CEDEL2
Corpus Escrito del Español L2 / L2 Spanish written corpus

Cristóbal Lozano
Universidad de Granada
www.ugr.es/~cristoballozano

International Seminar on Corpus Linguistics,
New Trends in Language Teaching and Translation Studies

Universidad de Granada, 24th Sept 2008

BACKGROUND: LEARNER CORPORA (1)

- Use of corpora in L1 acquisition is widespread since 80’s: CHILDES (McWhinney 2000).

- BUT not standard practice in L2 acquisition due to lack of learner corpora.

- Early this decade: L2 English
  - ICLE ‘International Corpus of Learner English’ (Granger et al. 2002a)

- Other L2 English corpora not commercially available.
  - LLC: Longman Learner Corpus
  - CLC: Cambridge Learner Corpus
  - SULEC: Santiago University Learner Corpus

BACKGROUND: LEARNER CORPORA (2)

- L2 Spanish: no such corpora available yet.

- BUT increased interest in L2 Spanish research, esp. in USA
  - Mostly experimental data.

- In this context:
  - CEDEL2 (this presentation)
  - SPLLOC (Spanish Learner Language Oral Corpus)
    - University of Southampton (Mitchell et al. 2008).
    - Oral corpus, L1 Eng - L2 Spa.
    - CHAT format, will be available in CHILDES.
    - Semi-natural oral tasks and controlled tasks (elicitation of citics and word order) problematic according to Sinclair.
    - No placement test: learners are classified according to age and course problematic for learner corpora.
CEDEL2: research context

- **WOSLAC** research group
  - ‘Word Order in Second Language Acquisition Corpora’
  - Amaya Mendikoetxea, Univ. Autónoma Madrid
  - [http://www.uam.es/woslac](http://www.uam.es/woslac)
  - See also: Chocano et al. (2007), Lozano (in press).

- **Aims:**
  - Acquisition of **formal aspects** in L2.
  - Role of **interfaces** in L2 acquisition of **syntax**
    - Syntax-discourse interface
    - Syntax-lexicon interface
    - Syntax-phonology interface
  - Learners show problems when syntax **interfaces** with other domains (see Sorace 2005, 2006).
  - So, 2 corpora are being created to test whether deficits at the interfaces with word order stem from **L1 transfer** or rather universal **developmental patterns**.
    - WRICLE (L1 Spa – L2 Eng)
    - CEDEL2 (L1 Eng – L2 Spa)

CEDEL2

- ‘Corpus escrito del español L2’ (L2 Spanish written corpus)
- **Written** corpus
- **L1 Eng – L2 Spa**
- Data are being collected **online**
- **All proficiency levels**
  - Beginner, intermediate, advanced
  - Standartized placement test (Univ. of Wisconsin 1998).

Problems in learner corpus design

- Some learner corpora use **ad hoc methodology**.
  - Corpus is designed to elicit and collect specific linguistic structures

- **CEDEL2** design follows some methodological principles used in monitor & normative corpora
  (Sinclair 2005 ‘Corpus and text: Basic principles’)
  - Sinclair proposes 10 basic **design principles**:

1. **PRINCIPLE 1: Content selection**

   “The contents of a corpus should be selected without regard for the language they contain, but according to their **communicative function**”

   [Sinclair 2005:1]
1. Sinclair distinguishes 2 content criteria:
   - **Internal criteria**: language of texts.
   - **External criteria**: communicative function of texts.

   CEDEL2 design follows **external criteria** only.

2. **PRINCIPLE 2: Representativeness**

   Corpus builders should strive to make their corpus as **representative** as possible of the language from which it is chosen.

   Sinclair 2005:2

3. **PRINCIPLE 3: Contrast**

   Only those components of corpora which have been designed to be **independently contrastive** should be contrasted.

   Sinclair 2005:3
3

- CEDEL2 is not only a learner corpus but also a contrastive corpus.
- CEDEL2 allows these contrasts:
  - Different levels of learners’ interlanguage:
    - E.g., intermediate vs. Advanced
  - Interlanguage vs. Native language
    - Comparable subcorpus of Spanish native speakers.
    - Subcorpus follows same design principles as learner corpus.
    - See next principle (Structural criteria)

4

PRINCIPLE 4: Structural criteria

"Criteria for determining the structure of a corpus should be small in number, clearly separate from each other ...
Sinclair 2005:5"

5

PRINCIPLE 5: Annotation

"Any information about a text ... should be stored separately from the plain text and merged when required in applications"
Sinclair 2005:5
5

- CEDEL2: **UAM CorpusTool** software
  - Developed by Mick O’Donnell at UAM.
  - Freely downloadable
- Works on **raw text** file.
- Stores tagged information on separate **XML** file.

6

PRINCIPLE 6: Sample size

"Samples of language for a corpus should wherever possible consist of **entire documents** or transcriptions of **complete speech events** .... This means that samples will differ substantially in size"

Sinclair 2005:7

6

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

6

- CEDEL2: Some **variability** in sample size.
  - **Beginners**: relatively shorter texts (from one long paragraph to several hundred words).
  - **Advanced**: usually, longer texts (up to 400 words max.)
  - Word limit for each composition: **500 words**.

- **Bottom line**:
  - All texts are **complete speech events**.
7 PRINCIPLE 7: Documentation

The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

Sinclair 2005:7

7 PRINCIPLE 8: Representativeness & balance

The corpus builder should retain, as target notions, representativeness and balance.

Sinclair 2005:9

8 For Sinclair, balance refers to proportion spoken vs written language.

CEDEL2:
- Only written language.
- Written language is as reliable as spoken language to study interlanguage grammars.
- Idea of learner written lang being more ‘controlled’ and less natural than spoken lang dates back to Krashen (1982) Monitor Hypothesis perpetuated over the decades.
- BUT: production of postverbal subjects with unaccusative verbs: ‘It arrived the day of his departure’ (Lozano & Mendikoetxea 2007, 2008)
  - Attested in experimental data in L2 Eng
  - Attested in corpus data in L2 Eng, ICLE corpus
- So, no principled reason to believe that written lang is less reliable than spoken lang for research purposes.
PRINCIPLE 9: Topic

Any control of the **subject matter** [i.e., **topic**] in a corpus should be imposed by the use of external, and not internal, criteria.

Sinclair 2005:10

- **Topic** is ruled out by external criteria.
- **Vocabulary** is also ruled out by external criteria!
  - « ...it seems strange to many people that it is essential that the vocabulary should not be directly controlled. But vocabulary choice is clearly an internal criterion » (p. 9)
- Therefore **topic** and **vocab** should not be controlled, as they represent internal criteria.

**CEDEL2**: as stated earlier, CEDEL2 is designed to elicit representative learner language: no control over vocab or structures to be produced.

PRINCIPLE 10: Homogeneity

A corpus should aim for **homogeneity** in its components while maintaining adequate coverage, and **rogue texts** should be avoided.

Sinclair 2005:14

- **CEDEL2**:
  - Rogue and atypical texts are discarded as they are collected.
  - After data collection: double checking of atypical texts.
CEDEL2: DATA COLLECTION

- **Online data collection:**
  - [http://www.uam.es/wosac/start.htm](http://www.uam.es/wosac/start.htm)

- **Learners** fill in 3 forms:
  - Learning background
  - Standardised placement test
  - Composition

- **Natives** fill in 2 forms:
  - Learning background
  - Composition

---

Placement test

- University of Wisconsin Spanish Placement Test (1998)

---

Learning background

- Please estimate your ability in Spanish:
  - Very advanced
  - Advanced
  - Intermediate
  - Lower Intermediate
  - Elementary
  - Beginner

- Do you speak any languages in addition to English and Spanish?
  - Please choose.

Composition

- Instructions for composition:
  - Write a short essay about a topic you've chosen.
  - Word limit: 750 words.
  - Do not use internet search or dictionary as verifying.
TAGGING

- **UAM CorpusTool** software
  - Developed by Mick O’Donnell at Univ. Autónoma de Madrid.
  - Freely available: [http://www.waqsoft.com/CorpusTool](http://www.waqsoft.com/CorpusTool)

- **Tagging process**
  1. Define tags via a tag scheme.
  2. Select segment (text, paragraph, sentence, word)
  3. Assign tags to segment.

Tagging example: coding scheme

- See Lozano (forthcoming 2008)

Tagging example: Actual tagging

- **UAM Corpus Tools** – also:
  - Corpus search
  - Stats (descriptive & inferential)
  - XML format.

DATA: evolution

- **CEDEL2: Grand total number of words**
  - **Aim:** 1 million words
**DATA: words by group**

CEDEL2: TOTAL NUMBER OF WORDS (by group)

- **Natives**: 131488, 27%
- **Learners**: 357732, 73%

**DATA: Number of participants**

CEDEL2: NUMBER OF PARTICIPANTS

- **Natives**: 448, 27%
- **Learners**: 1227, 73%

**DATA: Location**

Learners: Location of institution

- **USA**: 920, 74%
- **UK**: 80, 7%
- **SPAIN**: 43, 4%
- **NEW ZEALAND**: 19, 2%
- **AUSTRALIA**: 8, 1%
- **CANADA**: 9, 1%
- **OTHER/UNKNOWN**: 148, 11%

**Calls for participation**

<table>
<thead>
<tr>
<th>2006</th>
<th>2007 &amp; 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2006 Portal del Hispanismo (Instituto Cervantes)</td>
<td>Mar 2007 AEDIPAN</td>
</tr>
<tr>
<td>May 2006 ToddELE.net</td>
<td>Mar 2007 AESLA</td>
</tr>
<tr>
<td>May 2006 INFOLING</td>
<td>Mar 2007 Corpora List</td>
</tr>
<tr>
<td>May 2006 AEDIPAN</td>
<td>Mar 2007 Linguist List</td>
</tr>
<tr>
<td>May 2006 Wordpress.com</td>
<td>Apr 2007 Infoling</td>
</tr>
<tr>
<td>May 2006 Centro Virtual Cervantes</td>
<td>May 2007 Democratic Underground.com</td>
</tr>
<tr>
<td>May 2006 FORMESPAN</td>
<td>May 2007 ELE inicios.es</td>
</tr>
<tr>
<td>June 2006 DEI (Instituto Cervantes)</td>
<td>Oct 2007 Linguist List</td>
</tr>
<tr>
<td>June 2006 Corpora List</td>
<td>Oct 2007 FORMESPAN</td>
</tr>
<tr>
<td>June 2006 Linguist List</td>
<td>Oct 2007 AESLA</td>
</tr>
<tr>
<td>June 2006 Corpus4you</td>
<td>Oct 2007 AEDIPAN</td>
</tr>
<tr>
<td>June 2006 AlteTECH</td>
<td>Oct 2007 Corpora List</td>
</tr>
<tr>
<td>Oct 2006 AESLA</td>
<td>March 2008 Linguist List</td>
</tr>
<tr>
<td>Oct 2006 Linguist List</td>
<td>May 2008 Linguist List</td>
</tr>
<tr>
<td>Oct 2006 Infoling</td>
<td>Sept 2008 Linguist List</td>
</tr>
<tr>
<td>Oct 2006 DeExtranais blogspot</td>
<td></td>
</tr>
<tr>
<td>Oct 2006 FORMESPAN</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- **Next steps:**
  - Create online **taster version**
  - Complete **data collection**
    - Our target: 1 million words
    - So far: half a million words
  - **Tagging** (syntax-discourse)
  - CEDEL2 will be **freely available** on the WOSLAC webpage.

- **Our aim** – to be used by:
  - Researchers
  - Practitioners of ELE (Español Lengua Extranjera)
  - Students of ELE

- « **It is important to avoid perfectionism** in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like. » (Sinclair 2005:81)

References

- Spanish Learner Language Oral Corpus (SPLLOC) [http://www.splloc.soton.ac.uk](http://www.splloc.soton.ac.uk)
- Word Order in Second Language Acquisition Corpora (WOSLAC), Universidad Autónoma de Madrid [http://www.uam.es/woslac](http://www.uam.es/woslac)

Corpus Escrito Del Español L2 (CEDEL2) [http://www.uam.es/woslac/cedel2.htm](http://www.uam.es/woslac/cedel2.htm)
- French Learner Language Oral Corpus (FLLOC), [http://www.flloc.soton.ac.uk](http://www.flloc.soton.ac.uk)

**THANK YOU !**