Abstract Syntactic Knowledge or Limited-Scope Formulae: A Computational Study of Children's Early Utterances

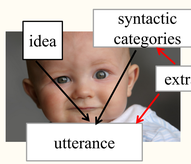
Qihui Xu, Martin Chodorow, Virginia Valian, Xiaomeng Ma

Contact: qxu@gradcenter.cuny.edu

Introduction

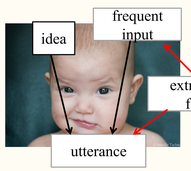
Where do children's early multi-word utterances come from?

Possibility 1: abstract syntactic knowledge



Children map concrete words to their abstract syntactic knowledge at onset of combinatory speech

Possibility 2: slot-filler formulae



Children primarily imitate adult input at onset of combinatory speech and only later acquire syntactic knowledge

Previous research about children's early multi-word utterances

- Abstract syntactic knowledge —> word combinations[1, 2]
- Word combinations —> abstract syntactic knowledge[3, 4]
- Look at children too old to provide evidence of early language acquisition
- Unable to tease apart syntactic knowledge from lexical development

Methods

Data

Training: mother's utterances (N = 1,272,139) from all Eng-NA corpora in CHILDES[5]

Testing: 7 children's multi-word utterances from longitudinal corpora[6, 7, 8, 9] in CHILDES

Table 1

utterances information of each child.

Child	Corpus	Utterances	Word types	Utterances by 20	Word types by 20
Ethan	Providence	9803	2084	4471	908
Naima	Providence	24304	3687	6874	1341
L	Brauwald	21345	2944	5708	969
Ross	MacWhinney	1747	620	1033	372
Cameron	Davis	4280	883	2130	484
Rebecca	Davis	14474	864	13430	802
June	Higginson	2458	340	1840	240

Analyses

Both the aggregated data of all children and individual children's data were analyzed.

- **Word-based Language Model:**
- **POS (part-of-speech)—based Language Model**

  Retrieve POS tags of training/testing utterances

POS tagging in CHILDES
MOR: automatic POS tagging
Accuracy: Adult language - 98%
Child language - 97%

- **Control testing on:**

  Randomized word/POS sequences

  Adult word/POS sequences

Statistical Language Model(LM)

A probability distribution over word/pos sequences

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

Training: adult lexical/ syntactic knowledge

Testing: children's lexical/syntactic knowledge

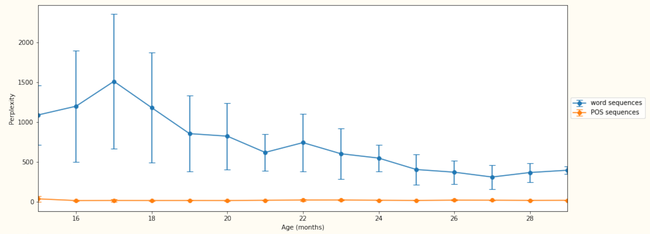
E.g. blue/adj car/n broken/part down/adv. (Ethan at 15 months)
P('blue car broken down') = P('broken|blue car) * P(down|car broken) * ....
P('adj n part adv') = P(part|adj n) * P(adv|n part) * ...

Results

Possibility 1: High word perplexity with low POS perplexity
Possibility 2: Word perplexity goes together with POS perplexity

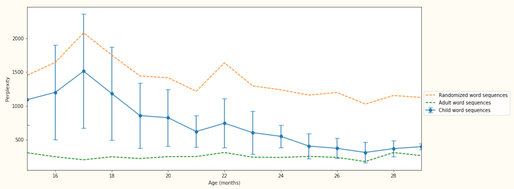
Aggregated data averaged across all children

Word combinations vs. POS combinations



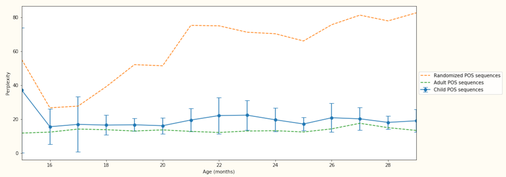
children have adult-like POS sequences even at a very early age when their word sequences are still not adult-like

Word combinations



Children's early word combinations have more similar perplexity with Randomized words than adult word combinations. Perplexity of word sequences keeps decreasing and reaches adult level at round 26 months.

POS combinations

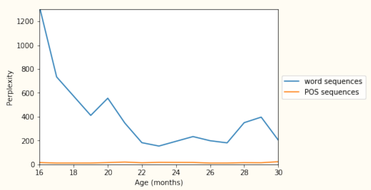


Children's early POS combinations are adult-like from 16 months on.

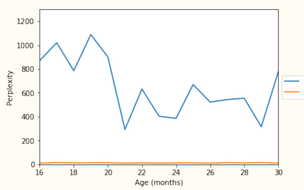
Individual children's data

Two Examples

Cameron



Ethan



# of children following the overall pattern:

7/7 children - POS

4/7 children - word

Discussion

Where do children's early multi-word utterances come from?

- High word perplexity with low POS perplexity - Abstract syntactic knowledge
- Word combinations: not adult-like at the beginning; gradually reaches adult level at around 26 months.
- POS combinations: constantly adult-like since 16 months.
- Children have adult-like syntactic knowledge even at a very early age when their word sequences are still not adult-like.

Limitations and future directions

- Accuracy of POS tags: need a more accurate tagger
- Sample size: increase sample size
- Age range: investigate data with wider age range
- Language: Replicate in different languages.

Take-home message

Children map concrete words to their abstract syntactic knowledge at onset of combinatory speech

References

1. Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of child language, 36*(4), 743-778.
2. Bencini, G. M., & Valian, V. V. (2008). Abstract sentence representations in 3-year-olds: Evidence from language production and comprehension. *Journal of memory and language, 59*(1), 97-113.
3. Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. A. Moravscik, H. Ouali, & K. M. Wheatley (Eds.), *Formulaic language*, Vol. *2*, pp. 297-322. Amsterdam, The Netherlands: John Benjamins.
4. Bannard, C., & Lieven, E. (2012). Formulaic language in L1 acquisition. *Annual Review of Applied Linguistics*, *32*, 3-16.
5. MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
6. Börschinger, Benjamin, Johnson, Mark, & Demuth, Katherine. 2013. A Joint Model of Word Segmentation and Phonological Variation for English Word-final /t/-Deletion. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol 1: Long Papers, 1508–1516, Sofia, Bulgaria.
7. Braunwald, S. R. (1978). Context, word and meaning: Toward a communicational analysis of lexical acquisition. In A. Lock (Ed.), *Action, gesture and symbol: The emergence of language*. London: Academic Press, pp. 485-527.
8. Davis, Barbara L., Peter F. MacNeilage & Christine L. Matyear (2002). Acquisition of serial complexity in speech production: A Comparison of Phonetic and Phonological Approaches to First Word Production. *Phonetica*, 59, 75-107.
9. Higginson, R. P. (1985). Fixing-assimilation in language acquisition. Unpublished doctoral dissertation. Washington State University.