Forming Syntactic Categories From String-Distributional Data

Adam Leif
UCLA
Contact: adamleif@ucla.edu

How close can we get to adult-like syntactic categories using only string-distributional data?

Introduction

- Category: all elements that occupy shared positions in a distribution [1]
- Frames [2] form precise proto-categories: **left __ right**
- Guiding Questions:
  - How do children form categories? How much can we depend on string-distributional data?
  - How can we redeploy frame-based knowledge? [3]
  - Can frames (good for lexical categorization) be used for function word categorization?

Method

- Granular initial category system: categories = word types
- Form frames from categories, not words themselves
- When categories share ≥ 20% of types [2], merge them
- Using new category system, label corpus
- Repeat process

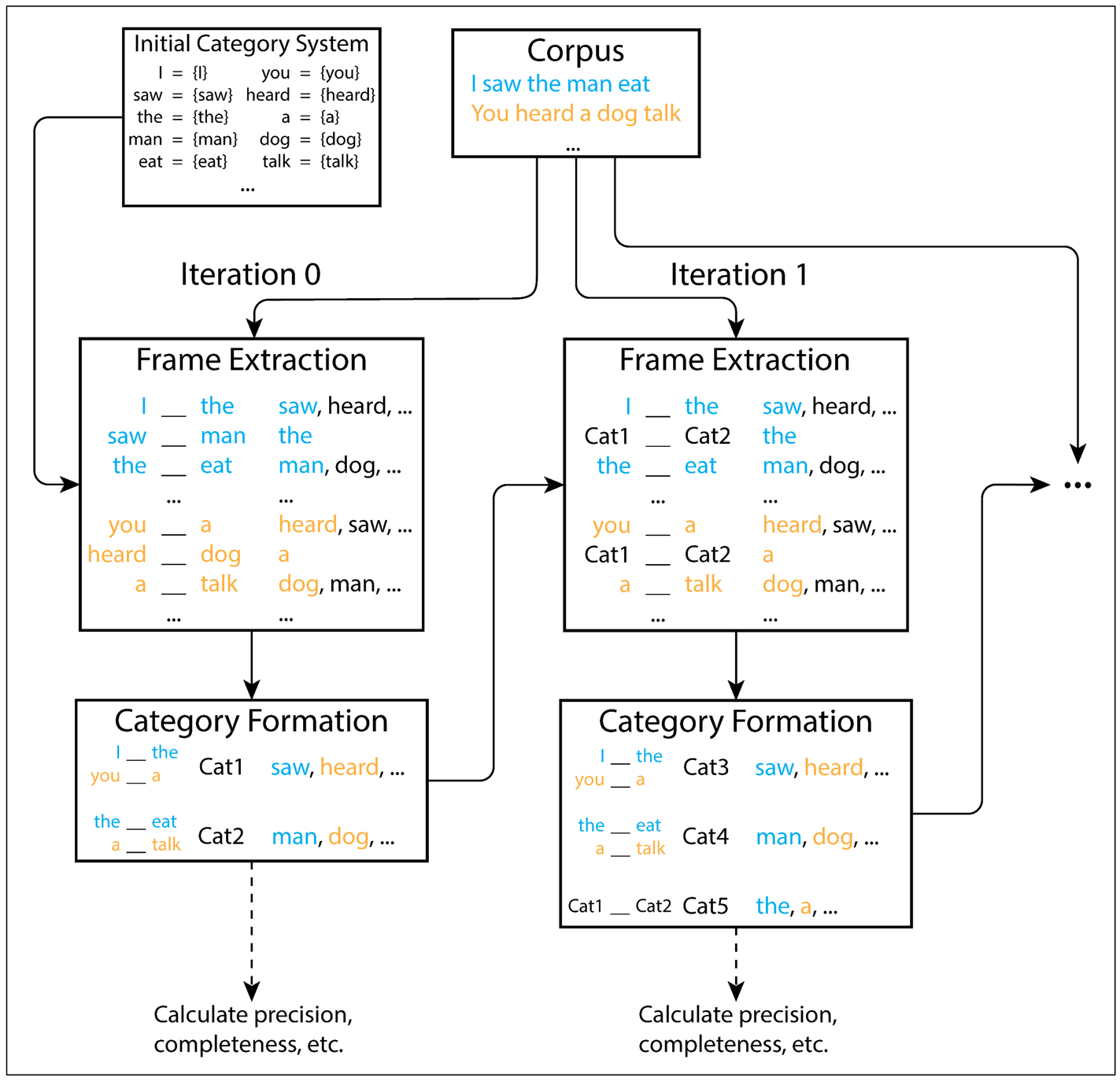


Figure 1

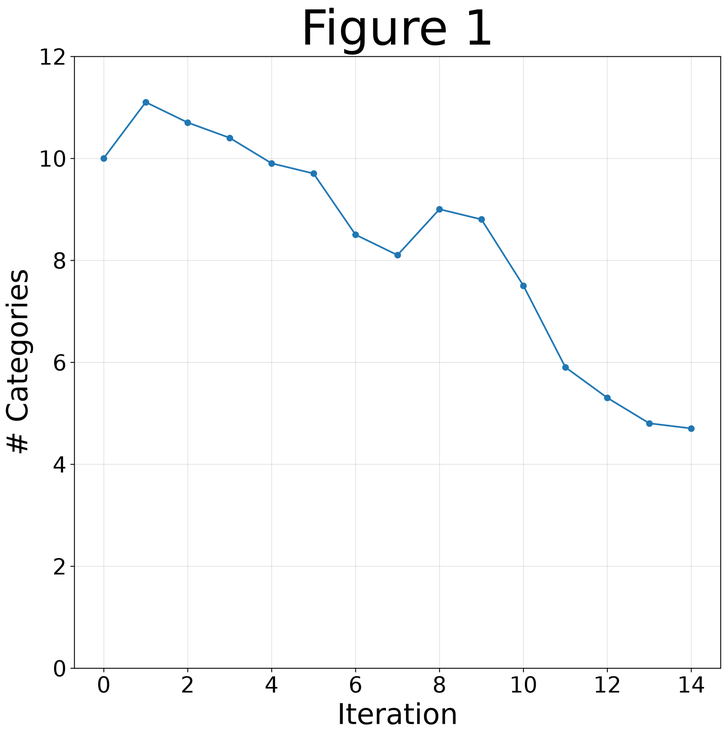


Figure 2

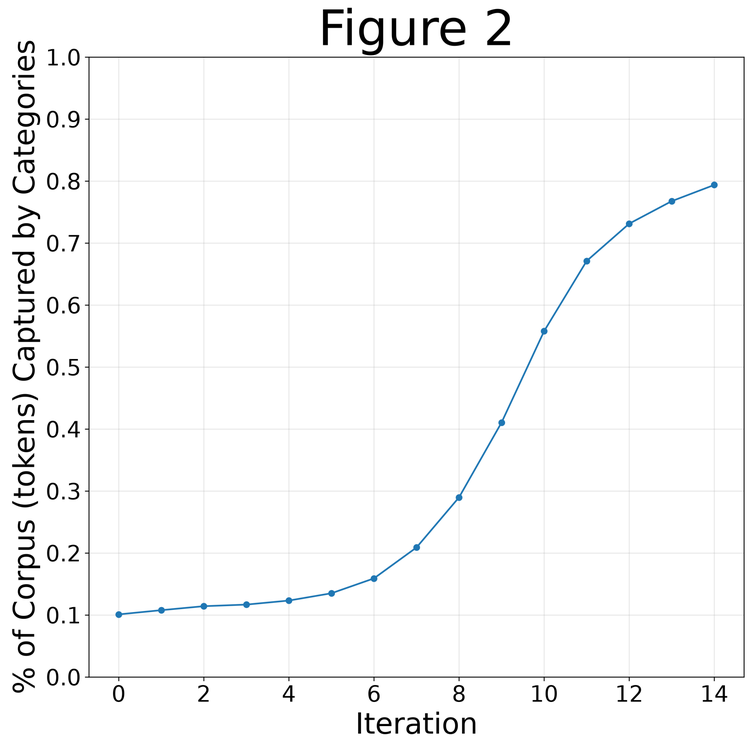


Figure 3

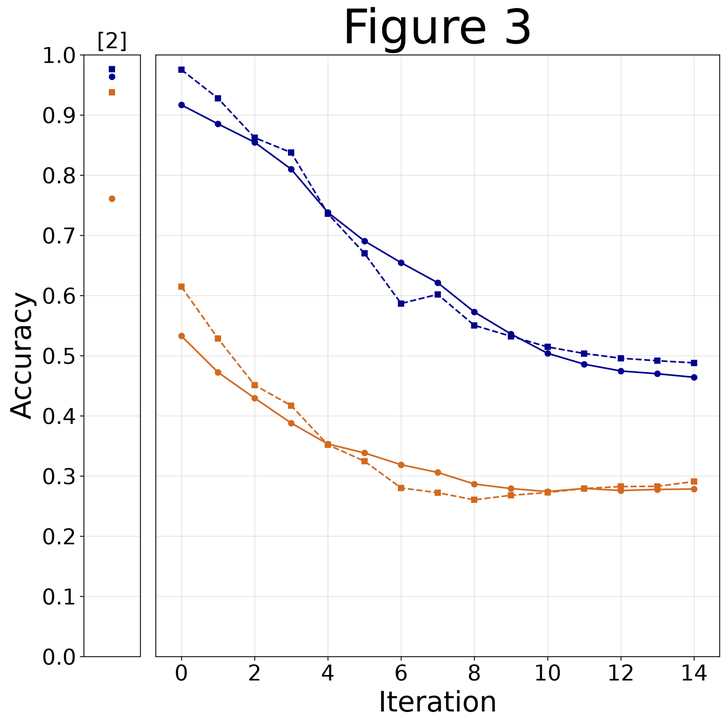
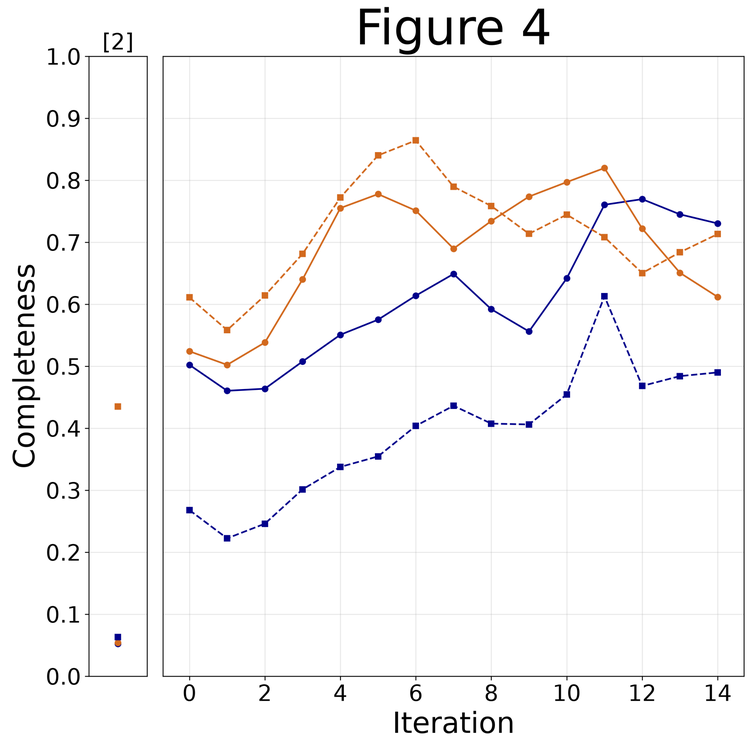


Figure 4





Results

- Performance as learner: # of categories decrease (Fig. 1), but cover more of the corpus (Fig. 2)
- Comparison to CHILDES POS: categories are less precise (Fig. 3), but better capture words with shared gold labels (Fig. 4)
  - Precision: $\frac{hits}{hits + false\ alarms}$ (categories ≈ gold labels?)
  - Completeness: $\frac{hits}{hits + misses}$ (gold's $C(w_i) = C(w_j)$?)
- This process improves categories by some metrics

Discussion

- Manage to better-categorize functional words
- How many categories *should* a young learner have?
  - Unknown and unclear
  - What is the best precision for *n* categories?
- Other free parameter settings (overlap %, labeling function, etc.)
  - No change per iteration
  - Decrease precision, increase completeness
  - # categories is critical for precision + completeness
- Possibility: integrate this into richer category-formation pipeline (e.g. morphemic decomposition)

[1] Harris, Z. S. (1954). *Distributional Structure*. WORD, 10(2–3). [2] Mintz, T. H. (2003). *Frequent frames as a cue for grammatical categories in child directed speech*. Cognition, 90(1). [3] Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). *Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquisition strategies*. Developmental Science, 12(3).

This work was funded by UCLA's Graduate Summer Research Mentorship Program. Special thanks to Laurel Perkins, Tim Hunter, and UCLA's CompPsych Seminar for their invaluable feedback!