.
- Such corpora must be compiled according to explicit design criteria which make them useful to conduct SLA research: they must be compiled by SLA researchers (or in collaboration with them).
  - Most available corpora are 'opportunistic'.
  - No formal measurement of proficiency is provided.
- Corpora must be fully documented: e.g. to select texts for subcorpora.

CLC in SLA Research: The way forward

- Analysis tools must be developed which are suitable for learner data and are not reliant on manual tagging.
- It would be useful to use standard annotation to make it possible for data to be shared: Rutherford & Thomas (2001) advocate the use of CHILDES, but is CHILDES really suitable?
- Methodologies have to be developed to combine corpus data with experimental data in search for converging evidence.
- There is a need for a clearer relationship between (learner) corpus linguists and SLA, with more hypothesis-testing, more explanatory studies.
- This line of research has to be made more visible.

Thank you!!!

References

- Aston, Guy, Bernardini, Silvia, and Stewart, Dominic. 2004. *Corpora and Language Learners.* Amsterdam: John Benjamins
- Barlow, M., 2005. Computer-based analysis of learner language. In R. Ellis and G. Barkhuizen *Analysing Learner Language.* Oxford: OUP.
- De Mönnink, I. 2000. *On the move.* Language and Computers 31. Rodopi.
- De Mönnink, I. 1997. Using Corpus and Experimental Data: a Multi-method Approach'. In M. Ljung (ed.). *Papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17).*
- Gilquin, G., 2007. To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. Zeitschrift für Anglistik und Amerikanistik 55, 273-291.
- Granger, S, 2002. A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Language Learning and Language Teaching 6. Amsterdam & Philadelphia: Benjamins, 3-33.
- Granger, S. 2004. 'Computer learner corpus research: current status and future prospects.' in G. Aston et al. (eds.)
- Granger, S. J. Hung and S. Petch-Tyson (eds.), 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Language Learning and Language Teaching 6. Amsterdam & Philadelphia: Benjamins, 3-33.
- Hasselgard, H., 1999. review of S. Granger (ed.) *Learner English on Computer. Icame journal,* 23, 148-152.
- Mackey, A. & S. M. Gass. 2005. *Second Language Research. Methodology and Design.* London: Lawrence Erlbaum.
- Lozano, C. & A. Mendikoetxea, in press. Postverbal subjects at the interfaces in English and Italian learners of English: a corpus study. In B. Díaz, G. Guilquin and S. Papp (eds.) *Linking up Contrastive and Learner Corpus Research.* Amsterdam: Rodopi.
- Mitchell, R. , L. Domínguez, M. Arche, F. Myles and E. Marsden, 2008. SPLLOC: A new corpus for Spanish second language acquisition research. Paper submitted to *EUROSLA Yearbook 8.*
- Myles, F. 2005. 'Review article. Language Corpora and Second Language Acquisition Research' *Second Language Research* 21, 4, 373-391.
- Nesselhauf, N., 2004. *Collocations in a Learner Corpus.* Amsterdam: John Benjamins.
- Sinclair, J., 1996. *EAGLES. Preliminary Recommendations on Corpus Typology.* Online manuscript.