Formalizing the Second Language Learner Corpus by Means of Automatic Analysis

Aurélia Marcus marcus@sinequa.com
Sinequa
Marzena Watorek watorek@univ-paris8.fr
Université de Paris VIII (SDL) & UMR 7023-CNRS
OUTLINE

- Claims of this study
- The theoretical Second Language Acquisition framework
- The automatic analysis of an L2 adult oral corpus
  - Data
    - A first trial
- The results
- Discussion
- Perspectives
- Concluding remarks
Claims of this study

- Automatic analysis and formalisation of L2 learner productions
- Comparison with manual qualitative analysis
- Creation of efficient NLP Tools for L2 data
Theoretical framework for the learner language study

- **Learner variety approach**
  (idiosyncratic dialect, Corder 1967, interlanguage, Selinker 1972, learner variety, Klein 1987)

  « **Interlanguage** is the type of language produced by second- and foreign- language learners who are in the process of learning a language. In language learning, learner’s errors are caused by several different processes. These include:
  
a. borrowing patterns from the mother tongue
b. extending patterns from the target language
c. expressing meanings using the words and grammar which are already known » (Richards & al. 1992)
LEARNER VARIETIES

Example: Italian learner of English
Another man prove for open the door
(another man tries to open the door)
ESF Project
(cf. Perdue, 1993)

Second Language Acquisition by Adult Immigrants
- The functionalist research project
  - how do learners gradually acquire an approximative system (IL) in order to accomplish their communicative needs/goals?

- Longitudinal and cross-linguistic data
  - Corpus available on the IMDI portal at the MPI for psycholinguistics:
  - CHAT format (transcription convention CHILDES)
ESF Database
The main result of the ESF project

Three acquisitional stages

- **Pre-basic variety** (nominal structure)
  ex. 1  *I lift* (I took the elevator)

- **Basic variety** (non-finite verb structure)
  ex. 2  Another man *prove* for open the door
  (another man tries to open the door)

- **Post-basic variety** (finite verb structure)
  ex. 3  he *working* only for ten minutes
The basic variety

The interaction of three types of constraints
> phrasal constraints - three basic syntactic patterns with non finite verb

NP1
Proper Noun
(Det) Noun
Pro
∅

NP1 V (NP2)? (NP2)?
NP1 Cop (ADJ or NP2 or PP)
(V or Cop) NP2

NP2
Proper Noun
(Det) Noun

> Semantic constraints - ‘controller first’
*The NP-referent with highest control comes first*

> Pragmatic constraints - ‘focus last’
*The focus expression is last in the utterance*
Automatic analysis with a tagger

- **POS-Tagger**
  1. tokenises words, punctuation and sentences
  2. makes a lexical analysis of each word (with the help of lexicons)
  3. Disambiguates according to the language model of the program

- **Our goal**
  - Good results with small corpus
  - Guesser for unknown words
  - Resources for several languages (cf. TL in the ESF corpus: French, English, Dutch, German and Swedish)
  - Possibility to add additional dictionary for the interlanguage
    - Ex. *sepa* (standard French « Je ne sais pas », « I don’t know ») lemmatized VER pas_savoir
  - Keep morpho-syntactic ambiguity for eventual parsing
The automatic analysis:
*a sample of ESF corpus*

The Data: Spanish LS>French LC (Berta & Alfonso)

<table>
<thead>
<tr>
<th>Acquisitional stage</th>
<th>Corresponding Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-basic variety (pre BV)</td>
<td>Berta 1st cycle</td>
</tr>
<tr>
<td>Basic variety (BV)</td>
<td>Berta 2nd and 3rd cycle</td>
</tr>
<tr>
<td></td>
<td>Alfonso 1st and 2nd cycle</td>
</tr>
<tr>
<td>Post-basic variety (post BV)</td>
<td>Alfonso 3rd cycle</td>
</tr>
</tbody>
</table>
Automatic Analysis: first trial

1. **CORPUS**

2. **Script formatting for TreeTagger**

3. **Tagging (French)**

4. **TreeTagger format**
   - la dama@ &sapel

5. **TreeTagger output**
   - la DET:ART le (the)
   - dama@ NOM dame (lady)
   - &sapel VER se_appeler (be called)

6. **Semi-manual correction of tagged corpus**
   - based on the qualitative analysis led by Klein and Perdue (1997)

*TreeTagger, University of Stuttgart: [http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/](http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/)

- word@ = SL
- &word = IL
Results of the trial

- Percentage of good guesses on unknown words (ie IL) - ≈ 50%

- Distribution of the IL forms in the four main categories:
  > Verb
  > Nominal elements
  > Ambiguous V/N
  > Various grammatical elements (Adv-Det-Prep)

- Evolution across the 3 acquisitional stages
Results of the trial

- Automatic formalisation of TreeTagger output

&el PRO il elle ils elles (he, she, they)
&parla VER:simp parler (talk)
avec PRP avec (with)
moi PRO:PER moi (me)

P -> N VER PP
Results of Trial

- Encountered problem
  - Ambiguity of grammatical categories and lemmatization

```
&se

V

Dem

Pro

savoir
know

être
be

c'est
ce
ces
se
ses

sais
sait

son
≈his, her
```
Discussion

- Automatic analysis matches with qualitative analysis
  - Decrease of the nominal elements
  - Correlation between the verb item increase and the acquisition process

*development from the pre BV to the post BV*
## Discussion

### Difficulties for a tagger

The differences between the IL and the TL

<table>
<thead>
<tr>
<th>Interlanguage</th>
<th>Target language</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;Instable system</td>
<td>&gt;Stable system</td>
</tr>
<tr>
<td>&gt;Simple basic syntactic patterns independent of LS and TL</td>
<td>&gt;Complex morpho-syntactic system partially shared by certain languages</td>
</tr>
<tr>
<td>&gt;No verbal morphology</td>
<td>&gt;Lexical items set predictable according to the language</td>
</tr>
<tr>
<td>&gt;Lexical items set: dependent on both the SL/TL systems and on the individual characteristics (IL=idiolect)</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

Two main questions for our future research:

- Which properties of the learners varieties are trainable for tagger resources specific to the IL?
- Which properties of the learners varieties can be tricky for a tagger?
Perspectives

- Use the ESF corpus (the entire corpus) as a training corpus in order to create resources for an IL tagger

- Grammatical formalisation of the obtained results

- Evaluate the possibility of making a tagger usable on L2 data similar to the ESF
  - Oral productions
  - Absolute beginners
  - Natural acquisition
Concluding remarks

- Performance of a tagger on an L2 corpus
  - Low score of good guesses not an obstacle
    - 70% good tags already very useful for pre-analysing and/or for a quantitative analysis tool to back-up qualitative analysis
  - Limits of NLP in SLA
    - Instability of IL
    - IL = idiolect