The Importance of Category Labels in Grammar Induction with Child-directed Utterances

Lifeng Jin and **William Schuler**
Department of Linguistics
The Ohio State University, Columbus, OH, USA
{jin, schuler}@ling.osu.edu

Abstract

Recent progress in grammar induction has shown that grammar induction is possible without explicit assumptions of language-specific knowledge. However, evaluation of induced grammars usually has ignored phrasal labels, an essential part of a grammar. Experiments in this work using a labeled evaluation metric, RH, show that linguistically motivated predictions about grammar sparsity and use of categories can only be revealed through labeled evaluation. Furthermore, depth-bounding as an implementation of human memory constraints in grammar inducers is still effective with labeled evaluation on multilingual transcribed child-directed utterances.

1 Introduction

Recent work in probabilistic context-free grammar (PCFG) induction has shown that it is possible to learn accurate grammars from raw text (Jin et al., 2018b, 2019; Kim et al., 2019), which is significant in addressing the issue of the *poverty of the stimulus* (Chomsky, 1965, 1980) in linguistics. Although phrasal categories and morphosyntactic features can be induced from raw text (Jin and Schuler, 2019; Jin et al., 2019), most unsupervised parsing work has been evaluated using unlabeled parsing accuracy scores (Seginer, 2007; Ponvert et al., 2011; Jin et al., 2018b; Shen et al., 2018, 2019; Shi et al., 2019). This is potentially distortative because children and adults can distinguish categories of phrases and clauses (Tomasello and Olguin, 1993; Valian, 1986; Kemp et al., 2005; Pine et al., 2013), and much of acquisition modeling research has been directed at simulating the development of abstract linguistic categories in first language acquisition (Bannard et al., 2009; Perfors et al., 2011; Kwiatkowski et al., 2012; Abend et al., 2017; Jin et al., 2018b).

Recent work proposed a labeled parsing accuracy evaluation metric called Recall-V-Measure (RVM) as a method for evaluating unsupervised grammar inducers (Jin et al., 2019), but this metric counts categories as incorrect if they are finer-grained than reference categories or if they represent binarizations of n-ary branches in reference trees, which may be linguistically acceptable. We therefore further modify it to Recall-Homogeneity (RH) calculated as the homogeneity (Rosenberg and Hirschberg, 2007) of the labels of matching constituents of the induced and gold trees, weighted by unlabeled recall. This work uses transcribed child-directed utterances from multiple languages as input to a grammar inducer with hyperparameters tuned using either unlabeled F1 or labeled RH. Results show that: (1) the induced grammars capture the preference of sparse concentrations in human grammars only when using labeled evaluation; (2) grammar accuracy increases as the number of labels grows only when using labeled evaluation; (3) depth-bounding (Jin et al., 2018a, limiting center embedding) is still effective when tuned to maximize labeled parsing accuracy.

2 Model

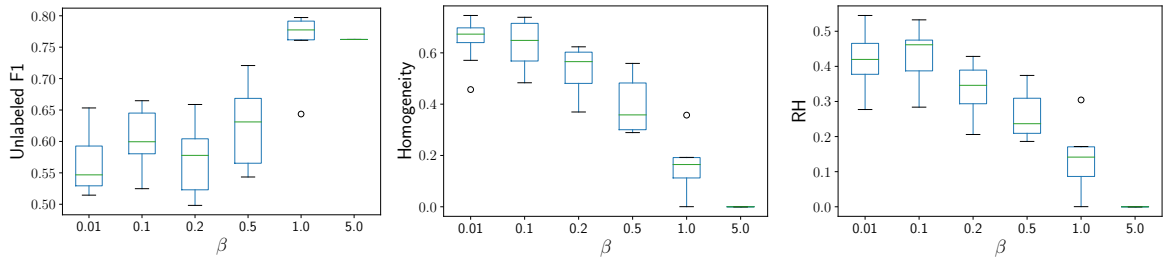
All experiments described in this paper use a Bayesian Dirichlet-multinomial model (Jin et al., 2018a) to induce PCFGs without assuming any language specific knowledge. This model defines a Chomsky normal form (CNF) PCFG with C nonterminal categories as a matrix $\mathbf{G}$ of binary rule probabilities which is first drawn from the Dirichlet prior with a concentration parameter β:

$$\mathbf{G} \sim \text{Dirichlet}(\beta) \quad (1)$$

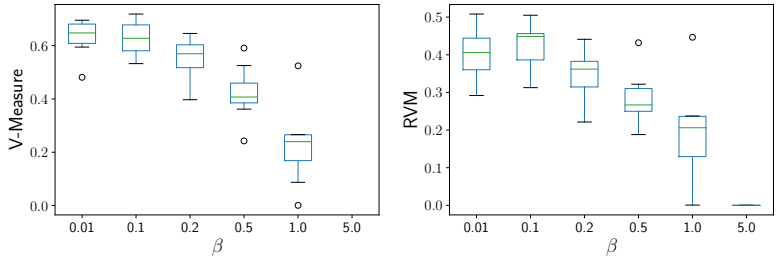
Trees for sentences $1..N$ in a corpus are then drawn from a PCFG parameterized by $\mathbf{G}$:

$$\tau_{1..N} \sim \text{PCFG}(\mathbf{G}), \quad (2)$$

Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task, pages 145–150
Virtual Meeting, July 9, 2020.



(a) Unlabeled F1 scores with βs (b) Homogeneity scores with βs (c) RH scores with βs



(d) V-measure scores with βs (e) RVM scores with βs

Figure 1: Different evaluation metrics on the Adam dataset with different β values.

and each tree τ is a set $\{\tau_\epsilon, \tau_1, \tau_2, \tau_{11}, \tau_{12}, \tau_{21}, ...\}$ of category node labels τ_η where $\eta \in \{1, 2\}^*$ defines a path of left or right branches from the root to that node. Category labels for every pair of left and right children $\tau_{\eta 1}, \tau_{\eta 2}$ are drawn from a multinomial distribution defined by the grammar $\mathbf{G}$ and the category of the parent τ_η:

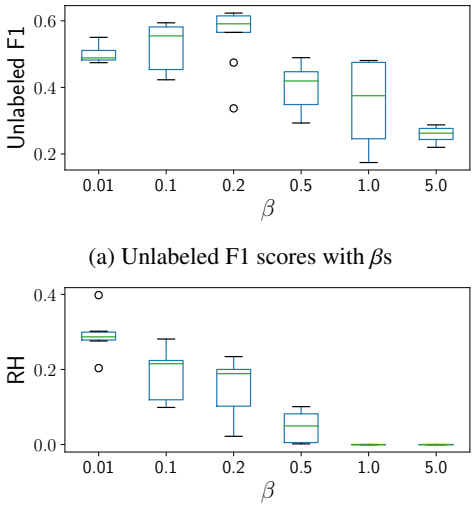
$$\tau_{\eta 1}, \tau_{\eta 2} \sim \text{Multinomial}({\delta_{\tau_\eta}}^\top \mathbf{G}) \quad (3)$$

where δ_x is a Kronecker delta function equal to 1 at value x and 0 elsewhere. Terminal expansions are treated as expanding into a terminal node followed by a special null node.

Inference in this model uses Gibbs sampling to produce samples of grammars and trees with the most probable parses obtained with the Viterbi algorithm.

3 Data and hyperparameters

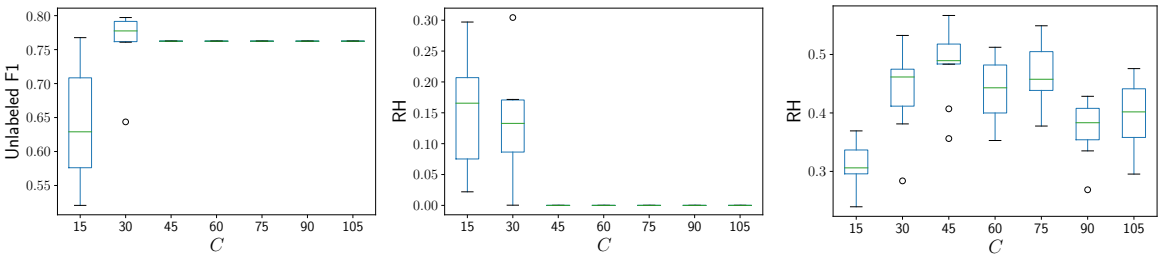
Experiments here use transcribed child-directed utterances from the CHILDES corpus (Macwhinney, 1992) in three languages with more than 15,000 sentences each. English hand-annotated constituency trees are taken from the Adam and Eve portions of the Brown Corpus (Brown, 1973). Mandarin (Tong, Deng et al., 2018) and German (Leo, Behrens, 2006) data are collected from CHILDES with reference trees automatically generated using the state-of-the-art Kitaev and Klein (2018) parser. Disfluencies are removed, and only sentences spoken by caregivers are kept in the data. Models are run 10 times with 700 iterations with random seeds following previous work (Jin et al., 2018a). The last sampled grammar is used to generate Viterbi parses for all sentences. All punctuation is retained during induction and then removed in evaluation. Significance testing uses permutation tests on concatenations of Viterbi trees from all test runs. We use Adam for exploratory experiments and the other three sets for confirmatory experiments.



(a) Unlabeled F1 scores with βs

(b) RH scores with βs

Figure 2: Different evaluation metrics on the WSJ20Dev dataset with different β values.



(a) Unlabeled F1 with Cs and $\beta = 1$ (b) RH scores with Cs and $\beta = 1$ (c) RH scores with Cs and $\beta = 0.01$

Figure 3: Different evaluation metrics on the Adam dataset with different C values at high and low βs.

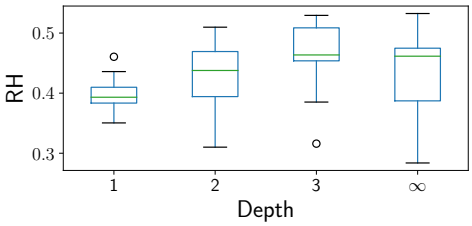


Figure 4: Depth-bounding on Adam

3.1 Recall-Homogeneity

RH is calculated by multiplying unlabeled recall of bracketed spans in the predicted Viterbi trees with the homogeneity score (Rosenberg and Hirschberg, 2007) of the predicted labels of the matching spans, This is different from RVM (Jin et al., 2019), which is the product of unlabeled recall and V-measure. The metric is insensitive to the branching factor of the grammar by the use of unlabeled recall. Unlike RVM, it is also insensitive to the precision of predicted labels to gold labels, indicating that models are not penalized by hypothesizing more refined categories, as long as these categories all fall into the confines of a gold category. RVM, on the other hand, would penalize both underproposing and overproposing categories compared to the ones in the annotation, but the gold categories, like nouns and verbs, are defined on a very high level that languages almost always further specify, represented usually as subcategories or features in linguistic theories. Unary branches in gold and predicted trees are removed, and the top category is used as the category for the constituent.

4 Experiments

4.1 Experiment 1: Labeled evaluation shows preference of grammar sparsity

Human grammars are sparse (Johnson et al., 2007; Goldwater and Griffiths, 2007). For example, in the Penn Treebank (Marcus et al., 1993), there are 73 unique nonterminal categories. In theory, there can be more than 28 million possible unary, binary and ternary branching rules in the grammar. However, only 17,020 unique rules are found in the corpus, showing the high sparsity of attested rules. In other frameworks like Combinatory Categorial Grammar (Steedman, 2002) where lexical categories can be in the thousands, the number of attested lexical categories is still small compared to all possible ones.

The Dirichlet concentration hyperparameter β in the model controls the probability of a sampled multinomial distribution concentrating its probability mass on only a few items. Previous work using similar models usually sets this value low (Johnson et al., 2007; Goldwater and Griffiths, 2007; Graça et al., 2009; Jin et al., 2018b) to prefer sparse grammars (i.e. grammars in which most of the probability mass is allocated to a small number of rules), with good results. The prediction based on the preference of sparsity is that the best β value should be much lower than 1.

Figure 1a shows unlabeled F1 scores with different β values on Adam.[1] Contrary to the prediction, grammar accuracy peaks at high values for β when measured using unlabeled F1. However, these grammars with high unlabeled F1 are almost purely right-branching grammars, which performs very well on English child-directed speech in unlabeled parsing evaluation, but the right-branching grammars have phrasal labels that do not correlate with human annotation when evaluated with Homogeneity, shown in Figure 1b. This indicates that instead of capturing human intuitions about syntactic structure, such grammars have only captured broad branching tendencies. The same grammars are evaluated again with RH, shown in Figure 1c.

[1]The results shown in the figure use C=30. We also tested other C values from 15 to 105 and the trend is almost identical.

When both structural and labeling accuracy is taken into account, results correctly capture the intuition that grammar accuracy has a low peaking concentration hyperparameter. Figure 1d and 1e shows the same experiments evaluated with the labeled evaluation metric RVM. Because of the sensitivity to labeling accuracy, results in VM and RVM also show the similar trend as Homogeneity and RH where labeling quality decreases as β increases. Jin et al. (2018b) noted that induced grammars high in unlabeled bracketing scores are low in NP discovery scores, which is a category-specific evaluation metric. This can also be explained by the induced grammars with high bracketing scores only capture a broad right-branching bias without accurately clustering words and phrases based on their distributional properties.

Figure 2 shows the same experiments on a corpus of formal English written text, the WSJ20dev[2] dataset. The pattern is similar but less extreme than on CHILDES. The higher βs at the range of 0.1-0.2 still show better performance on unlabeled F1 than the sparser models, consistent with previous results in Jin et al. (2018b). However RH scores reveal that the labels induced by the denser models are less accurate, manifesting as the overall lower peak for β using RH than using unlabeled F1.

4.2 Experiment 2: Performance increases with the number of categories

Previous research (Jin et al., 2018a) also reported that the number of categories C used by the induction models was relatively low compared to the number of categories in human annotation. For example, there are 63 unique tags in the Adam dataset. This is in contrast to 30 or fewer categories used in previous induction work. The bias brought by high β values and unlabeled evaluation together may be masking the real relationship between the number of categories and grammar accuracy.

Figures 3a and 3b show unlabeled and labeled evaluation on different grammars induced with the best performing β on Adam tuned by unlabeled F1. With F1, increasing the number of categories beyond 30 yields no improvement as most of the induced grammars are purely right-branching grammars. RH results confirm this: as grammars approach the pure right-branching solution when C increases, the similarity between induced and gold labels of constituents deteriorates quickly. RH scores from grammars induced with $\beta = 0.01$ are more indicative of the interaction between the number of categories and grammar accuracy. Grammar accuracy increases as C gets larger initially and peaks at $C = 75$. The results confirm the importance of labeled evaluation, because the trend from labeled evaluation shows that there should be a sufficient number of categories to account for different syntactic structures, and models with small numbers of categories are limited in their ability to do this.

4.3 Experiment 3: Depth-bounding is still effective with RH

Previous work showed that depth-bounding is effective in helping grammar inducers induce more accurate grammars (Shain et al., 2016; Jin et al., 2018a), because it removes the parse trees with deeply nested center-embeddings, which cannot be produced by humans due to memory constraints (Chomsky and Miller, 1963), from grammar induction inference. However the unlabeled evaluation metric used in previous work may lead to unhelpful conclusions. In order to revisit this claim with labeled evaluation, experiments are first conducted on Adam exploring the interaction between depth and labeled performance, and subsequently on the Eve (English), Tong (Chinese Mandarin) and Leo (German) portions of the CHILDES corpus. All experiments use hyperparameters tuned with RH.[3]

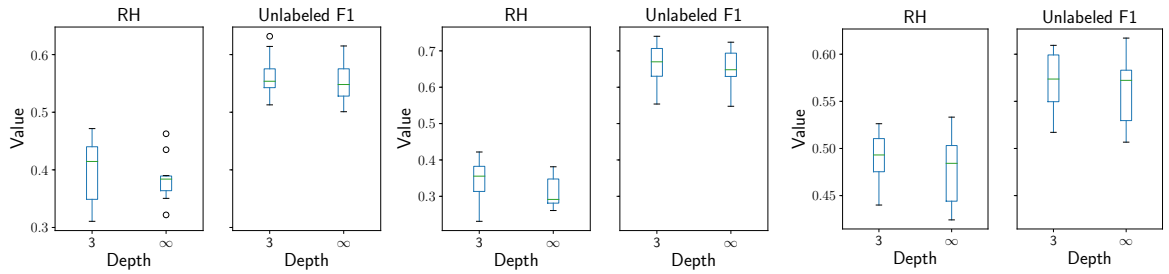
Figure 4 shows the interaction between depth and RH scores on Adam. Performance of the unbounded models can be lower than all bounded models, showing that unbounded inducers can induce grammars inconsistent with human memory constraints. The labeled performance peaks at depth 3, which is significantly more accurate ($p < 1 \times 10^{-3}$) than unbounded models. This is consistent with previous results that over 97% of trees in English contain 3 or fewer nested center embeddings (Schuler et al., 2010).

Experiments on Eve, Tong and Leo replicate this result. Figure 5 shows that the models bounded at depth 3 are more accurate than unbounded models with both unlabeled and labeled evaluation metrics. Significance testing with unlabeled F1[4] shows the

[2]The first half of the Wall Street Journal part of the Penn Treebank with sentences with 20 words or fewer.

[3]The optimal C is 75 from previous experiments, but we used 30 in all depth-bounding experiments due to hardware constraints at high depth bounds.

[4]Neither RH nor RVM were used in permutation significance testing, because labels with the same values from different induced grammars may represent different linguistic categories, therefore two parses of the same sentence from



(a) Depth-bounding on Eve (b) Depth-bounding on Tong (c) Depth-bounding on Leo

Figure 5: Comparison of labeled and unlabeled evaluation of grammars bounded at depth 3 and unbounded grammars on English (Eve), Chinese Mandarin (Tong) and German (Leo) datasets from CHILDES.

performance differences across three datasets are all highly significant ($p < 0.001$). Therefore, the claim that depth-bounding is effective in grammar induction is still supported when the models are developed and evaluated with labeled evaluation.

5 Conclusion

Unlabeled evaluation has been used in grammar induction, but experiments presented in this paper show that unlabeled evaluation can reveal unexpected bias in the data which may lead to unhelpful conclusions compared to labeled evaluation. Results show that trends of preference of sparsity and use of categories that are consistent with linguistic annotation can only be discovered with labeled evaluation. Furthermore, human memory constraints are still effective in grammar induction when labeled evaluation is used throughout all stages of development.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. In *Cognition*, volume 164, pages 116–143. Elsevier B.V.

Colin Bannard, Elena Lieven, and Michael Tomasello. 2009. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17284–9.

Heike Behrens. 2006. The input–output relationship in first language acquisition. *Language and Cognitive Processes*, 21(1-3):2–24.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

different runs are not exchangeable.

Noam Chomsky. 1980. On cognitive structures and their development: A reply to Piaget. In Massimo Piattelli-Palmarini, editor, *Language and learning: the debate between Jean Piaget and Noam Chomsky*, chapter 49, pages 751–755. Harvard University Press.

Noam Chomsky and George A Miller. 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology*, pages 269–321. Wiley, New York, NY.

Xiangjun Deng, Virginia Yip, Brian Macwhinney, Stephen Matthews, Mai Ziyin, Zhong Jing, and Hannah Lam. 2018. A Multimedia Corpus of Child Mandarin: The Tong Corpus. *The Journal of Chinese Linguisticsvol*, 46(1):69–92.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751.

João V. Graça, Kuzman Ganchev, Taskar Ben, and Fernando Pereira. 2009. Posterior vs. Parameter sparsity in latent variable models. In *Advances in Neural Information Processing Systems*, pages 664–672.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019. Unsupervised Learning of PCFGs with Normalizing Flow. In *ACL*.

Lifeng Jin, Finale Doshi-Velez, Timothy A Miller, William Schuler, and Lane Schwartz. 2018a. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lifeng Jin, Finale Doshi-Velez, Timothy A Miller, William Schuler, and Lane Schwartz. 2018b. Unsupervised Grammar Induction with Depth-bounded PCFG. *Transactions of the Association for Computational Linguistics*.

Lifeng Jin and William Schuler. 2019. Variance of average surprisal: a better predictor for quality of grammar from unsupervised PCFG induction. In *ACL*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov chain Monte Carlo. *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–146.

Nenaugh Kemp, Elena Lieven, and Michael Tomasello. 2005. Young Children's Knowledge of the "Determiner" and "Adjective" Categories. *Journal of Speech, Language, and Hearing Research*, 48(June):592–609.

Yoon Kim, Chris Dyer, and Alexander M Rush. 2019. Compound Probabilistic Context-Free Grammars for Grammar Induction. In *ACL*.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In *ACL*.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244.

Brian Macwhinney. 1992. *The CHILDES Project: Tools for Analyzing Talk*, third edition. Lawrence Elrbaum Associates, Mahwah, NJ.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118:306–338.

Julian M. Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. 2013. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage parsing using human-Like memory constraints. *Computational Linguistics*, 36(1):1–30.

Yoav Seginer. 2007. Fast Unsupervised Incremental Parsing. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

Cory Shain, William Bryce, Lifeng Jin, Victoria Krakovna, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2016. Memory-bounded left-corner unsupervised grammar induction on child-directed input. In *Proceedings of the International Conference on Computational Linguistics*, pages 964–975.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In *ICLR*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks. In *ICLR*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually Grounded Neural Syntax Acquisition. In *ACL*.

Mark Steedman. 2002. Formalizing Affordance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Michael Tomasello and Raquel Olguin. 1993. Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8(4):451–464.

Virginia Valian. 1986. Syntactic Categories in the Speech of Young Children. *Developmental Psychology*, 22(4):562–579.