# Modeling Child Syntactic Category Learning: Applying the Tolerance Principle to Prototype-Driven Part-of-Speech Tagging*

**Diana Marsala**
University of Pennsylvania
marsalad@sas.upenn.edu

## 1 Introduction

State-of-the-art part-of-speech (POS) tagging models are highly accurate (FLAIR from Akbik et al., 2019 is 97.85% accurate), but cognitively implausible because they rely on large annotated corpora and/or are structured as complex statistical optimization problems (Haghighi & Klein, 2006; Christodoulopoulos et al., 2010). Meanwhile, POS tagging is a problem children are extremely skilled at solving when in the process of learning a language (Tomasello, 2000). The average American child hears only about 5 million words per year (Hart & Risley, 1995), none of which are labeled with their syntactic category like they are in supervised models such as FLAIR. Children also receive no negative feedback (Brown & Hanlon, 1970), and surely have lesser computational capacities than that of modern computing systems.

In this paper, we develop a model that can be used for both child language acquisition and minimally-supervised POS tagging. The key principles it is based upon are semantic bootstrapping, distributional frames, reinforcement learning, and the Tolerance Principle–a threshold for the productivity of linguistic rules (Yang, 2016). While our goal is not to outperform existing POS taggers, we present accuracy, precision, recall, and F-score metrics to demonstrate the our model's validity.

This paper is organized as follows. Section 2 contains background on the principles our model is based upon. Section 3 details our prototype-driven model. Evaluation methods are explained in section 4, followed by results in section 5. Section 6 discusses conclusions and areas of potential future research.

*Thank you to Dr. Charles Yang for advising my work on this project, Dr. Julie Anne Legate & Ryan Budnick for teaching the class this project was originally developed for, John Hewitt for his insights and advice, and my LING 300 classmates for their invaluable feedback.

## 2 Background

While it is still unknown exactly how children learn syntactic categories, we focus on four key theories of child language acquisition when developing our model.

### 2.1 Semantic Bootstrapping

The semantic bootstrapping hypothesis (Pinker, 1984) suggests that children innately map salient semantic categories to syntactic ones. For example, children learn that tangible objects are nouns, actions are verbs, visible properties are adjectives, etc. An analysis by Rondal et al. (1987) supports this hypothesis by showing, in an English corpus of child-directed speech, all things are nouns and all actions are verbs. Further, we know children's early vocabulary tends to contain mostly tangible, concrete, salient words (Pinker, 1984; Gleitman et al., 2005, Carlson et al., 2014). Presumably learning these salient words enables children to use other tools such as distributional information to expand their vocabulary further.

### 2.2 Distributional Frames

There is wide support that distributional cues play a role in learning the syntactic categories of novel words (Brown, 1957, Mintz, 2003, Chemla et al., 2009, Reeder et al., 2013, Schuler et al., 2017). For example, upon hearing the phrase "This is a sib", four-year-olds will assume "sib" is a noun, but upon hearing the phrase "I sib the dog" they will assume "sib" is a verb (Brown, 1957). Adults have also been shown to make these assumptions when learning artificial languages (Schuler et al., 2017). Distributional frames, defined by Mintz (2003) to be the words immediately preceding and immediately following the target word, are at play here. For example, in "I sib the dog", the lexical frame surrounding "sib" is "I _ the" and the POS

frame is "PRO _ D"[1]. A language learner can use these context words to infer the POS tag of the target word.

### 2.3 Reinforcement Learning

Children have been shown to overgeneralize linguistic rules, over time learning exceptions and refining their assumptions (Marcus et al., 1992; Yang, 2016). For example, the English rule in which past-tense forms of verbs have the suffix "-ed" is initially overgeneralized in young children: they will incorrectly say "I goed to the park" instead of "I went to the park". The ability to unlearn (or learn exceptions to) these overgeneralized rules is critical to child language learning.

Reinforcement learning is an artificial intelligence concept concerning how an agent learns and acts. As more information is taken in by the agent, the agent's assumptions about the world change. This process of taking in more information over time and updating assumptions (both to learn rules and unlearn rules) is common to both reinforcement learning and child language acquisition.

### 2.4 The Tolerance Principle

The Tolerance Principle, proposed by Yang (2016), defines a threshold for the productivity of linguistic rules. If $n$ is the number of words a rule applies to and $e$ is the number of words which are exceptions to that rule, the rule is only productive if

$$e \leq \theta_n = \frac{n}{\ln(n)}$$

If $e > \theta_n$, the rule is not productive and thus is not learned by a language learner.

Schuler et al. (2016) demonstrated the Tolerance Principle's effectiveness for a variety of rules. For example, the study found 1022 unique past-tense verbs in their child-directed speech data ($n = 1022$), of which 127 did not end in "-ed" ($e = 127$). Here $\theta_n = 147.5$, thus $e \leq \theta_n$ and the Tolerance Principle holds for the rule asserting past-tense verbs end in "-ed". Children will assume that for an unknown verb, adding the suffix "-ed" will make the verb past-tense.

## 3 Model

Our model begins with a small set of POS-tagged seed words. The model learns prototypical frames by leveraging this seed list. Learned frames are then used to do the reverse and learn the POS of the novel words they target.

### 3.1 Seed Words

Based on the semantic bootstrapping hypothesis, we choose a small set of common, salient words to kick-start our model's learning.

The Chicago corpus consists of child-directed speech samples from 64 child-caregiver dyads, observed in the home every 4 months from 14 to 50 months (Rowe & Goldin-Meadow, 2009). Carlson et al. (2014) lists the 536 words that most of these children recognized by 50 months and highlight 86 as especially common and salient. We first manually label these 86 words[2] with their POS. From these 86 words, the 3 most frequent words (in the CHILDES dataset) for each of the 7 POS categories are given as input to our model. These 21 words are the only labeled data the model will receive.

For the Chinese CHILDES dataset, we use the same 7 POS categories but do not restrict seed words to be from a predefined list of 86 salient words; We simply find the 3 most frequent words in the dataset for each tag and use those as the seeds, consistent with the methodology used by Haghighi & Klein (2006).

### 3.2 Iterative Learning

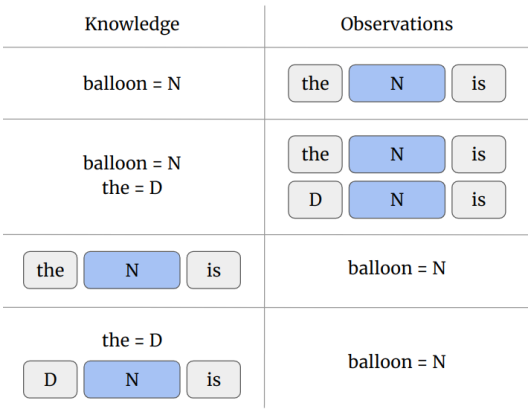
Words are provided 1-by-1 to the model. Consistent with the principle of distributional frames proposed by Mintz (2003), the frame surrounding each word is also provided (i.e. the word immediately preceding and the word immediately following the target word). As the model learns possible frames and POS tags for words, it uses these prototypes to dynamically learn new ones. For example, assume "the balloon is" is fed to the model. Here, "balloon" is the target word and "the _ is" is the context. Figure 1 shows an example of observations the model makes for this input based on the model's current knowledge. The learning process works in both directions: words can be leveraged to learn frames and frames can be leveraged to learn words.

As the model is fed more data, it will keep track of all frames for "balloon", e.g. "the N is" and "big N on" as it observes them. Since "balloon" is a noun, these are potential frames for N. Initially the frames are lexical only (Lexical-Lexical), but as the model learns more words, frames may become

[1] PRO = pronoun, D = determiner

[2] The final labeled list contains 32 verbs, 25 nouns, 9 prepositions, 8 adjectives, 5 adverbs, 4 pronouns, and 3 determiners.

Figure 1: Frame Generalization Process

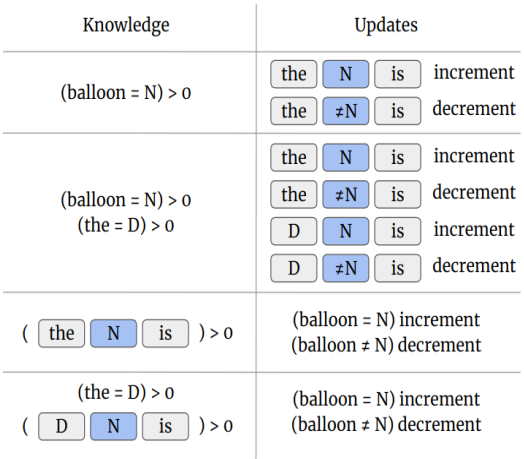



generalized, e.g. "the N V" (Lexical-POS) and "ADJ N P" (POS-POS). If multiple frames are valid when inferring a tag, the more specific one is used, e.g. "the N is" takes precedence over "the N V".

### 3.3 Learning Rules

At each step the model is not just making observations as shown above. It is also keeping track of how many times a frame or POS tag for a word has been valid or invalid. This is the reinforcement learning component of the model. Returning to the example "the balloon is", Figure 2 shows how the scores are updated based on the model's current knowledge. If a POS tag for a word or frame has a positive score, it is considered valid. The tagged seed words are always considered valid.

Figure 2: Reinforcement Learning




Since scores are constantly being incremented and decremented, the model is able to learn and unlearn frames or words. However this simple threshold does not apply when generalizing POS-POS frames.

### 3.4 Generalizing POS-POS Frames

POS-POS frame generalization only occurs if the Tolerance Principle holds for both the left and right sides of a frame. As an example, consider the potential POS-POS frame "D N V".

For the left side of the frame, we find all valid (i.e. positive scores from the reinforcement learning step) Lexical-POS frames where a word the model thinks is a determiner is on the left and V is on the right. The model finds 5 such frames: "the N V", "this N V", "an N V", "a N V", and "that ADV V". Using the equation for the Tolerance Principle (Yang, 2016), we have $n = 5$, $\theta_n = 3.1$, and $e = 1$. $e \leq \theta_n$, thus the left side holds.

We do the same for the right side. Find all valid Lexical-POS frames where a word the model thinks is a verb is on the right and D is on the left. The model finds 4 such frames: "D N is", "D P run", "D ADV see", and "D ADV play". $n = 4$, $\theta_n = 2.9$, and $e = 3$. $e > \theta_n$, thus the right side does not hold. Since the Tolerance Principle holds for only one side of "D N V", the frame is not learned.

### 3.5 POS Ambiguity

Occasionally, the model will learn incorrect frames, such as "ADV D N". The source of these errors is the POS ambiguity of some words. The seed word "down" is likely the source of this specific error since, despite not always being an adverb, it is labeled as an adverb in the seed list. In the phrase "it's just down the road", "down" is not an adverb but rather a preposition. This reveals two limitations of the model:

1. Seed words should be unambiguous.
2. The model can learn at most 1 tag for a word.

The second point is indeed a limitation, but it does not mean the model will only ever tag an ambiguous word with a single tag. If the model fails to learn the tag for an ambiguous word, it will tag each instance of the word based on the frame surrounding the word. If the frame is different for different tags of the word (as we would expect), it can be tagged differently depending on the context.

We experimented with different methods of allowing ambiguity while developing this model. One approach allowed the model to tag a word with a less likely tag when the target word for the relevant frame doesn't match the tag the model already believes the word has. This did not improve accuracy, so we tried extending it by allowing the

model to look back 1 word: to use the tag it just predicted for the left context word (the previous target word) when predicting the current target word.

We also tried multiple adjustments to the thresholds for when ambiguity would be allowed, but ultimately while none of the methods we tried decreased accuracy by more than 5%, they also didn't improve the accuracy or F-score of our model. At least in the case of the CHILDES data, we believe this is because the data itself has relatively low ambiguity and adding this feature unnecessarily increased the complexity of the model.

## 4 Evaluation Methods

For all experiments, we hold out 10% of the data for testing and compare to a baseline model that tags only seed words.

### 4.1 Data

We primarily evaluate our model on child-directed portions of four corpora from the CHILDES database for 7 POS tags[3], totaling just under 2 million words (Brown, 1973; Gelman et al., 1998, Brent & Siskind, 2001; Gelman et al., 2004; Gelman et al., 2014; Newman et al., 2016). Since our model is based on principles of child language acquisition, evaluation on child-directed speech will demonstrate its cognitive plausibility. To demonstrate the our model's efficacy for langauges other than English, we also evaluate on child-directed portions of ten corpora from the Mandarin Chinese section of the CHILDES database, for the same 7 POS tags as English, totaling just under 1 million words (Zhou, 2001; Chang, 2003; Li & Zhou, 2004; Luo et al., 2012; Li & Zhou, 2015; Cheung & Chang, 2017a; Cheung & Chang, 2017b; Deng & Yip, 2018; Zhou, 2018; Li, 2019).

We also evaluate our model on the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1993) for all 45 POS tags in order to compare our model to the prototype-driven model by Haghighi & Klein (2006). The more complicated tagset used for this dataset allows us to demonstrate our model's extensibility. When evaluating on this dataset, we use the same seed list as Haghighi & Klein (2006), which contains 112 words. We present the accuracies on both English and Chinese data.

[3]POS tags: ADJ=adjective, ADV=adverb, D=determiner, N=noun, P=preposition, PRO=pronoun, V=verb

### 4.2 Metrics

We use 1-to-1 token accuracy for most tests in order to evaluate the quality of the tags: count the number of times the model tagged a word correctly and divide that by the total number of words tagged.

For the remaining tests, we evaluate the validity of our model by calculating pairwise metrics that define precision and recall for a clustering task (Christodoulopoulos et al., 2010). Typically for some category $X$, precision is the number of tokens correctly tagged as $X$ divided by the number of tokens tagged as $X$ and recall is the number of tokens correctly tagged as $X$ divided by the number of tokens that actually have tag $X$.

However when there are more than 2 categories, as is the case here, we can define pairwise precision and recall in which each unique pair of words is an instance. If the two words should be in the same category, then the pair's tags are correct if they are given the same tag by the model. If the two words should not be in the same category, then the pair's tags are correct if they are given two different tags by the model. Since these pairs are generated from all unique words instead of from all tokens, a word's tag for pairwise metrics is its most frequent tag.

## 5 Results

A comparison of our model to the baseline and models by Chemla et al. (2009) and Freudenthal et al. (2013) on CHILDES datasets is shown in Figure 3. Evaluation on the WSJ dataset and comparison to models by Haghighi & Klein (2006) is shown in Figure 4.

Figure 3: CHILDES Model Accuracies

| Accuracy | English | Chinese |
|---|---|---|
| Baseline | 14.7% | 24.7% |
| Our Model | **74.8%** | **52.1%** |
| Chemla | 53% | - |
| Freudenthal | 55% | - |

On the CHILDES dataset, our model far outperforms similar models which leverage distributional frames by Chemla et al. (2009) and Freudenthal et al. (2013). The Chinese CHILDES dataset is half the size of its English counterpart, but our model still more than doubles the accuracy of the baseline.

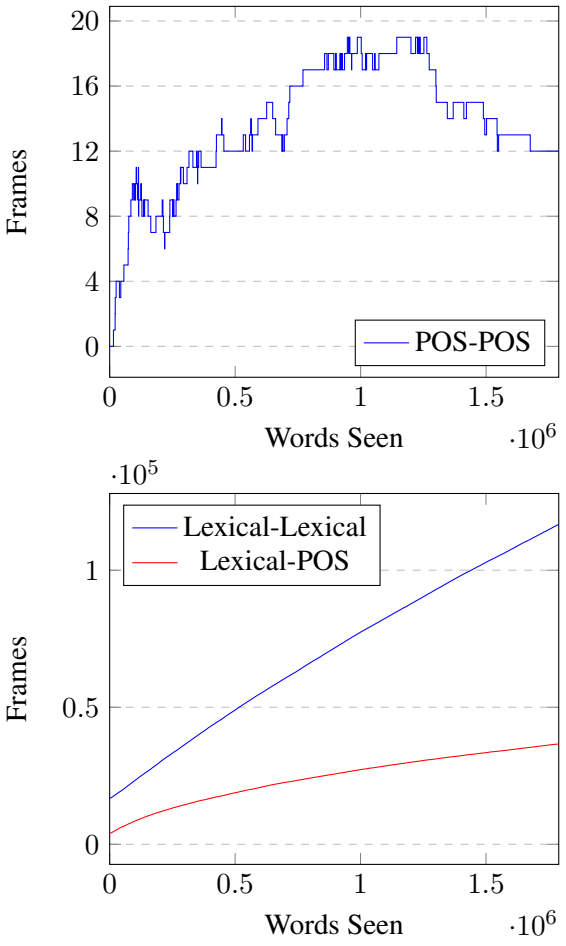
Our model far outperforms the baseline on the WSJ dataset and performs similarly to the PROTO

Figure 4: WSJ Model Accuracies

| Accuracy | English | Chinese |
|---|---|---|
| Baseline | 42.3% | 29.4% |
| Our Model | 64.3% | 56.4% |
| H&K BASE | 41.3% | 39.0% |
| H&K PROTO | **68.8%** | **57.4%** |

model by Haghighi & Klein (2006) on both the English and Chinese WSJ datasets. In fact, our model achieves much higher accuracies than the simpler BASE model by Haghighi & Klein (2006) for both English and Chinese which, unlike the PROTO model, uses morphological features only and does not perform complex statistical optimizations to cluster words when building prototypical frames. While our goal was not necessarily high accuracy, the large gains over the baseline POS tagger and similar performance to the more complicated prototype-driven model by Haghighi & Klein (2006) demonstrates the effectiveness of our approach.

Figure 5: Frame Generalization




The number of frames the model learns over time for CHILDES is shown in Figure 5. The model is indeed learning more frames as it takes in more data. Interestingly, the POS-POS frames are not strictly increasing. Our model is learning and unlearning these frames just like a child would overgeneralize and eventually unlearn the rule or learn exceptions.

Examples of frames that were unlearned include "ADJ V D" and "V D ADJ". It is easy to speculate why these frames would be unlearned: the two context words do not always predict the same target word. "ADJ _ D" could indeed target V, such as in "cats that are white are the best", but a target of P may be more likely, such as in "are you ready for a snack". "V _ ADJ" could target D, as in the case of "hit the red button", but it targets ADJ in "I like tiny pink bows" and P in "you asked for plain pasta". These frames weren't necessarily incorrect, but they weren't good enough to accurately predict the POS of the target word according to the tolerance principle, so they were unlearned by the model.

The full list of frames when the model has run to completion is shown in Figure 6. All frames the model learns are valid with the exception of "ADV D N". As demonstrated in the example for this frame, the ambiguity of the seed word "down" is likely the source of error here.

Figure 6: POS-POS Frames

| | Frame | | | Example |
|---|---|---|---|---|
| | ADV | V | PRO | you can calmly tell her |
| | P | D | N | swimming in the pool |
| * | ADV | D | N | it's just down the road |
| | PRO | V | D | can you taste that |
| | PRO | V | P | he cried in his room |
| | ADV | V | ADV | we do not talk fast |
| | N | V | P | dinner is on the table |
| | ADV | V | D | quietly get a book |
| | P | D | ADJ | look in this blue bag |
| | D | N | P | is the dog on the couch |
| | ADV | V | P | quickly eating at lunch |
| | ADJ | N | P | a red bird on a branch |

Finally, we calculate pairwise metrics for both the English CHILDES dataset as well as the 536-word corpus the English CHILDES seed words are from (Carlson et al., 2014). Evaluation on both is shown in Figure 7.

Comparing the baseline to our model, precision does decrease, but this is expected since a model

Figure 7: Pairwise Metric Evaluation

(a) 10% Held-Out CHILDES

| | P-Precision | P-Recall | F-Score |
|---|---|---|---|
| Baseline | **1.000** | $2.73 \cdot 10^{-6}$ | $5.46 \cdot 10^{-6}$ |
| Our Model | 0.646 | **0.522** | **0.577** |
| Cartwright | 0.853 | 0.178 | 0.295 |
| Mintz | 0.91 | 0.13 | 0.228 |

(b) Carlson Potential Seed List

| | P-Precision | P-Recall | F-Score |
|---|---|---|---|
| Baseline | **1.000** | $5.30 \cdot 10^{-4}$ | $1.06 \cdot 10^{-3}$ |
| Our Model | 0.680 | **0.803** | **0.737** |

is almost guaranteed to be right when tagging seed words. Both recall and F-score however see dramatic improvements. We've also included metrics from models by Cartwright & Brent (1997) and Mintz (2003), two frame-based minimally-supervised models of childhood syntactic category learning, to show that our model outperforms (based on F-score) existing models in this area.

Success on the CHILDES data demonstrates our model's ability to learn the words in child-directed speech and success on the Chicago corpus from Carlson et al. (2014) demonstrates our model's ability to learn words children would normally learn.

## 6 Conclusions

We present a model that can be used both for child language acquisition and minimally-supervised POS tagging. Unlike existing POS taggers, our model is cognitively motivated and surprisingly simple for the high accuracy it achieves.

Not only are the POS-POS frames generated by our model consistent with theories of child language learning, but our model also greatly outperforms the baseline and performs similarly to the cognitively implausible Haghighi & Klein (2006) PROTO model. It actually far outperforms the more similar Haghighi & Klein (2006) BASE model that does not rely on solving a complex statistical optimization problem.

We believe future work should investigate the effectiveness of this approach in other languages, especially those which have less morphology than English since the models that rely on morphological features (e.g. Haghighi & Klein, 2006) would be at a disadvantage on this data. On the other hand, since our model operates exclusively at the word-level, inclusion of morphological features may improve the performance of our own model for English. It may also be beneficial to track finer-grained learning data for the model, such as learning speed over time.

The success of our model provides support for the principles of child language acquisition which it is built upon: semantic bootstrapping, distributional frames, reinforcement learning, and the Tolerance Principle.

Additionally, this work has relevancy to POS tagging in that it provides a promising new way to approach minimally-supervised POS tagging. The accurate labeled data necessary to train supervised POS taggers is both expensive to create and likely unavailable for most languages. Since our model requires less than 2% of the data as the current state-of-the-art POS tagger (Akbik et al., 2019) requires, and the vast majority of this data does not need to be labeled, future work that builds upon our model may finally make POS tagging for such languages possible.

## References


Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations)*, 54–59. Minneapolis, MN: Association for Computational Linguistics.

Brent, Michael R. & Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition* 81(2). B33–B44.

Brown, Richard W. 1957. Linguistic determinism and the part of speech. *Journal of Abnormal Psychology* 55(1). 1–5.

Brown, Roger. 1973. *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Brown, Ronald & Cleo Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes (ed.), *Cognition and the development of language*, 11–53. New York: Wiley.

Carlson, Matthew, Morgan Sonderegger & Max Bane. 2014. How children explore the phonological network in child-directed speech: A survival analysis of childrens first word productions. *Journal of Memory and Language* 75. 159–180.

Cartwright, Timothy A. & Michael R. Brent. 1997. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition* 63(2). 121 – 170.

Chang, Chien-Ju. 2003. Talking about the past: How do chinese mothers elicit narratives from their young children across time. *Narrative Inquiry* 13(1). 99–126.

Chemla, Emmanuel, Toben H. Mintz, S. William Fernando Bernal & Anne Christophe. 2009. Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science* 12(3). 396–406.

Cheung, Hintat & Chien-Ju Chang. 2017a. Mandarin TCCM-Reading corpus. *TalkBank* doi:10.21415/QV4V-1Q83.

Cheung, Hintat & Chien-Ju Chang. 2017b. TCCM mandarin corpus. *TalkBank* doi:10.21415/T5Z616.

Christodoulopoulos, Christos, Sharon Goldwater & Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 575–584. Cambridge, MA: Association for Computational Linguistics.

Deng, Xiangjun & Virginia Yip. 2018. A multimedia corpus of child mandarin: The tong corpus. *Journal of Chinese Linguistics* 46(1). 69–92.

Freudenthal, Daniel, Julian Pine, Gary Jones & Fernand Gobet. 2013. Frequent frames, flexible frames and the noun-verb asymmetry. In *Proceedings of the annual meeting of the cognitive science society*, vol. 35, 2327–2332. Austin, TX: Cognitive Science Society.

Gelman, Susan A., John D. Coley, Karl S. Rosengren, Erin Hartman, Athina Pappas & Frank C. Keil. 1998. Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development* 63(1). i–148.

Gelman, Susan A., Marianne G. Taylor, Simone P. Nguyen, Campbell Leaper & Rebecca S. Bigler. 2004. Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development* 69(1). I–142.

Gelman, Susan A., Elizabeth A. Ware, Felicia Kleinberg, Erika M. Manczak & Sarah M. Stilwell. 2014. Individual differences in children's and parents' generic language. *Child development* 85(3). 924–940.

Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou & John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1(1). 23–64.

Haghighi, Aria & Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the human language technology conference of the NAACL, main conference*, 320–327. New York City, USA: Association for Computational Linguistics.

Hart, Betty & Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young american children*. Paul H Brookes Publishing.

Li, Hongmei & Jing Zhou. 2015. *Study on dinner table talk of preschool children family in shanghai*. East China Normal University MA thesis.

Li, Linhui. 2019. Chinese li shared reading corpus. *TalkBank* doi:10.21415/YV4E-DT03.

Li, X.Y. & Jing Zhou. 2004. *The effects of pragmatic skills of mothers with different education on childrens pragmatic development*. Shanghai, China Nanjing Normal University MA thesis.

Luo, Ya-Hui, Catherine E. Snow & Chien-Ju Chang. 2012. Mother-child talk during joint book reading in low-income american and taiwanese families. *First Language* 32(4). 494–511.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu & Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* 57(4). 1–182.

Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19(2). 313–330.

Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90(1). 91–117.

Newman, Rochelle S., Meredith L. Rowe & Nan Bernstein Ratner. 2016. Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language* 43(5). 1158–1173.

Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Reeder, Patricia A., Elissa L. Newport & Richard N. Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology* 66(1). 30–54.

Rondal, Jean A., Martine Ghiotto, Serge Bredart & Jean-Franois Bachelet. 1987. Age-relation, reliability and grammatical validity of measures of utterance length. *Journal of Child Language* 14(3). 433–446.

Rowe, Meredith L. & Susan Goldin-Meadow. 2009. Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science* 323(5916). 951–953.

Schuler, Kathryn D., Patricia A. Reeder, Elissa L. Newport & Richard N. Aslin. 2017. The effect of zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development* 13(4). 357–374.

Schuler, Kathryn D., Charles Yang & Elissa L. Newport. 2016. In Papafragou, Anna, Daniel Grodner, Daniel Mirman & John Trueswell (ed.), *Proceedings of the 38th annual conference of the cognitive science society*, 2321–2326. Austin, TX: Cognitive Science Society.

Tomasello, Michael. 2000. Do young children have adult syntactic competence? *Cognition* 74(3). 209–253.

Yang, Charles. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: The MIT Press.

Zhou, Jing. 2001. *Pragmatic development of mandarin speaking young children: from 14 months to 32 months*: The University of Hong Kong dissertation.

Zhou, Jing. 2018. Chinese zhou3 corpus. *TalkBank* doi:10.21415/T5G40S.