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

Tomonori Nagano and Virginia Valian
CUNY Graduate Center
contact: tnagano@gc.cuny.edu

The CHILDES database (MacWhinney, 2000) potentially offers researchers the ability to analyze patterns of acquisition across a large number of children, but two issues stand in the way of realizing the full potential of the database. One issue is the scarcity of published studies on the validity of automated corpus studies. Many information layers have been added to CHILDES (e.g., part-of-speech, morphological, tokenization layers), but what they contribute is rarely addressed. Various evaluation metrics are employed with the untested working assumption that high scores on those metrics directly translate into the validity of the categorized database. For example, Sagae et al. (2010) report that their argument-structure tags (e.g., SUBJ, OBJ, LOC etc.) achieved 92.79% accuracy in the children’s utterances, but a closer look shows that “accuracy” is an overall performance metric that encompasses many irrelevant and high-frequency tags such as PUNCTUATION and ROOT.

To investigate the areas of greater and lesser usefulness of CHILDES, we integrated the CHILDES database with Natural Language Toolkit (NLTK; Bird, Klein, & Loper, 2009) and conducted a fully automated replication of several findings from Valian (1991). Valian investigated the use of syntactic subjects and objects by 21 two-year-olds. Her data consist of frequency counts and percentages of subjects, objects, expletive subjects, modals, semi-auxiliaries, and verbs. Valian’s results, achieved manually, constitute a gold standard against which a fully automated analysis can be compared. The comparison sheds light on “usability” and possible applications of the CHILDES tools, such as the argument-structure layer.

The results from Valian’s original study and CHILDES/NLTK data were comparable in many analyses, but in some cases were wildly different. Strong correlations for mean length of utterances (MLU), the number of verbs, and, to a lesser degree, the number of children’s utterances were observed (MLU: $r(19)=0.91$, $p<.001$; verb: $r(19)=0.96$, $p<.001$; utterances: $r(19)=0.74$, $p<.001$). The correlation of the number of verbs that appeared with subject also showed a high correlation, $r(8)=0.95$, $p<0.001$. As shown in Fig. 1, the proportions of verbs, semi-auxiliary, and modals exhibit the same overall patterns by MLU group. Thus, overall tabulations based on individual word categories were very similar to the gold standard.

In contrast, the verb types with which the children used objects showed quite different patterns (Figs. 2a and 2b). In general, CHILDES does not do well when intuitive judgment is required. For example, Valian (1991) excluded imperatives and imitations from her utterance count, but it is nearly impossible to determine those utterance types with the information currently available in CHILDES. The automated analyses also have difficulty determining the transitivity of verbs. Many verbs are polysemous between transitive and intransitive uses and it requires careful examination of the contexts to correctly determine the transitivity of verbs. These issues, however, may be resolved by an addition of more information layers to CHILDES.

A highly automated FLA analysis with the CHILDES database can complement current practice in the observational FLA corpus research, but only if additional information is included, and only if the investigator is primarily concerned about patterns rather than accuracy for each data point.
References: