# Using entropy to evaluate child language performance

Zhong Chen, Marisa Ferrara Boston, and John T. Hale
Department of Linguistics, Cornell University, USA

Cornell University
Department of Linguistics

## Introduction

- Examine dependency uncertainty of utterances as an index of child language development in production.
- The uncertainty is measured by Entropy (Shannon, 1948).
- The entropy of the unlabeled, directed dependency relation (Tesnière, 1959) given the head, naturally quantifies the information content of a child's grammar.

## We use CHILDES annotated by Dependency Rules...

- Longitudinal data from 139 transcripts from the CHILDES informant Sarah (Brown, 1973; MacWhinney, 2000)
- Transcripts have been pre-annotated with part of speech (POS) tags and dependency grammatical relations (Sagae et al., 2007).
- Dependency rules are extracted that describe possible dependency relations.
- Dependency rules are based on word types (POS), not word tokens.
- Dependency rules represent linear relations and consequently do not include phrase structures.

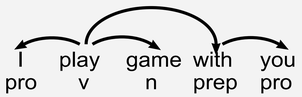



Figure 1. Dependency Rule Relations

$$\Downarrow$$

| | | |
|---|---|---|
| $pro \leftarrow v$ | | "a VERB may have a PROnoun to its left" |
| $v \rightarrow n$ | $\Longrightarrow$ | "a VERB may have a Noun to its right" |
| $v \rightarrow prep$ | | "a VERB may have a PREPosition to its right" |
| $prep \rightarrow pro$ | | "a PREPosition may have a PROnoun to its right" |

- Figure 1 shows a sentence with four directed dependency rules, where the arrow indicates a dependency relation from a governor to a dependent.

## ...to calculate Grammar Entropy

- Our motivation of choosing Grammar Entropy comes from the intuition that a child uses the grammar in its current state, when he or she has to determine the grammatical relations.
- Deciding which unlabelled grammatical relations emanate from a particular word is at least as hard as picking a directed dependency rule for that word's POS.
- Grammar Entropy $H(G)$: the sum of entropies (uncertainty) of deciding an appropriate dependency rule for each governor in the grammar.
- In the grammar $G$ with $n$ governors (represented by word types (POS)), $H(G)$ is calculated as a collection of an individual governor $i$'s entropy. (Cover and Thomas, 2006)
- The entropy for the governor $i$, which has $m$ dependency rules, consists of all conditional entropies of a dependency rule $j$ given governor $i$.

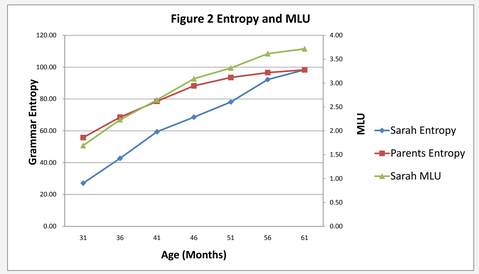
$$H(G) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(Rule_j|Governor_i) \times \log\left(P(Rule_j|Governor_i)\right)$$

## Data preparation

- We split the 139 files Sarah corpus (Age: 2;3 - 5;1) into seven consecutive parts. Each part represents a five-month time period. We remove any annotation or comments. Parental and Sarah's utterances are kept, although their grammar entropies are calculated separately. We also discard the final punctuation.
- We assume the acquisition of dependency rules is cumulative. Thus, for example, dependency rules acquired before 36 months are collected from a combination of two parts of data: "27-31 months" and "32-36 months".

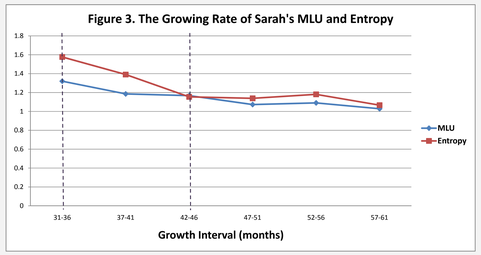
## Both Grammar Entropy and MLU increase.

- Figure 2 plots the longitudinal change of the grammar entropy at each of seven time points of Sarah compared with the adult speech provided by her parents.
- Language complexities for both groups increases as Sarah's language development unfolds more quickly. Thus, by 61 months, the entropies for both the child speech and the adult speech are almost equal in quantity, which means that their grammar complexities reach the same level.
- Figure 2 also illustrates an increase on Sarah's average mean length of utterance (MLU; Brown, 1973) for each period of time throughout the complete time span.



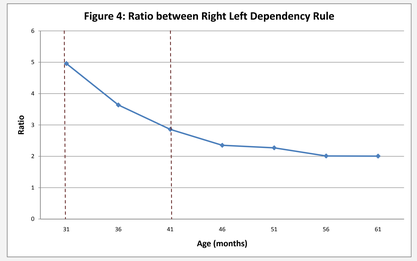

## However, Grammar Entropy and MLU are not the same...

- Figure 3 examines the growing rates compared between Sarah's grammar entropy and her MLU. As in Figure 2, the growth slows in later stages.
- The grammar in Sarah's speeches is diversifying into new dependency structures (Entropy) faster than her utterances are lengthening (MLU).
- The result indicates that grammatical relations might become much more diversified without a significant increase in MLU, especially in the early stage of children.




## A story behind this disparity...

- Figure 4 plots the decreasing ratio between the number of right and left dependency rules. We also notice that this ratio reduces 73.4% between 31 months and 41 months.
- This increasing use of left-attachment suggests that the decision of choosing a dependency rule's direction is more balanced, particularly in the early stages.




- It hence comfirms figure 3 that the grammar is diversifying even faster than the sentences are increasing in length.

## Discussion

- The dependency relations we compute here are based on the grammar represented by word types, not particular word tokens.
- We are investigating the "flexibility" of the grammar if the grammar were to just cover the sentences that have actually been uttered by a child.
- The increase in grammar entropy thus acknowledges children's grammatical structural enrichment.

## Conclusion

- The grammar entropy is able to reflect the growth of children's language performance and can be used as an alternate to MLU.
- The grammar complexity of children's production grows faster than the parental speech, which decreases the gap between the two.
- The grammar entropy quantifies the way that a child's grammatical relations diversify more quickly than her utterances lengthen (MLU), especially in the early months, since the "default" right branching analyses give way to certain left-branching analyses.

## References

Brown, R., 1973. A first language: The early stages. Cambridge, MA: Harvard University Press.
Cover, T., Thomas, J., 2006. Elements of Information Theory. 2nd Edition, Wiley, New York
MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd Edition, New Jersy: Lawrence Erlbaum
Sagae, K., Davis, E., Lavie, A., MacWhinney, B. and Wintner S., 2007. High-accuracy annotation and parsing of CHILDES transcripts. In Proceedings of workshop on cognitive aspects of computational language acquisition, 25-32.
Shannon, C. E., 1948. A math. theory of communication. Bell Sys. Tech. Journal, 27, 379-423, 623-656.
Tesnière, L., 1959. Élements de syntaxe structurale. Editions Klincksiek.

{zc77,mfb74,jthale}@cornell.edu
http://www.people.cornell.edu/pages/zc77/