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 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 frequencies 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.

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 that expression encoded 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., “drank the tea of” by the number of occurrences of “drank the”). 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 make 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, “pick[a+article]a+[noun] up” = “pick”+ “[a+article]”+ “[a+article]”+ “[noun]”+ “[noun]”+ “[noun]”+ “[noun]”. Because we kept the individual frequencies of words constant across corresponding 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 model—indeed, independent-probability, syntactic, and Markov—corresponded with the distributional model, using a log-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 expressions

**Distributional Property**

- Predicts potential differences in comprehension or repetition across corresponding chunks.

**Syntactic Property**

- Because the parts of speech that make 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.

EMPIRICAL RESULTS

- **Repetition accuracy**
  - Statistically significant differences only for the distributional model
  - 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

**Model Correspondence Measure**

- % of Correspondence

**Model**

- Independent
- Syntactic
- Markov

**% of Correspondence**

- 52.6%
- 49%
- 58.6%

**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**

**SELECTED BIBLIOGRAPHY**


