Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis

Leonardo Campillos Llanos

Computational Linguistics Laboratory
Autonomous University of Madrid (UAM)

Learner Corpus Research Conference (LCR2013) – Bergen, 28th September 2013
Index

- Background of Spanish learner corpus research
- Goals
- Methodology
  - Corpus design
  - Error typology
  - Error analysis
- The corpus search interface
- Results and discussion
- Conclusions
Background of Spanish learner corpus research

- Most learner corpora comprise written data: e.g. **International Corpus of Learner English**.
- Few research projects on spoken learner corpus:

<table>
<thead>
<tr>
<th>ENGLISH</th>
<th>FRENCH</th>
<th>SPANISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINDSEI (Louvain International Database of Spoken English; Gilquin et al., 2010)</td>
<td>FLLOC (French Learner Language Oral Corpus; Myles, 2005)</td>
<td>The Díaz Corpus (Díaz Rodríguez, 2007)</td>
</tr>
<tr>
<td>NICT JLE (NICT Japanese Learner Corpus; Izumi et al., 2004)</td>
<td></td>
<td>SPLLOC (Spanish Learner Language Oral Corpora; Mitchell et al., 2008)</td>
</tr>
</tbody>
</table>
Goals

- Fulfil the lack of **oral corpus** and **computerised resources** for Learner Corpus Research

- To understand the acquisition of the oral expression by different groups of learners of Spanish at **A2** and **B1 levels** (CEFR).
Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis

Methodology

1. Interview and recording
2. Transcription
3. POS-tagging
4. Text-sound synchronisation
5. Interlanguage analysis
6. Computer processing
7. Error tagging
8. Error analysis

Spanish Learner Oral Corpus
Corpus design

- Cross-sectional corpus.
- Participants: Foreign students of Spanish (20-26 years old).
- Low-intermediate level: A2 (N=20) and B1 (N=20) (*CEFR*)
- N=40, clustered in:
  - 9 groups of 4 students with the same L1:
    - Italian, English, Japanese
    - French, German, Chinese
    - Portuguese, Dutch, Polish
  - 1 mixed group of 4 students with other L1s:
    - Korean, Finnish
    - Turkish, Hungarian
- Control group of native speakers (N=4): 2 men and 2 women.
## Corpus design

<table>
<thead>
<tr>
<th>Romance languages</th>
<th>File</th>
<th>L1</th>
<th>Length (mm:ss)</th>
<th>Length L1 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>PORMA2</td>
<td>Portuguese</td>
<td>25:10</td>
<td></td>
<td>1:26:52</td>
</tr>
<tr>
<td>PORWA2_1</td>
<td>Portuguese</td>
<td>20:09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PORWA2_2</td>
<td>Portuguese (Brazilian)</td>
<td>19:51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PORWB1</td>
<td>Portuguese (Brazilian)</td>
<td>21:42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITAMA2</td>
<td>Italian</td>
<td>20:45</td>
<td></td>
<td>1:13:25</td>
</tr>
<tr>
<td>ITAWA2</td>
<td>Italian</td>
<td>13:09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITAMB1</td>
<td>Italian</td>
<td>23:16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITAWB1</td>
<td>Italian</td>
<td>16:15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FREMA2</td>
<td>French</td>
<td>24:08</td>
<td></td>
<td>1:23:17</td>
</tr>
<tr>
<td>FREWA2</td>
<td>French</td>
<td>20:31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FREMB1</td>
<td>French</td>
<td>21:56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FREWB1</td>
<td>French</td>
<td>16:46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sino-Tibetan languages</th>
<th>File</th>
<th>L1</th>
<th>Length (mm:ss)</th>
<th>Length L1 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIWA2_1</td>
<td>Chinese</td>
<td>18:48</td>
<td></td>
<td>1:17:27</td>
</tr>
<tr>
<td>CHIWA2_2</td>
<td>Chinese</td>
<td>18:45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHIMB1</td>
<td>Chinese</td>
<td>18:56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHIW1</td>
<td>Chinese</td>
<td>20:58</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Languages from Japan</th>
<th>File</th>
<th>L1</th>
<th>Length (mm:ss)</th>
<th>Length L1 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAPWA2</td>
<td>Japanese</td>
<td>28:52</td>
<td></td>
<td>1:32:41</td>
</tr>
<tr>
<td>JAPWB1_1</td>
<td>Japanese</td>
<td>16:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAPWB1_2</td>
<td>Japanese</td>
<td>20:59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAPWB1_3</td>
<td>Japanese</td>
<td>26:22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Other languages

<table>
<thead>
<tr>
<th>File</th>
<th>L1</th>
<th>Length (mm:ss)</th>
<th>Length L1 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENGWA2</td>
<td>English</td>
<td>15:04</td>
<td></td>
</tr>
<tr>
<td>ENGMB1</td>
<td>English</td>
<td>18:44</td>
<td></td>
</tr>
<tr>
<td>ENGWB1_1</td>
<td>English</td>
<td>18:02</td>
<td></td>
</tr>
<tr>
<td>ENGWB1_2</td>
<td>English</td>
<td>28:49</td>
<td></td>
</tr>
<tr>
<td>DUTMA2</td>
<td>Dutch</td>
<td>18:19</td>
<td></td>
</tr>
<tr>
<td>DUTWA2_1</td>
<td>Dutch</td>
<td>17:33</td>
<td></td>
</tr>
<tr>
<td>DUTWA2_2</td>
<td>Dutch</td>
<td>23:05</td>
<td></td>
</tr>
<tr>
<td>DUTWB1</td>
<td>Dutch</td>
<td>17:49</td>
<td></td>
</tr>
<tr>
<td>GERMA2</td>
<td>German</td>
<td>18:23</td>
<td></td>
</tr>
<tr>
<td>GERWA2</td>
<td>German</td>
<td>19:45</td>
<td></td>
</tr>
<tr>
<td>GERWB1_1</td>
<td>German</td>
<td>15:35</td>
<td></td>
</tr>
<tr>
<td>GERWB1_2</td>
<td>German</td>
<td>19:41</td>
<td></td>
</tr>
<tr>
<td>POLMA2_1</td>
<td>Polish</td>
<td>22:20</td>
<td></td>
</tr>
<tr>
<td>POLMA2_2</td>
<td>Polish</td>
<td>30:28</td>
<td></td>
</tr>
<tr>
<td>POLMB1</td>
<td>Polish</td>
<td>26:46</td>
<td></td>
</tr>
<tr>
<td>POLWB1</td>
<td>Polish</td>
<td>12:51</td>
<td></td>
</tr>
<tr>
<td>FINWA2</td>
<td>Finnish</td>
<td>20:27</td>
<td></td>
</tr>
<tr>
<td>HUNWA2</td>
<td>Hungarian</td>
<td>21:28</td>
<td></td>
</tr>
<tr>
<td>KORWB1</td>
<td>Korean</td>
<td>21:14</td>
<td></td>
</tr>
<tr>
<td>TURWB1</td>
<td>Turkish</td>
<td>15:56</td>
<td></td>
</tr>
</tbody>
</table>

### Control group

<table>
<thead>
<tr>
<th>File</th>
<th>Sex</th>
<th>L1</th>
<th>Level</th>
<th>Length (mm:ss)</th>
<th>Length L1 group</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPAM_1</td>
<td>M</td>
<td>Spanish</td>
<td>-</td>
<td>18:57</td>
<td></td>
</tr>
<tr>
<td>SPAM_2</td>
<td>M</td>
<td>Spanish</td>
<td>-</td>
<td>26:47</td>
<td></td>
</tr>
<tr>
<td>SPAW_1</td>
<td>W</td>
<td>Spanish</td>
<td>-</td>
<td>16:49</td>
<td>1:22:29</td>
</tr>
<tr>
<td>SPAW_2</td>
<td>W</td>
<td>Spanish</td>
<td>-</td>
<td>19:56</td>
<td></td>
</tr>
</tbody>
</table>

### Total:
13 hs 36’
Data collection method

- One-to-one **semi-controlled** spoken interviews.
- **15-20** minutes long each recording.
- **Tasks:** (similar to foreign language examinations)
  - Description of two photographs about food.
Data collection method

Tasks (cont.):

- **Story retelling task** from pictures.
- A question about two **speech acts**.
- **Spontaneous dialogue**: opinion about topics related to food.
The corpus search interface

- [http://cartago.lllf.uam.es/corele/index.html](http://cartago.lllf.uam.es/corele/index.html)
Error typology

- Classification according to several criteria:
  - **Linguistic level**: e.g. Grammar: *la casa *blanco → blanca
    ('the white house')
  - **Target modification**: e.g. Unnecessary: *un mi amigo ('a my friend')
  - **Category**: e.g. verb: *tiengo → tengo ('I have')
  - **Type**: e.g. ser/estar: *soy satisfecho → estoy ('I am satisfied')
  - **Etiology**: e.g. interlinguistic: e.g. to realise ('darse cuenta')
    ≠ realizar ('to make')
Error analysis

- Error analysis of:
  - Grammar.
  - Lexis.
  - Pronunciation.
  - Pragmatics-Discourse.

- Word counts for each morphological category were obtained to normalise error frequencies.
Results

- 6,838 errors in 52,688 lexical units $\rightarrow$ 12.98 errors per 100 lexical units
- A mean of 170.63 errors per interview (SD = 90.36).
- Progress from A2 to B1 shows a diminution of errors:
Results

- These data only **partially** reflect the acquisition process:
  - They can be related to the **avoidance** of difficult structures.
  - Learners at intermediate levels would be **expected** to make **more errors** than students at lower levels.

→ Students are trying or practising new structures.
Results

- Most errors affect:
  - **Grammar** (48.61%)
  - **Lexis** (29.37%)

- Fewer errors in:
  - **Pronunciation** (14.19%)
  - **Pragmatics-Discourse** (3.58%)
Results

- Around 4.21% are **ambiguous** errors.
- 49.21% of errors would be due to **interference**.
Results

- **Lexical errors** at A2-B1 levels:
  - **Formal errors** are more frequent (80.73% of lexical errors)
    - e.g. borrowings, misformations, malapropisms, gender, calques…
  - **Semantic errors** are less frequent (19.12% of lexical errors)
    - e.g. semantic relation errors, false friends, collocations, register…

Note that in the following I will show only figures for non-ambiguous errors.
Results

- The rate of formal errors decreases at B1
- The rate of semantic errors persists and slightly increases at B1

→ Semantics is more difficult to acquire.
Results

- At A2, the most frequent **lexical errors** are:
  - **Borrowings**: $M = 21.87$ (SD = 46.99)
    - e.g. *tocaba *guitare* (‘I played guitar’) → guitarra
  
  → There is a **large standard deviation** due to the fact that borrowings are very frequent among Portuguese, German, and Dutch learners.
  
  - **Misformations**:
    - e.g. *melijones* → mejillones (‘mussels’)

Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis
Lexical errors decline at B1, but some persist or hardly decrease:

- Semantic relation:
  
  e.g. confusion *ir* (‘to go’) ~ *venir* (‘to come’)

![Graph showing the percentage of non-ambiguous and ambiguous lexical units across CEFR levels]

Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis
Results

- **Lexical errors decline** at B1, but some **persist** or **hardly decrease**:
  - **Gender:**
    - e.g. *el bolso* (‘handbag’) ~ *la bolsa* (‘bag’)

![Graph showing the decrease of lexical errors from A2 to B1 for non-ambiguous and ambiguous words.](image)
Results

- **Grammar**: the most frequent and generalised **errors** affect:
  - **Articles**:
    - e.g. *y Ø camarero está contento* → *el camarero*
      - (‘and [the] waiter is happy’)
  - **Agreement**:
    - e.g. *la comida *famoso* → *famosa*
      - (‘the famous food’)

Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis
Results

- **Grammar**: the most frequent and generalised errors (cont.):
  - **Prepositions**:
    
    e.g. *estoy aquí a Madrid* → **en Madrid**
    
    (‘I am here in Madrid’)
  
  - **Pronouns**:
    
    e.g. *a mí encanta la pizza* → **me encanta** (‘I love pizza’)}
Results

- **Grammar**: the most frequent and generalised errors (cont.):
  - **Sentence structure**:  
    e.g. Blends: *estudio algo como se llama Estudios de cultura* → *estudio algo que se llama…* or *estudio algo como …*  
    (‘I study something called Culture studies’ or ‘I study something like…’)
  - **Past tense**:  
    e.g. *hace 30 años las mujeres no *trabajaron → *trabajaban*  
    (‘women did not use to work 30 years ago’)
Results

- Certain **grammar errors persist** or **hardly decrease** at B1:
  - **Pronouns**: e.g. Él no sabe qué *se quiere → Él no sabe qué quiere
    ('He does not know what he wants')
Results

- Certain grammar errors persist or hardly decrease at B1 (cont.):
  - **Prepositions**: e.g. *He venido en Madrid* → *He venido a Madrid*
    ('I have come to Madrid')
Results

- Certain grammar errors persist or hardly decrease at B1 (cont.):
  - **Subordination**: e.g. \( *que \) entiendo qué pasa
    \[ \rightarrow Espero entender qué pasa \]
    (‘I hope to understand what happens’)

![Graph showing ambiguous and non-ambiguous errors across levels](image)
Results

- The characteristics of spoken discourse may explain the high number of the following grammar errors:
  - **Sentence structure** errors, especially:
    - Omission: e.g. *su restaurante *Ø* muy bien → *está* muy bien
      ('his restaurant is very nice')
    - Word order: e.g. *no realmente* sé → *realmente no* sé
      ('I really do not know')
  - **Agreement**: e.g. *unos amigas* → *unas amigas* ('some friends')
  - **Overuse of present tense**.
Results

- **Pronunciation errors: interference phenomena** tend to strongly persist at B1
  
  → The **L1** maybe has the **greatest** influence.

- However, learners from **every language background** commit certain errors: e.g. the articulation of /r/: *perro* /'pero/ ~ *pero* /'peɾo/ (**‘dog’** ~ **‘but’**)

- **Pragmatics-Discourse errors** show a **wide individual variability**
  
  → each **learner’s rhetoric skills** in the L1 may explain these results
Discussion

- Limitations of the study:
  - Only oral data have been used → it is difficult to diagnose:
    - the type or the linguistic level of certain deviations
    - whether they are due to competence or performance
  - Low number of participants per L1 group, and only at A2-B1 levels:
    - results cannot be generalised
    - conclusions as to the possibility of acquiring an almost bilingual proficiency cannot be inferred
Discussion

- Some **results** are **similar** to error analyses of **written learner corpora of Spanish** (Fernández 1997) and **English** (Díez Bedmar 2011b):
  - The **most frequent** errors affected **grammar**, especially:
    - articles
    - verbs
    - pronouns
  - The **second most frequent** types of errors were **lexical errors**.
  - **BUT statistical significance** does not imply **pedagogical significance** (Díez Bedmar 2011a)
Thank you for your attention!

Contact

Leonardo Campillos Llanos:
leonardo.campillos@uam.es          leonardo.campillos@gmail.com

Corpus interface: http://cartago.illf.uam.es/corele/index.html