Using learner corpus tools in SLA research: the morpheme order studies revisited

Ana Díaz-Negrillo
anadiaznegrillo@ugr.es

Cristóbal Lozano
cristoballozanno@ugr.es
http://wdb.ugr.es/~cristoballozanno

Corpus Linguistics Conference
Lancaster, UK
22th – 26th July, 2013
Revisiting the morpheme order studies (MOS) (1)

- The MOS (70s-80s) have been crucial in our understanding of IL in the SLA of English:
  - A remarkably consistent sequence independently of ...:
    - the learners’ mother tongue (L1), age and learning environment
    - the testing method and the measuring instrument

1. progressive –ing
2. contractible copula –’s
3. plural –s

4. articles a(n)/the
5. contractible auxiliary (be) –’s
6. irregular past

7. regular past –ed
8. 3rd person singular –s
9. possessive –’s

- Similar sequencing in child L1 English
- Different theoretical explanations: nativism (natural order), perceptual saliency, grammatical factors, etc.
- For overviews: Hawkins & Lozano 2006; Kwon 2005; Goldschneider & DeKeyser 2001
Revisiting the morpheme order studies (MOS) (2)

- Why are the MOS relevant for SLA and LCR?

- The MOS is a recently revived and controversial topic in SLA research (Goldschneider & DeKeyser 2001; Kwon 2005; Luk & Shirai 2009; Tono 2000)

  “The order that learners follow constitutes one of the most important ‘facts’ that any theory of L2 acquisition must account for”
  (Ellis & Barkhuizen 2005: 91-92)

  “The study of learners’ use of morphemes through obligatory occasion analysis still has much to offer SLA. The descriptive information it provides serves as a basis for testing the validity of different explanations of the order of acquisition.”
  (Ellis & Barkhuizen 2005: 79)
Objectives

Present an approach in LCR considering SLA as a point of departure

- replicating MOS
  - replication in SLA is a necessary condition to (dis)confirm previous findings and to eliminate possible biases in the research method (Porte 2012)

- using a different methodology
  - learner corpora and corpus tagging

Promote dialogue and synergies between LCR and SLA research (Tono 2003, Myles 2007)

“Learner corpus researchers should exchange ideas with SLA researchers in a more structured and systematic way. Many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them effectively, while SLA researchers typically know little about what corpora can do for them. By improving the communication lines, we will be able to learn from each other.” (Tono 2003: 806)
Methodological limitations of previous research (1)

- (Quasi)experimental methods have traditionally been used in the MOS:
  - **small L2 samples** under controlled conditions (except for Tono 2000; McEnery, Xiao & Tono 2006)
  - **native-oriented** approach (Ellis & Barkhuizen 2005: 92): unable to tell us about the forms that arise in learners’ Interlanguage (IL)
    - Bley-Vroman’s (1983) **Comparative Fallacy** (see slides later)
  - **coarse-grained** in their analysis of learner productions since they do not fully explore all the subtypes of errors typical of learners’ IL (*stealed, *stoled, *foots, *feets, etc.)
    - We consider: **U-shaped learning** and the **Dual Mechanism** (see slide later)
    - We consider: **Asymmetry in irreg. vs. reg. forms:**

![Diagram showing examples of irregular and regular forms]

- **Irregular forms**: stole, steal__, children, child__, stole, stealed
- **Regular forms**: worked, work__, books, book__
Methodological limitations of previous research (2)

- Our approach aims to compensate limitations of MOS and LCR
  - by combining the methodological strengths of LCR and the theoretical explanatory power of SLA in MOS
  - for a fully-rounded picture of the acquisition of L2 English morphemes we need to triangulate:

  previous findings
  (SLA theory and IL experimental data)

  learner corpus data
  (naturalistic data)

  corpus-based tools
  (fine-grained IL annotation)
Our approach (1)

1 Learner corpus data
   - COREFL, CORpus of English as a Foreign Language
   - Narrative written EFL texts, *Frog where are you?* - L1 Spanish
   - Age: 12-17 (secondary school)
   - Size: approx. 100,000 words
   - Ongoing (2012-)

2 Corpus-related methodology combined with SLA:
   - It moves away from bottom-up / corpus-driven / hypothesis-finding ← descriptive accounts of learner performance in LCR
   - It takes a top-down / corpus-based / hypothesis-testing approach (cf. Myles 2005, 2007)
Our approach (2)

Corpus techniques combined with SLA: IL Annotation (ILA)

It moves away from the coarse-grained, all-purpose tagging of learners’ errors. (cf., for example, Dagneaux et al. 1996; see Díaz-Negrillo & Fernández-Domínguez 2006 for an overview of error tagsets)

- **purposed-oriented**: designed for the study of morpheme acquisition.
- **fine-grained**: it categorises learner performance in detailed categories based on previous IL theory and findings.
- it considers both **non-target like (NTLU)** and **target-like (TLU)** uses.
✓ the tagset considers **all subtypes of NTLU uses**, some of which have been overlooked in previous tagging systems — rich tagset

<table>
<thead>
<tr>
<th>OC: Past irreg (Peter stole yesterday)</th>
<th>S: Supplied form</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target-like Use</strong></td>
<td></td>
</tr>
<tr>
<td>(correct form supplied)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-target-like Use</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Underuse</strong></td>
<td>Peter stole yesterday</td>
</tr>
<tr>
<td>(no form supplied)</td>
<td>[OC:past_irreg ]</td>
</tr>
<tr>
<td><strong>Misuse</strong></td>
<td>[S:past_irreg ]</td>
</tr>
<tr>
<td>(incorrect form supplied)</td>
<td></td>
</tr>
<tr>
<td><strong>Misselection</strong></td>
<td>Peter steal__ yesterday</td>
</tr>
<tr>
<td>(form exists)</td>
<td>[OC:past_irreg ]</td>
</tr>
<tr>
<td><strong>Misrealisation</strong></td>
<td>[S:Ø ]</td>
</tr>
<tr>
<td>(form does not exist)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IRREGULAR</th>
<th>PAST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OC: 3rd sing (Peter never stole [=steals])</strong></td>
<td><strong>SNOC</strong></td>
</tr>
<tr>
<td><strong>Overuse</strong></td>
<td>Peter never stole</td>
</tr>
<tr>
<td>(correct form supplied but in NOC)</td>
<td>[OC:3rd sing ]</td>
</tr>
<tr>
<td></td>
<td>[S:past_irreg ]</td>
</tr>
</tbody>
</table>

**Examples:**
- Peter stole yesterday
- Peter steal__ yesterday
- Peter stealing yesterday
- Peter stealed yesterday
- Peter stoled yesterday
It considers a bi-contextual approach: both **obligatory contexts (OC)** and **non-obligatory contexts (SNOC)**

*the boy and the dog falled into the river*

- **falled**: OC irregular past ➔ misrealization (=misformation)
- **falled**: NOC regular past ➔ overuse (SNOC)

It considers a **bi-layered approach**: the **native** and the **non-native (IL)** perspective so as to overcome the ‘Comparative Fallacy’ (Bley-Vroman 1983),

- **e.g. OC: reg. past**

  *And not wanted* (Target: “And he didn’t want”)
  
  **Native layer**: Overuse (SNOC)
  **IL layer**: TLU

*they climbed up into a tree*

**Native layer**: TLU
**IL layer**: TLU
IL Scoring (ILS) (frequency-based)

\[
ILS = \frac{N \text{ correct suppliance in obligatory contexts} + (N \text{ SMOC} \times 0.5)}{N \text{ obligatory contexts} + (N \text{ suppliance in non-obligatory contexts})} = \frac{SOC + (SMOC \times 0.5)}{OC + SNOC}
\]
Our learner corpus analysis with ILT

- **Corpus: COREFL**
  - sample of approx. 5,000 words
  - 44 texts
  - A2 and B1 levels (years 1-3, secondary education)

- **Interlanguage Annotion (ILA)**
  UAM corpus tool

<table>
<thead>
<tr>
<th></th>
<th>Irregular past</th>
<th>Regular past</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>94</td>
<td>80</td>
</tr>
<tr>
<td>B1</td>
<td>157</td>
<td>153</td>
</tr>
<tr>
<td><strong>TOTAL TAGS</strong></td>
<td><strong>251</strong></td>
<td><strong>233</strong></td>
</tr>
</tbody>
</table>

- **Interlanguage Scoring (ILS)**

\[
SOC + (SMOC \times 0.5) = \frac{OC + SNOC}{Work in progress}
\]
A bit of experimental evidence on regular vs. irregular past before interpreting the corpus evidence...

- **U-shaped learning** → **Dual Mechanism** for processing irregular vs. regular morphology (Pinker 1998).

- Observed in **L1** (Marcus et al. 1992, Pinker 1995) and **L2** (Zobl 1998, Birdsong & Flege 2001, Murphy 2004), *inter alia* --- but only L2 experimental evidence, **no corpus data**.

---

![Graph showing U-shaped learning](image)

- **Form in associative memory:**
  - *BOUGHT MICE*

- **Acquisition of symbolic rule**
  - STEM+ “ed” / “s”
    - *BUYED*
    - *MOUSES*

- **Blocking of rule:**
  - *BOUGHT MICE*
Learner corpus analysis with ILA: results (1)

Regular past

Irregular past

Let’s explore each of these in detail...
Learner corpus analysis with ILA: results (2)

➢ **TLU vs. NTLU**

**Regular past**
- *were sleeping the frog escaped from the vase*
- *while the boy was sleeping, the frog scape*

**Irregular past**
- *Then they saw a deer*
- *They leave the forest and moved the hand*

- **Development from A2 to B1**: significant and drastic decrease in NTLU for both regular and irregular past (p<0.05) ➔ L2ers start to acquire past tense from intermediate stages (B1 onwards).
Learner corpus analysis with ILA: results (3)

- **TLU**

- Inverted results for A2 and B1 groups
- It is only at **B1** (low intermediate) that *irregular* > *regular* past (p<0.05) ➔ irregular forms precede regular forms (in line with MOS)
### Learner corpus analysis with ILA: results (4)

#### NTLU

##### Regular past
- While the boy was sleeping, the frog **scape**
- He searches for all over the river
- A deer caught **ed** the boy.

#### Irregular past
- The boy **go** to sleep because was latter
- A deer **catch**ed the boy
- He don’t **found** the frog.

<table>
<thead>
<tr>
<th></th>
<th>A2</th>
<th>B1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underuse</td>
<td>78.3</td>
<td>78.1</td>
</tr>
<tr>
<td>Misuse</td>
<td>16.9</td>
<td>9.8</td>
</tr>
<tr>
<td>Overuse (SNOC)</td>
<td>4.8</td>
<td>10.9</td>
</tr>
</tbody>
</table>

- **Underuse** is by far the most frequent error at all levels and with both morphemes ➔ learners have not fully acquired yet the inflected forms (-ed) and the irregular forms.
- **Misuse**: irregular>regular at both levels ➔ to be discussed in detail later
- **Overuse** is the least frequent tag in all levels and in both morphemes

Let’s explore each of these in detail...
**Learner corpus analysis with ILA: results (5)**

> **NTLU 1: Underuse**

- **Regular past**
  - production of Ø morphology is **stable** across levels; more **likely** to appear with **regular verbs** (irregular morph. is listed in associative memory in the mental lexicon).
  - Not even the inflection for 3\(^{rd}\) ps. sing.. This is more frequent in A2 learners.

- **Irregular past**
  - a NTLU **decrease** from A2 (70\%) to B1 (63\%) **signals TLU of irregulars** (recall: irreg>reg in intermediates).
  - some **frequent irreg verbs** are **inflected** (saw, went vs. hold, fall) \(\rightarrow\) high frequency prevents overregularizations according to Blocking Principle in ‘**Dual mechanism**’ (Marcus et al 1992).
Learner corpus analysis with ILA: results (6)

- **NTLU 2: Overuse (SNOC)**

![Bar chart showing overuse of regular and irregular past tense forms. The chart shows that at A2 level, there are 4.8% regular past and 0% irregular past, while at B1 level, there are 11% regular past and 7% irregular past.]

**Regular past**
- An increase in overuse of **-ed** morpheme in irregular past contexts (4.8% at A2 and 11% at B1) reflects overregularisation at **intermediate** (B1) stages, as predicted by ‘Dual Mechanism’ model.

**Irregular past**
- All examples involve negative constructions (results to be taken cautiously).
- Double marking strategy??? **[PAST] → wasn’t + irreg_past**
Learner corpus analysis with ILA: results (7)

- **NTLU 3: Misuse (misselect vs. misrealization)**

  **Regular Past**
  - because tries to run away
  - The boy and the dog worrieds

  **Irregular Past**
  - then he falls
  - they broked their legs

<table>
<thead>
<tr>
<th></th>
<th>Misselect</th>
<th>Misrealiz</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>92,7</td>
<td>7,1</td>
</tr>
<tr>
<td>B1</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Misselect</th>
<th>Misrealiz</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>91,3</td>
<td>8,7</td>
</tr>
<tr>
<td>B1</td>
<td>47,4</td>
<td>52,6</td>
</tr>
</tbody>
</table>

  Imbalance regular vs irregular past:
  - **Regular past:**
    - misselection > misrealiz at all proficiency levels.
    - 93% + errors are misselect. of 3rd sing –s: escapes (=escaped) etc.
    - only 7% errors are misrealiz (agreement added to past tense): worrieds (=worried)
  - **Irregular past: Proficiency effect**
    - A2 (beginners): misselection > misrealiz: again 3rd singular –s: falls (=fell) etc.
    - B1 (low interm.): misselection ≤ misrealiz: failed (=fell) etc. ➔ overregularization clearly starts at intermediate stages (Dual Mechanism)
Conclusion

This study has illustrated a different approach in LCR which
- sets off from SLA theory...
- uses learner corpus research methods...
- proposes ILA (Interlanguage Annotation)

Future work
- annotation of the corpus for the rest of the morphemes
- further exploration of the bi-layered approach
- further specification of the annotation categories based on SLA findings:
  - tense-aspect categories: telicity, accomplishments, states, etc.
  - interface with other aspects: negation, passivization, etc.
- triangulation of corpus data with experimental data
References


References (2)


