"Eve Eat Dust Mop"

Measuring Syntactic Development in Child Language with Natural Language Processing and Machine Learning

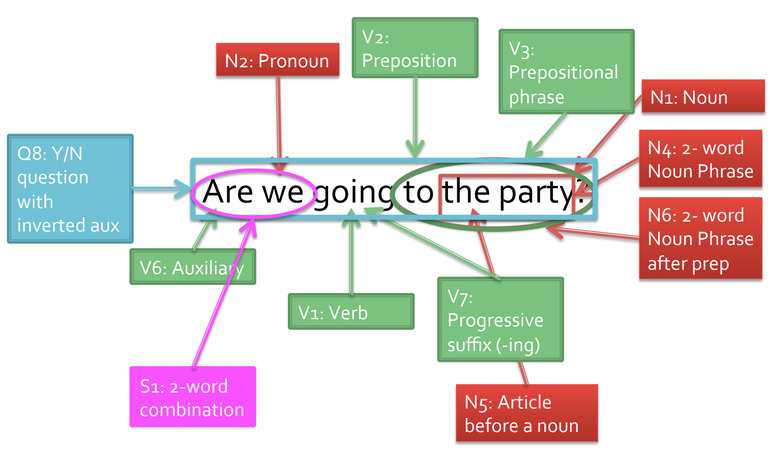
Shannon N. Lubetich (Pomona College '15)
Institute for Creative Technologies, University of Southern California

When measuring child language development, researchers often face choices between easily computable metrics focused on superficial aspects of language and more expressive metrics that rely on grammatical structure but require substantial labor. To advance research in child language development, we present an automatic scoring system facilitating easy analysis of large numbers of transcripts. Additionally, we explore a machine learning approach to produce scores of grammatical complexity based on extracting morphological and syntactic features of child utterances. Both techniques achieve accuracy similar to that of language researchers and reveal trends in syntactic development, offering promising results for future research and application. We can further apply our data-driven approach to predicting age, which does not require a previously-defined, language-specific inventory of grammatical structures.

Background: Index of Productive Syntax (IPSyn)

The higher the score, the more grammatically complex the language

IPSyn measures grammatical complexity of language by evaluating 100 successive utterances and awarding points based on defined "IPSyn Items."
There are 60 items that can earn up to 2 points each, resulting in a total possible score of 120 points.
This score can be broken down into subsections of Noun Phrases, Verb Phrases, Questions and Negations, and Sentence Structures.

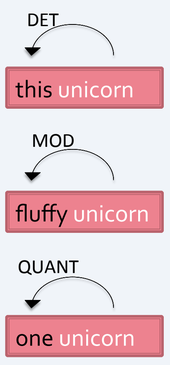


Automatic IPSyn

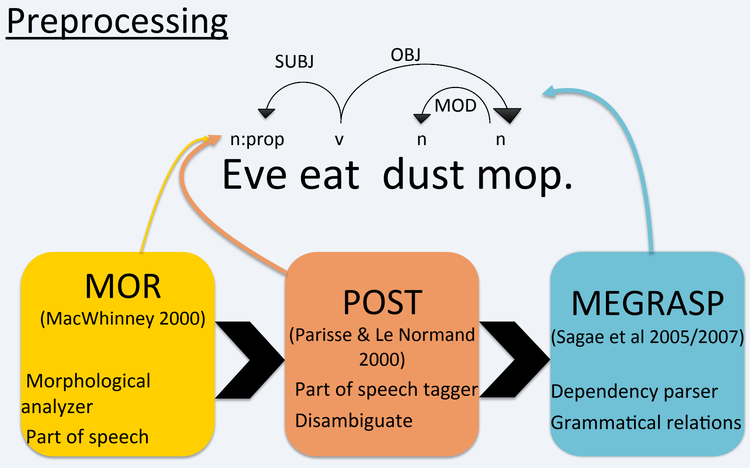
Goal: Create a program that, given a transcript, automatically calculates IPSyn score.

Defining IPSyn Items

Pattern: N4
Description: 2-word NP
GR: DET MOD QUANT
Parent requirements:
Index: forward-1
POS: n



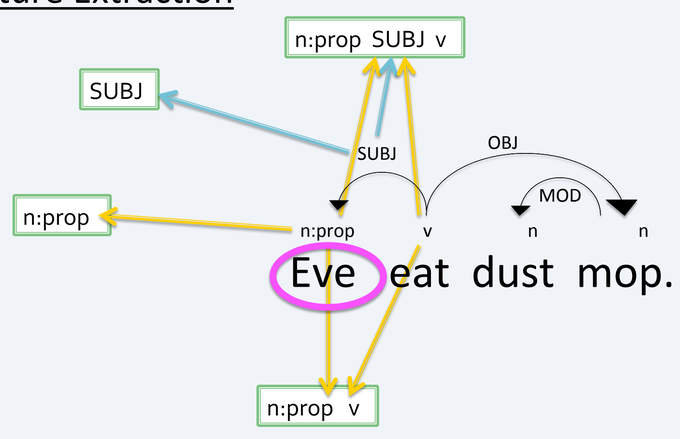
Preprocessing



Fully Data-Driven Measurement

Goal: Eliminate explicit IPSyn Items by training a machine on expert level annotations and scores so that it can produce an IPSyn score based solely on features of a transcript.

Feature Extraction

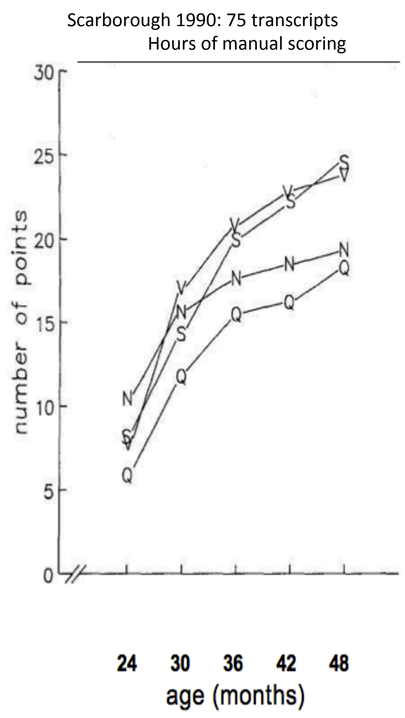


Results: Automatic IPSyn

Average absolute difference

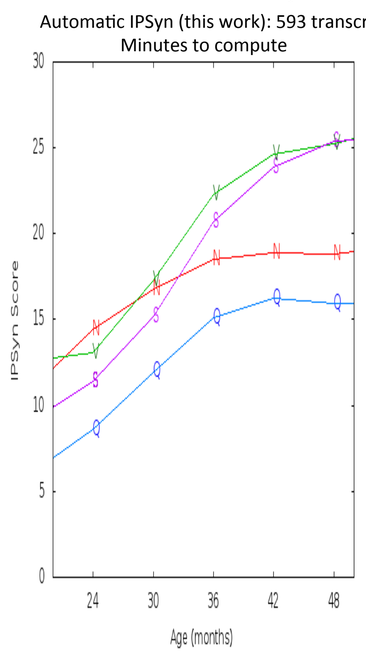
Manual: ~5 points

Scarborough 1990: 75 transcripts
Hours of manual scoring



Automatic: 3.6 points

Automatic IPSyn (this work): 593 transcripts
Minutes to compute



Data-Driven Approach to Age Prediction

Child (corpus)	Mean Abs Err	Pearson (r)
Adam (Brown)	2.5	0.93
Ross (MacWhinney)	3.7	0.84
Naomi (Sachs)	3.1	0.91

For children at least 3 yrs 4 months old

Child (corpus)	MLU *r*	IPSyn *r*	Regression *r*
Adam (Brown)	0.37†	0.53†	0.85†
Ross (MacWhinney)	0.19	0.34*	0.79†
Naomi (Sachs)	0.27	0.52	0.82†

†$p < 0.0001$
*$p < 0.05$

Future Work

- Improve patterns
- Train classifier on more data
- Expand to different languages (Spanish, Japanese, Hebrew)

Acknowledgements

I would like to thank Professor Kenji Sagae for introducing me to this project, and for providing his guidance, advice, support, and shared knowledge along the way. I thank Evan Suma for organizing my REU, and NSF for generously providing the funding for it. I would also thank to thank everyone at ICT for treating me as part of the family.