Of Frames and Frequencies:
How Early Language Production is Influenced by the Distribution

Revanth Kosaraju, Melody Dye & Michael Ramscar
Department of Psychology, Stanford University

ABSTRACT

Accounts of language acquisition differ significantly in their treatment of the role of distributional information in language learning and comprehension. In particular, nativist accounts posit that probabilistic learning about the distributions of words in a language has little to do with how children come to use and understand that language. **We examined the accuracy of this claim by testing how well 3-4 year olds were able to comprehend and repeat simple expressions (or "chunks")**. In our study, we contrasted performance on high frequency expressions (such as "poured tea into a cup") against performance on corresponding lower frequency expressions (such as "poured milk into a glass"). Corresponding chunks were the same length, expressed similar content, and were all grammatically acceptable, yet the results of our study showed marked differences in performance when the *overall* frequency of the expression varied, which persisted even when individual word frequency was kept constant. **We found that a distributional model of language predicted our empirical findings better than a number of other prominent models, including syntactic, independent-probability and Markov models.**

Extension of Prior Research

This work was designed as a follow-up to a study conducted by **Bannard & Matthews (2008)**, showing that a young child's ability to repeat brief expressions ("chunks") is moderated by overall chunk-frequency. We wanted to replicate this study with new materials & to test for sensitivity both in comprehension and production, with an <u>adult language corpus</u> (Bannard & Matthews had used a corpus of child-directed speech). We also wanted to examine whether our empirical findings would link up better with a distributional (chunk-based) model of language probability or a grammar-based model. In this, we hoped to determine: 1) whether – contrary to nativist claims – children were indeed sensitive to probabilistic information in the input and 2) whether a distributional model would predict this sensitivity more accurately than other models.

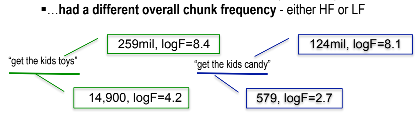
QUESTION / HYPOTHESIS

- **Question**: Will children be better at repeating higher-frequency (as opposed to lower-frequency) expressions from an <u>adult</u> corpus of language?
  - **Hypothesis**: If children are indeed sensitive to the distributional probabilities of language, higher frequency expressions from an adult corpus will be repeated with greater proficiency than lower frequency expressions
  - *Even when controlling for*: the overall length of the expression, individual word frequency, and the grammatical structure of sentence
  - We therefore predict that the <u>distributional model</u> will correlate best with results

EXPERIMENT

Methods: To test the effect of chunk frequency on proficiency of repetition, we created:

- 28 sets of corresponding High-Frequency (HF) and Low-Frequency (LF) Expressions, using non-locative alternating verbs
- In any given set, the two corresponding expressions:
  - …were both the same length
  - …were both grammatically correct
  - …made use of the same verb
  - …had the same individual word frequencies (**Zipf, 1935**)
  - **…had a different overall chunk frequency** - either HF or LF



Participants: 31 children, aged 3-4 took part in the experiment, with 15 in Condition 1 and 16 in Condition 2 (the conditions simply counterbalanced the position of the corresponding HF and LF expressions).

Measures: During testing, the 56 expressions were read to the child, one at a time, and the child was asked to repeat the expressions. Errors the children made during repetition were noted on the testing sheet for further observation and analysis after the experiment. In addition, an audio recorder was kept running over the entire course of the experiment. The audio recorder was used to measure delay times between comprehension of the expression and repetition. **The test thus measured both delay of repetition and accuracy of repetition.** We predicted that children would be faster and more accurate at repeating expressions with higher overall chunk frequencies.

PROBABILISTIC MODELING

CORPORA: We used Google and the Contemporary Corpus of American English (COCA) to determine individual word frequencies and larger phrase frequencies. Google mirrors COCA in frequency trends (**Ramscar, Matlock & Dye, 2010**).

Distributional Model

Models the probability of each expression as a function of its frequency as a whole unit. 'Chunk' frequency was established by determining the number of hits that appear on Google for that expression enclosed in quotes. **Predicts potential differences in comprehension and repetition across corresponding chunks.**

Markov Model

Models the probability of each expression as the sum of the transitional probabilities of the bigrams across the chunk. The transitional probability of each bigram was determined by dividing the number of occurrences of each bigram by the number of occurrences of the first word of the bigram (e.g., divide occurrences of "filled the" by the number of occurrences of "filled"). These probabilities were calculated for each bigram within the chunk and then summed. **Predicts potential differences in comprehension and repetition across corresponding chunks.**

Syntactic Model

Models the probability of each expression as the probability that a given part of speech ('grammatical class') will follow another part of speech. Because the parts of speech that made up our corresponding expressions were matched (e.g., "filled a glass with milk" and "filled a cup with tea" are both *verb+article+noun+preposition+noun* sentences), a syntactic model generates equal probabilities for the corresponding expressions used in our experiment, meaning that it **does *not* predict any differences in comprehension or repetition across corresponding chunks.**

Independent Probability Model

Models the probability of each expression as the sum of the chunk's individual word frequencies. For example, P["throw a ball at him"] = P[throw]+P[a]+P[ball]+P[at]+P[him]. Because we kept the individual frequencies of words constant across expressions, this model generates equal probabilities for corresponding expressions, meaning it **does *not* predict any real differences in comprehension or repetition across corresponding chunks**.

Model Correspondence Measure

We then determined how well each other model—independent-probability, syntactic, and Markov—corresponded with the distributional model, using a logs-odd equation. Specifically, we estimated the probabilities generated by each model for each expression and compared them to the chunk model's determination. As can be seen, *none of them corresponded well with the chunk model, or made similar predictions.*

- **Log-odds formula**
  - (Occurrences of high-frequency expressions/Total occurrences of high + low-frequency expression)*Log (total occurrences of high + low frequency expressions)

Model	% of Correspondence
Independent	**53.3%**
Syntactic	**53%**
Markov	**52.6%**

EMPIRICAL RESULTS

- **Repetition accuracy**
  - Statistically significant differences only for the distributional model
    $t(25)=2.18$, $p<0.05$
  - Higher-frequency expressions were repeated with much higher accuracy than lower-frequency expressions
- **Repetition delay**
  - Statistically insignificant differences for all four models; most children repeated 1-2 seconds after being told to do so regardless of frequency

ANALYSIS

Analysis of model fit with empirical findings: which model is most accurate?

Distributional Model

Measure	t Stat	t Critical	P value	Significance
Repetition accuracy	2.18	2.06	**0.04**	**Significant**
Repetition Delay	-1.4	2.06	0.18	Insignificant

Markov Model

Measure	t Stat	t Critical	P value	Significance
Repetition accuracy	0.426	2.06	0.67	Insignificant
Repetition Delay	0.9	2.06	0.36	Insignificant

Syntactic Model

Measure	t Stat	t Critical	P value	Significance
Repetition accuracy	-1.0	2.06	0.34	Insignificant
Repetition Delay	0.1	2.04	0.91	Insignificant

Independent Probability Model

Measure	t Stat	t Critical	P value	Significance
Repetition accuracy	0.72	2.06	0.48	Insignificant
Repetition Delay	0.1	2.06	0.92	Insignificant

Analysis of Model Fit

- Percentage of correspondence (POC) between models and results
  - POC formula derived from differences in repetition accuracy for high-frequency and low-frequency expressions

Model	POC value
Distributional	**73.5%**
Markov	41.6%
Independent Probability	33.3%
Syntactic	24.4%

****The Distributional Model is by far the best model for our findings****

DISCUSSION

- Our empirical work and analysis replicate and strengthen the findings of **Bannard & Matthews (2008)**. We find that children are sensitive to the frequency of whole sequences of words ('chunks') in their linguistic input, and not simply to individual word frequencies or formal grammatical properties. This finding – which suggests that children are attending to and learning about the <u>distributional properties</u> of English – is *not* predicted by a syntactic (nativist) account of language, and is much more readily explicable in terms of a predictive, probabilistic account (**Ramscar et al., 2010**).

SELECTED BIBLIOGRAPHY

--Bannard, C., & Matthews, D. (2008, March 3). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, 19(3), 241-48.
--Chomsky, N. (1959) Review of Verbal Behavior, by B.F. Skinner. *Language 35: 26-57.*
--Davies, M. (n.d.). Corpus of Contemporary American English. Retrieved August 14, 2009, from Brigham Young University Web site: http://www.americancorpus.org/
--Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press.
--Ramscar, M., Matlock, T., & Dye, M. (2010). Running down the clock: the role of expectation in our understanding of time and motion. *Language and Cognitive Processes*, 25(5), 589-615.
--Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). Feature-Label-Order effects and their implications for symbolic learning. *Cognitive Science*. DOI: 10.1111/j.1551-6709.2009.01092.x
--Yarlett, D. (2008). Language Learning Through Similarity-Based Generalization. PhD Thesis, Stanford University.
--Zipf G.K. (1935). The Psychobiology of Language. New York, NY: Houghton-Mifflin.