Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus

Workshop on Spanish Learner Corpus Research
Universidade da Coruña, 14th July 2015

Cristóbal Lozano
Universidad de Granada
http://wdb.ugr.es/~cristobalballozano
Aim

- To use corpus methods like **fine-grained annotation** and **learner corpus data**
- ...to reveal what has gone unnoticed in **experimental** studies on L2 Spa acquisition...
- ...about the distribution of overt/null pronominal subjects [=**anaphora resolution**] ...
- ...and to propose a **new pragmatic principle** that can account for the observed data.
In pro-drop languages like Spanish (and Italian, Greek, etc) both overt & null pronominal subjects are syntactically licensed:

(1)  

(Él / Ø) bebe cerveza  
He/ *Ø drinks beer  
‘He drinks beer’

...But this apparently “free” syntactic alternation is constrained discursively at the syntax-discourse interface by information structure: topic-continuity and topic-shift.
Pronominal distribution at
SYNTAX-DISCOURSE:

TOPIC-CONTINUITY

- Topic = known/old information to hearer from preceding context.
- Topic continuity is marked via null pronoun in native Spanish.

[Previous context: talking about the main character of a film]

En la película el protagonista tiene una familia y Ø trabaja en un programa de televisión. Un día Ø empieza a recibir videos anónimos...

[RSZ, Spa native, CEDEL2 corpus]

Theoretical literature on pronominal distribution in discourse:

Pronominal distribution at SYNTAX-DISCOURSE (2):

**TOPIC-SHIFT**

- Change of referent in discourse requires topic-shift $\Rightarrow$ use of overt pronoun in native Spa.
  
  [Previous context: talking about the main characters of the film “The illusionist”, a man and a woman]

  La última película que he visto es la de “El Illusionista” ... *Los protagonistas* son dos jóvenes que se conocen y se enamoran. *Él* es de clase baja, mientras que *ella* es de familia noble ...

- BUT also possible full NP in topic-shift contexts.
  
  - This fact has gone unnoticed in the L2 literature
  - Key question: when overt and when NP?? $\Rightarrow$ WE WILL SEE LATER.
Learners acquire the *syntactic alternation* of overt/null pronouns from *early stages* of acquisition → native-like knowledge

- Él/Ø bebe cerveza.

Licers 1989, Phinney 1987, Lozano 2002a, etc etc.

**BUT discursive constraints** are *persistently problematic until very advanced stages.*

- L2ers **overproduce** overt subjects in topic-continuity:
  - Pedro … #él bebe cerveza … #él bebe vino.
- L2ers rarely underproduce
  - Pedro … María … #Ø bebe cerveza … #Ø bebe vino.


What is the **cause** of learners’ persistent discursive problems?

- “**INTERFACE HYPOTHESIS**” (Sorace 2011) L2ers have **deficits/problems** when the syntax interfaces with other modules like discourse.
  - BUT we will see this explanation is not enough.
Previous L2 Spa findings (2)

- Previous L2 studies: mostly (quasi)experimental
  - Hence true pronominal distribution has been traditionally studied in artificial contexts (=comprehension) and NOT in naturalistic contexts with corpora (=production).

- Large learner corpus data rarely used to investigate anaphora resolution.
L2 CORPUS STUDY
Linguistic analysis based on the antecedent

**ANTECEDENT(S)**

- Pedro [+masc]
- Antonio [+masc]
- Pedro [+masc]
- María [-masc]
- Pedro [+masc]

**ANAPHOR**

1. **NP**: Pedro
   - #overt: él
   - #null: Ø
2. **NP**: Pedro
   - overt: él
   - #null: Ø
3. **NP**: Pedro
   - overt: él
   - #null: Ø
4. **NP**: Pedro
   - overt: él
   - #null: Ø
5. **NP**: Pedro
   - overt: él
   - #null: Ø
6. **NP**: Pedro
   - overt: él
   - #null: Ø
7. **NP**: Pedro
   - overt: él
   - null: Ø
8. **NP**: Pedro
   - overt: él
   - null: Ø
9. **NP**: Pedro
   - overt: él
   - null: Ø

**TOPIC-SHIFT**
- No gender diff
- Gender diff

**TOPIC-CONT.**

**OVERLOOKED IN PREVIOUS L2 RESEARCH**
- Ambiguous
- Uneconomical
- Ambiguous
- Redundant
CEDEL2 [Corpus Escrito del Español L2]
(Lozano 2008, Lozano & Mendikoetxea, 2013)

http://wdb.ugr.es/~zanos/cdw/

- **CEDEL2:**
  - L1 English – L2 Spanish
  - c. 750,000 words, 2,400 participants
  - All proficiency levels (beginner, intermediate, advanced)
  - Also: comparable native Spanish subcorpus.

- **Data tagged:**
  - Only [+human] [3rd person] [sing] /[plu] syntactic subjects
    - i.e., human characters in discourse, with shift of referents in description of a film (unknown to the reader to avoid bias in overuse of Ø in Topic-Shift contexts due to reader’s background knowledge of known stories like Little Red Riding Hood (cf. Montrul & R.Louro 2006)).

- **Software:** **UAM Corpus Tool** (O’Donnell 2009) ➔ tagging & stats.
## Participants’ biodata

- learners (very advanced)
- native (control group)

<table>
<thead>
<tr>
<th>Filename</th>
<th>Age</th>
<th>Proficiency</th>
<th>Self prof.</th>
<th>Stay length</th>
</tr>
</thead>
<tbody>
<tr>
<td>93_30_10_3_JM</td>
<td>30</td>
<td>40</td>
<td>4.5</td>
<td>9</td>
</tr>
<tr>
<td>93_20_4.5_3_KB</td>
<td>20</td>
<td>40</td>
<td>5.25</td>
<td>4.25</td>
</tr>
<tr>
<td>95_28_15_3_KDH</td>
<td>28</td>
<td>41</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>95_43_12_3_ML</td>
<td>43</td>
<td>41</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>95_58_3_3_SAR</td>
<td>58</td>
<td>41</td>
<td>5.25</td>
<td>12</td>
</tr>
<tr>
<td>98_49_10_3_GWB</td>
<td>49</td>
<td>42</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>98_20_8_3_JEL</td>
<td>20</td>
<td>42</td>
<td>5.5</td>
<td>7</td>
</tr>
<tr>
<td>98_48_11_3_OPE</td>
<td>48</td>
<td>42</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>98_47_29_3_TLS</td>
<td>47</td>
<td>42</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>100_28_8_3_KEM</td>
<td>28</td>
<td>43</td>
<td>5.5</td>
<td>10</td>
</tr>
</tbody>
</table>

**AVERAGE**  37.1  41.4  5.3  13.3

<table>
<thead>
<tr>
<th>Filename</th>
<th>Age</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>24_3_AAV</td>
<td>24</td>
<td>Spain</td>
</tr>
<tr>
<td>33_3_CMM</td>
<td>33</td>
<td>Spain</td>
</tr>
<tr>
<td>21_3_CPV</td>
<td>21</td>
<td>Spain</td>
</tr>
<tr>
<td>51_3_GDC</td>
<td>51</td>
<td>Spain</td>
</tr>
<tr>
<td>25_3_MCC</td>
<td>25</td>
<td>Argentina</td>
</tr>
<tr>
<td>32_3_MDD</td>
<td>32</td>
<td>Spain</td>
</tr>
<tr>
<td>58_3_MI</td>
<td>58</td>
<td>Mexico</td>
</tr>
<tr>
<td>45_3_RP</td>
<td>45</td>
<td>Spain</td>
</tr>
<tr>
<td>26_3_RSZ</td>
<td>26</td>
<td>Argentina</td>
</tr>
<tr>
<td>30_3_SPH</td>
<td>30</td>
<td>Spain</td>
</tr>
</tbody>
</table>

**AVERAGE** 34.5
## Corpus sample details

<table>
<thead>
<tr>
<th>Corpus</th>
<th># texts</th>
<th># words</th>
<th># tagged subjects</th>
<th>Tags/words ratio</th>
<th># terminal tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced L2ers</td>
<td>10</td>
<td>5,172</td>
<td>330</td>
<td>6.38%</td>
<td>5,874</td>
</tr>
<tr>
<td>Spanish natives</td>
<td>10</td>
<td>3,206</td>
<td>168</td>
<td>5.24%</td>
<td>2,976</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>20</strong></td>
<td><strong>8,378</strong></td>
<td>498</td>
<td><strong>5.24%</strong></td>
<td><strong>8,850</strong></td>
</tr>
</tbody>
</table>
Novelty of this study

- Fine-grained, theoretically motivated tagset (only possible via manual tagging).

- Tagging of both the anaphor and also the antecedent (← overlooked in previous studies).
  - In particular, 2 key properties of antecedents:
    - No. of potential (=competing) antecedents (2+)
    - Gender of antecedents
1) Syntax-discourse interface deficits are persistently problematic in very advanced L2ers even after long exposure to L2 \( \rightarrow \)
supports Interface Hypo (Sorace 2011).

2) BUT not all deficits are equally problematic.
   - In particular, L2ers’ deficits are better accounted for 2 basic pragmatic principles
     - Redundancy
     - Ambiguity
   - ...rather than by over-/under-production of pronominal subjects (as previously reported in L2 literature).

3) Specifically, L2ers: redundant > ambiguous
   - As also found in L1 Spanish development \( \rightarrow \) perhaps general tendency in language acquisition (?)
     - (Shin & Smith-Cairns 2009, 2011)
RESULTS

DATA ANALYSIS with in-built statistical software in UAM-Corpus Tools
• analysis of frequencies
• $X^2$ tests to check for significant differences
Syntax of TOPIC CONTINUITY

- **NATIVES** produce mostly **null** subject (expected), and just a few residual **overt** and **NPs** (unexpected...according to the literature!)

- **LEARNERS** produce:
  - **null**, as expected (80%)
  - but also redundant **overt** as previously reported (14%)
  - and, crucially, some redundant **NP**, previously unattested in L2 (5%)

Learners≠natives (sig), despite their advanced level and long exposure to L2 Spa.
Syntax of TOPIC SHIFT

- In top-shift, we find the opposite pattern than before:

  - **LEARNERS:**
    - many *overt* (23%), as expected and similarly to natives (19%)………… Pedro María … él
    - but they clearly prefer *NP* (61%), previously unattested………… Pedro María … #Pedro
    - natives also do so (71%) Pedro Antonio … Pedro
    - Also some ambiguous *null* pronouns (14%)…………………………… Pedro María … #Ø

- So, fine-grained annotation reveals that in topic shift both natives and L2ers prefer an *NP* in natural discourse ➔ we will discuss later why.

- Learners=natives with *overt and NP*, so learners are native-like in Top-Shift.
**Pragmatic infelicity of anaphor**

- **Natives**: just some occasional infelicity (only 4 tokens out of 168 = 2.4%)

- **Learners**: 42/330 = 13% infelicitous, out of these:

  - In less than 1/3 of the time, they are **ambiguous** in topic-shift contexts, by not producing informative anaphors when required:
    - Null-when-overt (54%) ................................................. Pedro María ... #Ø
    - Null-when-NP (38%) ...................................................... ...Pedro Antonio ... #Ø
    - Overt-when-NP (8%) .................................................... ...Pedro Antonio ... #él

  - In over 2/3 of the time, in **topic-continuity** contexts, they produce **redundant** material:
    - Overt-when-null (93%) ................................................... Pedro ... #él
    - Some NP-when-null (6.9%) .............................................. Pedro ... #Pedro
- In topic-cont L2ers are frequently redundant: many overt pronouns and occasional full NPs (unattested in L2 lit)

- In topic-shift L2ers are sporadically ambiguous

- So, L2ers are more redundant than ambiguous → we’ll see why.
In topic-shift contexts,
- we have observed a high production of NPs
- but traditional studies report high uses of overt pronouns.

So, what forces production of NP instead of overt?

Let’s consider two crucial factors which have been overlooked in the literature:

- No. of potential antecedents: 2 or 3
- Gender of those antecedents
No. of potential antecedents

- 2 antecedents: overt > NP
- 3 antecedents: NP >> overt

So, choice of anaphor (NP vs overt) depends on the no. of potential antecedents (previously unreported)
Let’s consider clear-cut topic-shift contexts: 2 antecedents only.

When no gender diff in antecedents (Pedro/Antonio), everybody prefers NP > overt anaphor to avoid ambiguity.
- Pedro Antonio … Pedro

When gender diff in antecedents (Pedro/María), everybody prefers overt > NP anaphor so as to be economical.
- Pedro María … él

So, choice of NP vs overt also depends on gender diffs between antecedents ➔ unexplored in L2 lit.

And learners are sensitive to this!
DISCUSSION AND CONCLUSIONS
L2ers show **pragmatic deficits**: 
- They prefer being redundant > ambiguous.
- i.e., overinformative > underinformative. WHY??

L2ers are more aware of **topic-shift** constraints because violation of these \(\rightarrow\) **communicative breakdown** (ambiguity).

L2ers are more relaxed with **[top-continuity]** constraints because violation of these \(\rightarrow\) **NOT communicative breakdown**, but simply **redundancy**.
Question is: Better to be **redundant** than **ambiguous** in SLA?? Also reported for ...

- **Spa children**, who learn first to avoid ambiguity (overt in top-shift) than to be redundant (overt in top-cont) (Shin & Smith-Cairns 2010, 2012)
- **Adult natives** of other langs (eg., English) (Engelhardt et al 2006)

**L2ers violate** Grice’s **Maxim of Quantity** (“Don’t say more than is required”).

- Based on neo-gricean approaches to AR in native Spanish (Blackwell 1998, Geluykens 2013):
  - Natives **occasionally violate Informativeness/Economy Principle** (“Produce minimal anaphoric forms to achieve your communicative ends”)
  - Natives **rarely violate Manner/Clarity Principle** (“Do not use complex anaphoric forms without reason”)
  - AND L2 LEARNERS ARE NO EXCEPTION TO THIS!

**Pragmatic Principles Violation Hypothesis (PPVH)**

- Mild pragmatic violations (**redundancy**) are typical even in advanced L2 grammars
- Strong pragmatic violations (**ambiguity**) are occasional

<table>
<thead>
<tr>
<th>Violation strength:</th>
<th>STRONG</th>
<th>MILD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle violated:</td>
<td>Manner/Clarity</td>
<td>Informativeness/Economy</td>
</tr>
<tr>
<td>Violation type:</td>
<td>Ambiguity</td>
<td>Redundant&lt;sub&gt;1&lt;/sub&gt;anteced Redundant&lt;sub&gt;2&lt;/sub&gt;anteced</td>
</tr>
</tbody>
</table>
Final conclusion: LCR and SLA

- Fine-grained, theoretically-motivated annotation ...

- coupled with powerful tagging analysis software & statistical package (UAM Corpus Tool) ...

- reveals findings that have gone previously unnoticed in experimental (and corpus) studies.

THANK YOU - GRACIAS


