# Searching for Determiners with an Infinite Relational Model

Avril Kenney
9.66

18 December 2009

## Introduction

The purpose of this project was to use an infinite relational model (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006) on the distributions of words in parents' speech to see whether it could infer part-of-speech categories, and in particular whether it would find a category of determiners. This is relevant to an interesting debate in the language acquisition literature about whether children initially represent determiners as an abstract syntactic category (as adults presumably do) and whether such a category representation must be innate or is learnable from the input (e.g., Valian, 1986; Pine & Martindale, 1996). If the model were able to categorize determiners as a distinct part of speech just from simple distributional data, this would be evidence that even if children do have a determiner category, it is not necessarily innate.

## Model

The model used was a simple version of the infinite relational model (IRM). This model takes a group of items and the values of a two-place relation on each pair of the items, and places them in categories based on the relation. In this case, the items were words, and the value of the relation $R(i, j)$ between words $i$ and $j$ was the number of times that word $i$ directly preceded word $j$ in the corpus data.

The posterior probability of any grouping is given by Bayes' Rule:

$$P(z|R) \propto P(R|z)P(z), \tag{1}$$

where the prior probability $P(z)$ of any grouping $z$ is distributed according to a Chinese Restaurant Process (CRP) with a preset parameter $\gamma$ (results are reported for $\gamma$=0.1, but were similar for other values). The probability of the relation occurring between any instance of a pair $(i, j)$ of items is given by the parameter $\eta(a, b)$, where $a$ is the category that $i$ is in and $b$ is the category that $j$ is in. The prior on each entry of $\eta$ follows a Beta distribution with both parameters preset to a value $\beta$ (which in this case was 1, a uniform distribution).

The likelihood $P(R|z)$ of the relational data given the grouping $z$ is therefore given by marginalizing over all possible values of $\eta$:

$$P(R|z) = \int_0^1 P(R|z,\eta)P(\eta)d\eta, \tag{2}$$

where

$$P(R|z,\eta) = \prod_{i=1}^{n}\prod_{j=1}^{n}\eta(z_i,z_j)^{R(i,j)}(1-\eta(z_i,z_j))^{\mathrm{zero}(R(i,j))} \tag{3}$$

and

$$\mathrm{zero}(R(i,j)) = \begin{cases} 1 & \text{if } R(i,j)=0 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Equation (2) then simplifies to

$$P(R|z) = \prod_{a,b\in\mathbb{N}} \frac{\mathrm{Beta}(\beta + \sum_{i\in a, j\in b} R(i,j), \beta + \sum_{i\in a, j\in b} \mathrm{zero}(R(i,j))}{\mathrm{Beta}(\beta,\beta)}. \tag{5}$$

Samples from the posterior were drawn using Markov Chain Monte Carlo: on each iteration, probabilities were calculated for the reassignment of each item to each category (or a new category), and the next grouping was chosen with those probabilities.

## Data

The data were taken from 2314 annotated transcripts of parent–child interactions from the CHILDES corpus (MacWhinney, 2000). Using the words spoken by the children's mothers or fathers, counts were obtained of how many times each word directly preceded each other word (not including across utterance boundaries). These were then restricted to only nouns, determiners, and adjectives, that occurred at least 600 times before or after each other. This left 43 words: 32 nouns, 5 determiners, and 6 adjectives.

## Results

The categories found by the IRM did not correspond to the parts of speech. It tended to group words together that have similar meanings (and hence occur with similar other words), but it found a lot more than three categories. The grouping with the highest posterior probability found by the model is shown in Figure (1).

Figure (2) shows that although words that were the same part of speech were often put in different categories, on virtually no iterations were words grouped with other words that were actually different parts of speech.

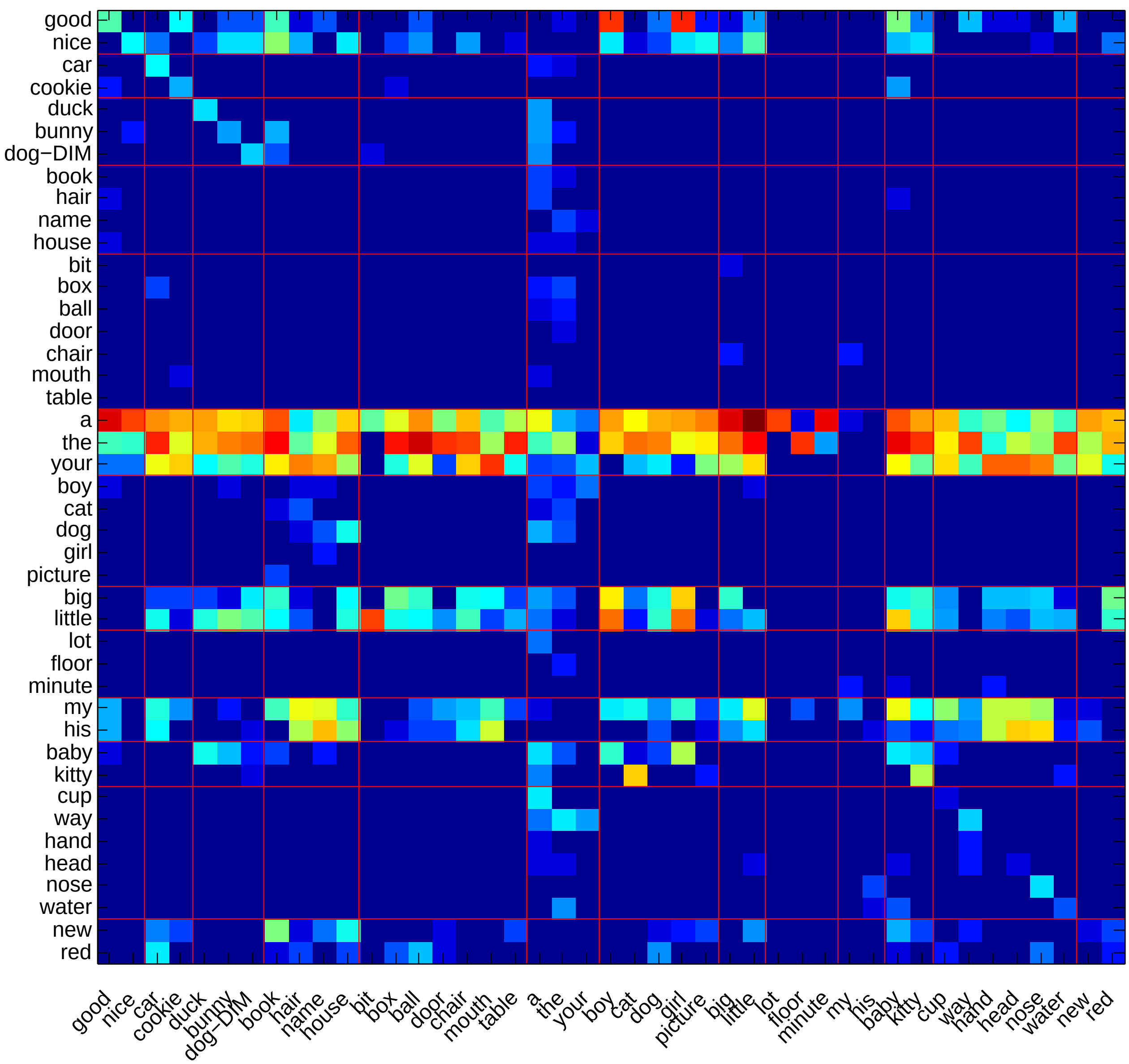


Figure 1: The grouping with the highest posterior probability found by the IRM. The red lines denote category boundaries. The color of each square represents the number of times in the data that the word on the vertical axis preceded the word on the horizontal axis (dark blue is low values, dark red is high values).

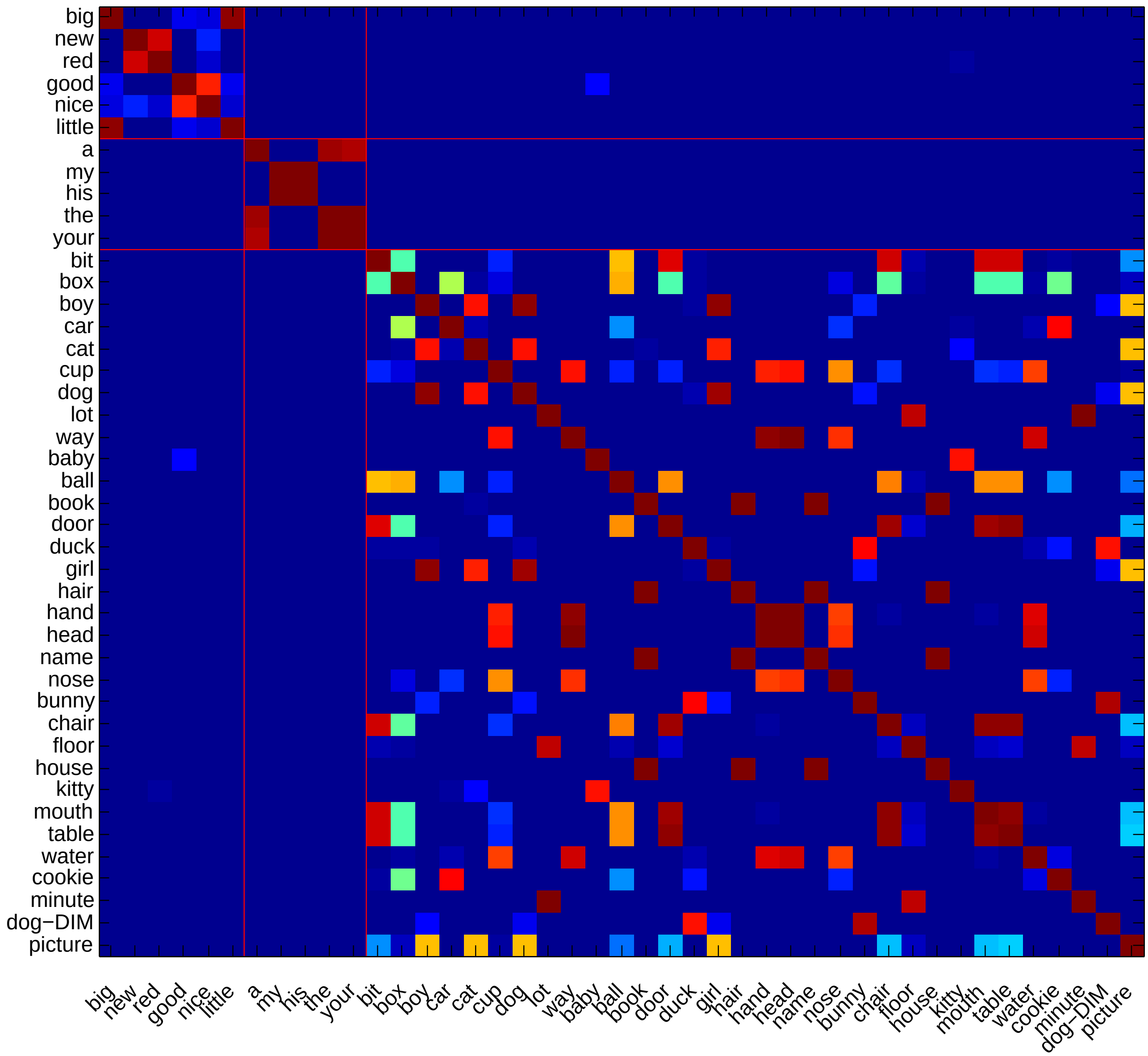


Figure 2: Probabilities of correct categorization. The red lines divide the words according to their actual part-of-speech categories. The color of each square represents the proportion of iterations in which the pair of words was placed in the same category.

Because the IRM was finding too many categories, the data were also put into a finite relational model (FRM), in which the number of categories is specified. When it was specified as three (the actual number – nouns, determiners, and adjectives), the categories found by the model did not correspond to the part-of-speech categories (Figure 3). All the determiners were put into the same category, but this category also contained three of the adjectives. The nouns were split across two categories, one of which also contained the other three adjectives.

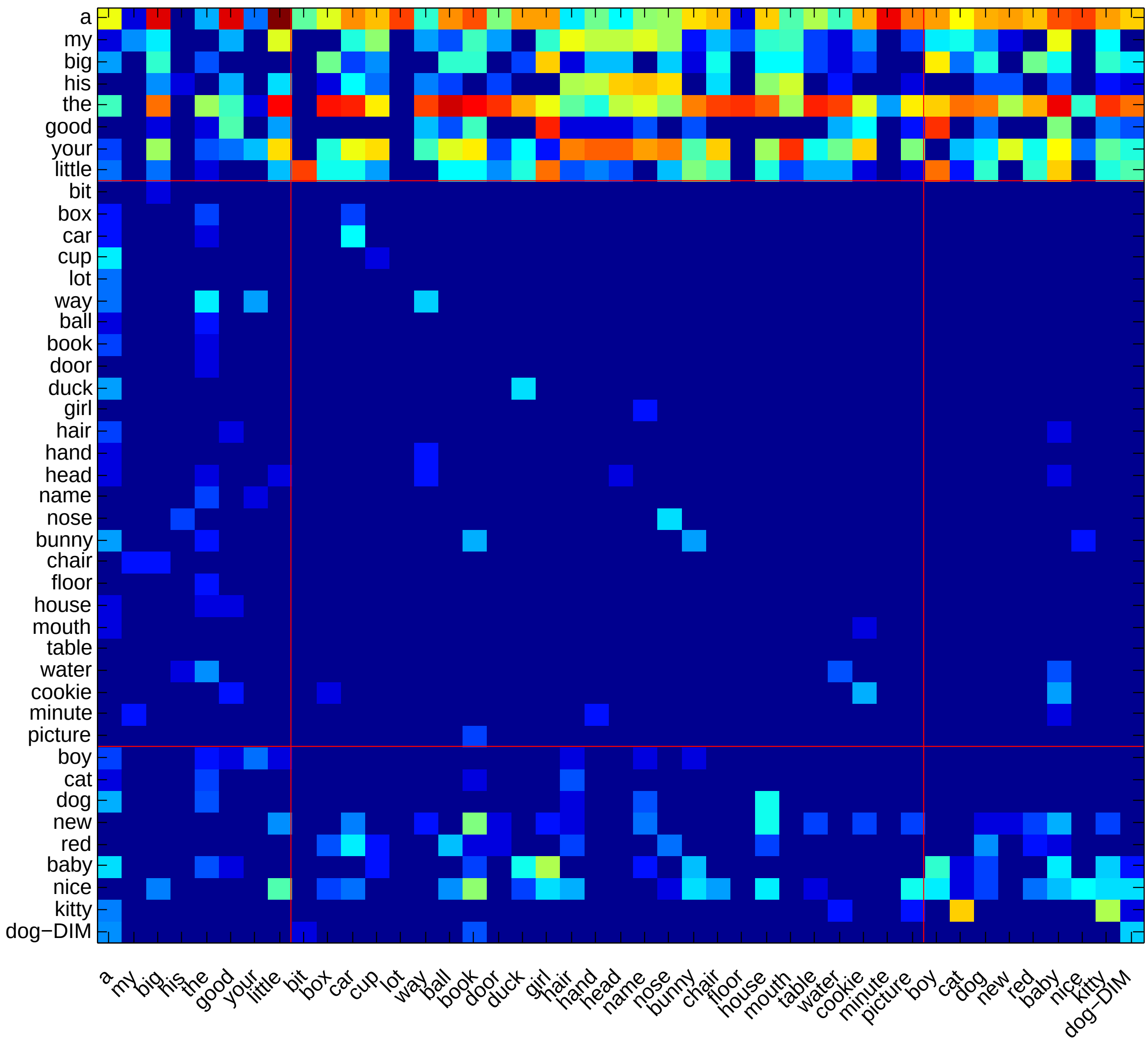


Figure 3: The grouping with the highest posterior probability found by the FRM with three categories.

However, when the model was told that there were four categories, its result aligned fairly closely with the part-of-speech categories (Figure 4). There was one category consisting of all and only the determiners. Another category contained all of the adjectives plus two nouns ("baby" and "kitty," both of which can often precede other nouns as modifiers). The remaining nouns were divided into two categories.

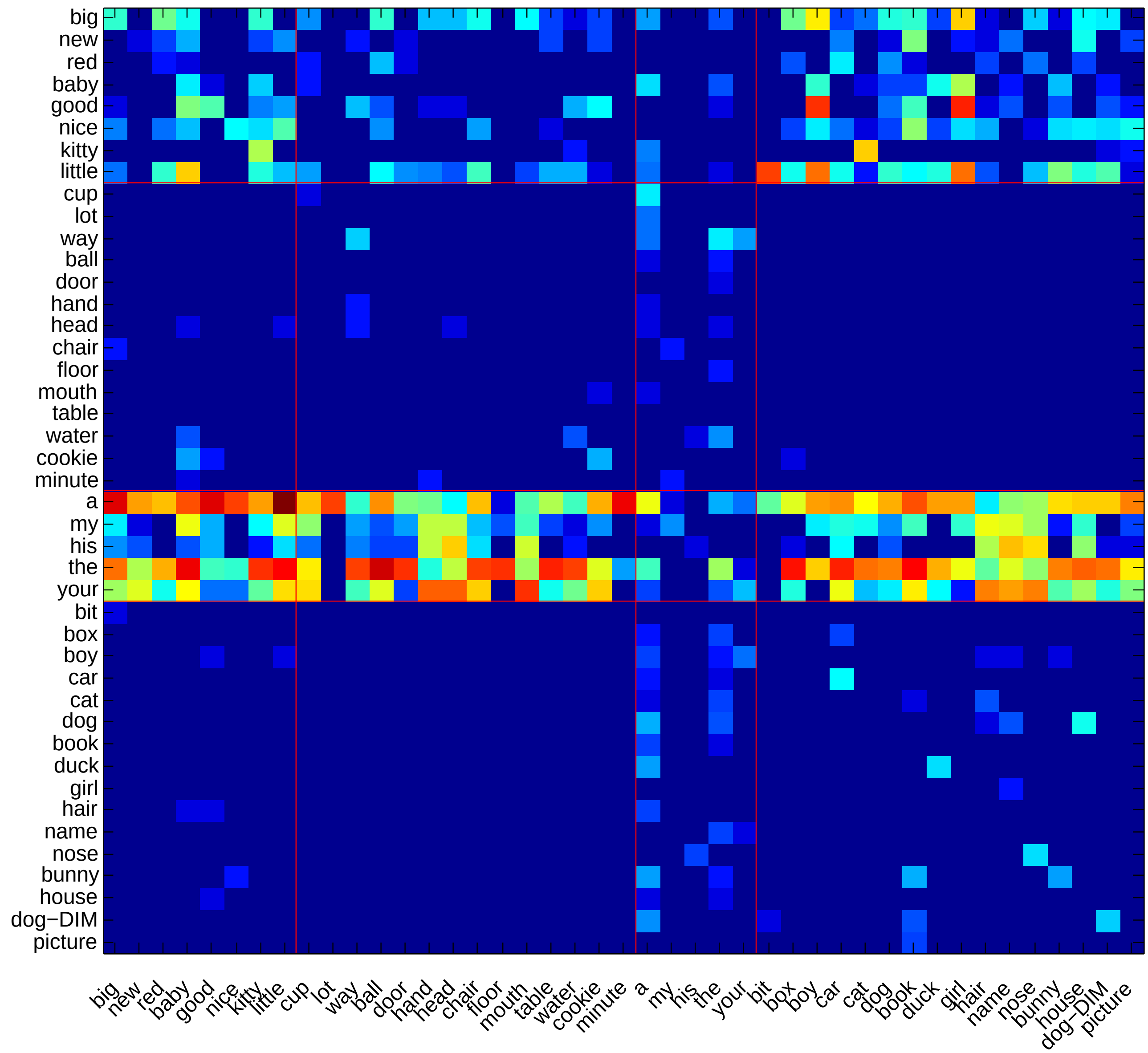


Figure 4: The grouping with the highest posterior probability found by the FRM with four categories.

## Discussion

The IRM did not infer part-of-speech categories from the distributional data. This is perhaps not surprising, because even if different words that are the same part of speech can all theoretically occur in the same positions, where they actually do occur is strongly affected by their meanings, as shown by the categories that the IRM did group the words into. The FRM was able to find a determiner category, but only when it was told there were four categories instead of three. This suggests that (in this data set) determiners do have similar enough distributions to be grouped together, but not enough to be established as a unique category without providing further information.

In addition to not considering the meanings of the words, the data were in several ways not representative of the actual input. For one thing, the relation was only on pairs of adjacent words, which is not always indicative of syntactic structure even among just these three parts of speech. The format of the data also did not take into account that occurrences of any individual word might not be the same type of usage, information which a learner could infer from knowledge about the semantics or prosody without any syntactic representation. For example, the word "bit" often occurs after "little" (in which case it is probably a noun) but it also occurs after "doggie" (in which case it is probably a verb). And there was no consideration of pauses in speech (which could signal phrase boundaries within utterances) or disfluencies (such as repetitions of the same word, or use of "a" as a filler syllable).

The results of this investigation do not provide strong evidence either way in terms of the learnability of the determiner category. Interesting extensions of the present work would be to include more specific relations in the model, such as semantic features and additional information about word order.

## Acknowledgements

I would like to thank Steve Piantadosi for suggesting the idea of working on determiners and for lots of helpful comments and always answering my questions.

## References

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence.*

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Pine, J. M., & Martindale, H. (1996). Syntactic categories in the speech of young children: the case of the determiner. *Journal of Child Language*, *23*, 369-395.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*(4), 562-579.