Is fully-automated corpus-based language acquisition research feasible?

Tomonori Nagano <tnagano@gc.cuny.edu> and Virginia Valian, The City University of New York

Introduction

The CHILDES database potentially offers researchers the ability to analyze patterns of acquisition across a large number of children, but the full potential of the database has not yet been realized. In this study, we investigated the "feasibility" and "accuracy" of fully-automated corpus-based FLA research through a comparison between a manually-conducted corpus study by Valian (1991) and CHILDES.

Valian (1991) Study 1

Age, MLUs, the number of utterances, and the number of verbs

- **Valian (1991) & CHILDES**: Age, MLUs, the numbers of utterances, and the numbers of verbs are highly correlated (MLU: $r(19) = 0.91, p < .001$; utterance: $r(19) = 0.74, p < .001$; and verb: $r(19) = 0.96, p < .001$) between the original study and the CHILDES.

Valian (1991) Study 2

Frequencies of expletive subjects and pronominal subjects

- **Valian (1991)**: Expletive subjects were rarely used across all developmental stages (only 12 expletive sentences were found). Also, the distribution of cases in pronominal subjects suggests that children can correctly assign nominative case to the noun in the subject position.

- **CHILDES**: Thirteen expletive subjects are found in CHILDES. The distributions of the expletive subjects’ cases are almost identical to the original study.

<table>
<thead>
<tr>
<th>MLU</th>
<th>Valian (1991)</th>
<th>CHILDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>3.0</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>4.0</td>
<td>0.44</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Valian (1991) Study 3

Children’s use of subjects/expletive subjects in sentences with verb

- **Valian (1991)**: American children provide subjects at a high rate from the onset of combinatorial speech (evidence that they are aware that overt subjects are obligatory), and their usage correlates strongly with verb use even when MLU is partialled out. Usage does not correlate with modals once MLU is partialled out, so the lack of full subject use is not due to lack of INFL.

- **CHILDES**: The numbers of modals, semi-auxiliaries, and verbs with subject for each MLU group are almost identical to the original study. However, the partial correlations do not show the same pattern.

**Discussion and Conclusion**

The results from Valian’s original study and CHILDES data were comparable in many analyses, but were sometimes wildly different. In general, CHILDES does not do well when intuitive judgment is required.

- **Sentence type**: Valian (1991) excluded imperative and imitations from her utterance count, but it was nearly impossible to correctly determine those utterance types with the information currently available in CHILDES.

- **Transitivity**: Many verbs are polysemous between transitive and intransitive uses (e.g., *Brenda opened the door*; *The door opened*), but not (e.g., *Elmer already ate his meal*; *Elmer already ate*). It requires careful examination of the contexts to correctly determine the transitivity of verbs.

Those two factors might have contributed to the divergence of CHILDES from Valian’s original study (i.e., partial corrections in Study 3 and the classification of verbs in Study 5). For example, the grammatical dependency information in CHILDES is primarily based on structural definitions, not logical/sematic definitions of subject/object. Thus, John (1a) and (1c) are identified as subject as well as the glass in (1b), which is identified as a predicate without an object.

1a) John broke the glass.
1b) The glass broke.
1c) John broke.

**References**

References


