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

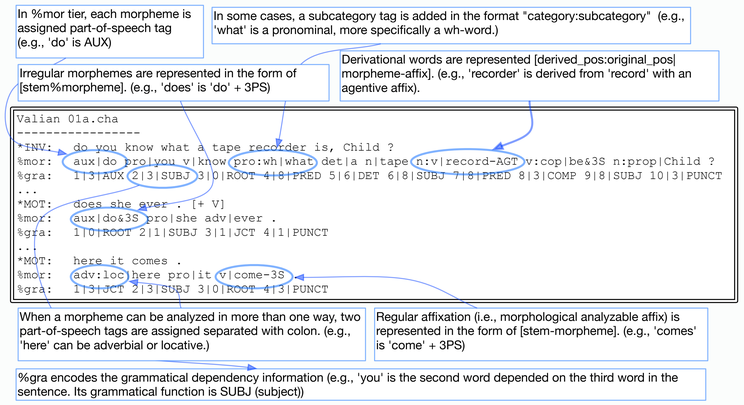
Tomonori Nagano <tnagano@gc.cuny.edu> and Virginia Valian, *The City University of New York*

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 (Valian, 1991) and CHILDES.

CHILDES (MacWhinney, 2000)

- morphosyntactic information with MOR (Hausser, 1989)
- ambiguity resolution with POST (Parisse & Le-Normand, 2000)
- grammatical dependency information with GRASP (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007, 2010)



- However, in many corpus-based FLA studies, only raw utterance data are used and these rich information layers (such as %mor and %gra) are practically discarded.

Research Question

- What are the "feasibility" and "accuracy" of automated analyses of the CHILDES data?

Methods

- Comparison between an automated analysis of CHILDES (using NLTK (Bird, Klein, & Loper, 2009)) and Valian's null-subject paper (Valian, 1991), a manually conducted corpus study that serves as a gold standard

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 with the original study.

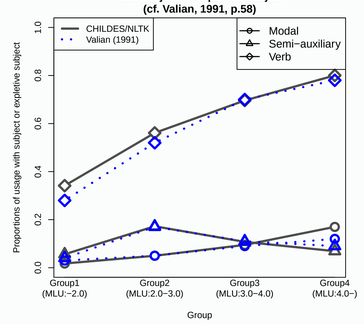
	Valian (1991)							CHILDES						
MLU	I	you	he/she	it	we	they	ACC	I	you	he/she	it	we	they	ACC
-2.0	0.63	0.05	0.15	0.13	0.02	0.02	0.00	0.52	0.06	0.18	0.19	0.02	0.02	0.00
2.0-3.0	0.66	0.06	0.07	0.12	0.02	0.04	0.02	0.63	0.07	0.08	0.14	0.02	0.04	0.03
3.0-4.0	0.49	0.17	0.14	0.12	0.03	0.05	0.00	0.45	0.18	0.15	0.13	0.03	0.04	0.00
4.0-	0.44	0.10	0.16	0.11	0.13	0.06	0.00	0.44	0.09	0.17	0.09	0.12	0.08	0.00

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 corrections do not show the same pattern.

Figure 1: Production of verbs, semi−auxiliaries, and modals with subject or expletive subject (cf. Valian, 1991, p.58)



	PARTIAL CORRELATION	
	VALIAN (1991)	CHILDES
MLU & modal (-age)	r=.43 (p=.056)	r=.73 (p<.001)
age & modal (-MLU)	r=.28 (p=n.s.)	r=.21 (p=.36)
subj & modal (-age & MLU)	r=-.04 (p=NA)	r=.29 (p=.32)
MLU & subject use (- age)	r=.48 (p=.03)	r=.43 (p=.04)
age & subject use (- MLU)	r=.41 (p=.075)	r=.24 (p=.28)
verb use & MLU (- age)	r=.81 (p<.001)	r=.28 (p=.22)
verb use & age (- MLU)	r=.20 (n.s.)	r=.28 (p=.21)
verb & subject (- age & MLU)	r=.78 (p<.001)	r=.34 (p=.13)

Valian (1991) Study 5

Children's use of objects with different types of verbs

- **Valian (1991)**: Children provided an object where it was necessary (i.e., transitive verb). The recognition of the distinction between obligatory and optional objects suggests that children pay attention to different verb uses in the input. Also, children increase their provision of objects for transitive/intransitive verbs between Group 1 and Group 2, suggesting a decrease of performance limitations.
- **CHILDES**: The data obtained from CHILDES show substantial divergence from the original study.

Figure 2a: Production of different types of verbs (cf. Valian, 1991, p.72)

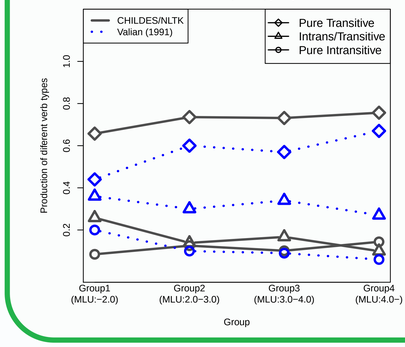
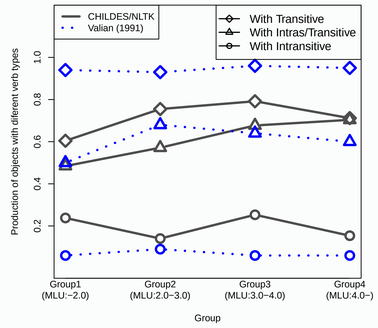


Figure 2b: Production of objects with different types of verbs (cf. Valian, 1991, p.73)



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 imperatives 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 **Brenda opened.*; cf. *Elmer already ate his meal.* ↔ *Elmer already ate.*) and 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/semantic definitions of subject/object. Thus, *John* in (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$_{\text{subject}}$ broke the glass.
(1b) The glass$_{\text{subject}}$ broke.
(1c) *John$_{\text{subject}}$ broke.

References

Valian, V. (1991). Syntactic subjects in the early speech of american and Italian children. *Cognition*, 40, 21-81.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Sebastopol, CA: O'reilly Media, Inc. (ISBN: 9780596516499; Course: ELX101; Price: $44.99)

Hausser, R. (1989). *Principles of computational morphology* (Vol. 47; Tech. Rep.). Pittsburgh, PA: Carnegie Mellon University.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum.

Parisse, C., & Le-Normand, M. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments and Computers*, *32*, 468-481.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *In proceedings of the acl 2007 workshops on cognitivee aspects of computational language acquisition* (p. xx-xx). Prague, Czech Republic.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, *37*(3), 705-729.

Valian, V. (1991). Syntactic subjects in the early speech of american and Italian children. *Cognition*, *40*, 21-81.