World Academy of Science, Engineering and Technology
International Journal of Psychological and Behavioral Sciences
Vol:4, No:6, 2010

The Predictability and Abstractness of Language: A Study in Understanding and Usage of the English Language through Probabilistic Modeling and Frequency

Revanth Sai Kosaraju, Michael Ramscar, and Melody Dye

***Abstract*—**Accounts of language acquisition differ significantly in their treatment of the role of prediction in language learning. In particular, nativist accounts posit that probabilistic learning about words and word sequences has little to do with how children come to use language. The accuracy of this claim was examined by testing whether distributional probabilities and frequency contributed to how well 3-4 year olds repeat simple word chunks. Corresponding chunks were the same length, expressed similar content, and were all grammatically acceptable, yet the results of the study showed marked differences in performance when overall distributional frequency varied. It was found that a distributional model of language predicted the empirical findings better than a number of other models, replicating earlier findings and showing that children attend to distributional probabilities in an adult corpus. This suggested that language is more prediction-and-error based, rather than on abstract rules which nativist camps suggest.

***Keywords*—**Abstractness, Child psychology, Language acquisition, Prediction and error

I. INTRODUCTION

THE question of how children acquire language is still very open. Many theories have been developed in an attempt to explain how children learn vast, complex languages within just a few short years. Many theories maintain that children are born with an innate language acquisition device. Another approach is to state that children use prediction and prediction-error to drive their learning, enabling them to learn language from their experience with the environment [13], [14]. While nativist theory argues that children are born with the abstract rules for language firmly imprinted within them, many proponents of the prediction-and-error theory suggest that children attend to the distributional properties of language and rely on this information in repetition and comprehension, rather than the more abstract representations proposed elsewhere [7], [16].

Revanth S.Kosaraju is with the Harker School, San jose, CA 95129 USA (e-mail: 12revanthk@students.harker.org).

Michael Ramscar is with the Psychology Department, Stanford University, Stanford, CA 94305 USA (e-mail: ramscar@stanford.edu).

Melody Dye is with the Psychology Department, Stanford University,Stanford , CA 94305 USA (e-mail:pkipsy@stanford.edu).

If children are indeed sensitive to probabilistic information in the input as several recent studies suggest [1], [9], [12], [15], it should be possible to test for this sensitivity. The author's particular interest, here, was to follow up from a recent study conducted by Bannard and Matthews [2], showing that a young child's ability to repeat brief phrases is moderated by distributional frequency. Specifically, Bannard and Matthews found that young children are better at reproducing high-frequency chunks than corresponding low-frequency chunks, even when the frequency of the individual words is held constant, and even when both chunks are both grammatically acceptable. This essentially showed that children attend to the distributional probabilities of language in a child-directed corpus. The authors wanted to replicate this study with new materials, and to test for sensitivity in reproduction for an adult corpus of speech, testing at the same time the question of abstractness versus predictability in language. The authors also wanted to examine whether the empirical findings would be more consistent with a distributional-based model of language probability or a grammar-based model. In this way, they hoped to test whether sensitivity to probabilistic information in the input was more in line with an empiricist or nativist account.

A. Brief Overview of Error-Driven Learning

In recent years, it was found that learning problems can be prediction-and-error driven processes (learning the past tense, [8]; learning irregular plurals, [9], [12]; learning color words, [13]; learning contextual cues, [10]). Further, these findings suggest that children do not learn words as determinate concepts, as suggested by logical theories, but rather learn probabilistic, predictive relationships between objects and sounds that develop over time. These results demonstrate how a simple model of error-driven learning can have tremendous predictive power over a range of learning tasks. It seems then, that children's knowledge of words may be probabilistic. But is their knowledge of word co-occurrence patterns similarly defined by distributional probabilities? This is the question the authors sought to examine.

II. Experiment

The aim of this study was to examine whether children would be better at reproducing high-frequency expressions as compared to lower-frequency expressions from an adult corpus of language, when overall length, individual word frequency, and semantics were controlled for. Further, the authors hoped to compare the empirical findings from the study with the predictions of a number of prominent linguistic models. It was hypothesized that a model of distributional frequency —which estimates the probability of the expression based on the occurrence frequency of the entire chunk in a distribution, rather than as the sum of the individual word frequencies—would be best at predicting our results. This would signify that children, rather than remembering specific strings of words, actually are sensitive to the distributional probabilities and thus the distributional model of language because they would be better at repeating higher-frequency expressions than corresponding lower-frequency ones.

A. Methods & Design

The main manipulation in the study is derived from a particular class of alternating verbs that display 'locative alternation' [6]. Locative alternating verbs have the peculiar property that the location of the words relative to the verb is said to define the grammaticality of the construction. For example, "filled a glass with milk" is an expression with the alternating verb *filled*. This expression is more grammatically acceptable than "filled milk into a glass."

In this experiment, locative alternating verbs were used to control for grammaticality while manipulating chunk frequency. For example:

"filled a glass with milk" vs. "filled a cup with tea"

and

"poured milk into a glass" vs. "poured tea into a cup."

The four constructions above are divided into two groups based on the alternating verb. In the first pair, the expression "filled a glass with milk" is more frequent than "filled a cup with tea." Conversely, in the second pair, "poured tea into a cup" is more frequent than "poured milk into a glass." The expressions used to test the children were designed in groups of four similar to the group above. The frequency of the nouns and prepositions was kept constant across chunks, to prevent any possible biasing effect (e.g., glass and cup are similar frequencies, as are milk and tea, as are into and with).

The expressions were created in this way so the performance as it related specifically to chunk-frequency could be easily compared. Fifty-six such expressions were created in fourteen groups of four. For each expression, Google and The Corpus of Contemporary American English (COCA) were used to determine the relative frequencies of both individual words and groups of words in a phrase. Prior research has shown that both corpora identify nearly identical frequency trends in English and express virtually the same comparative frequency patterns [10], lending the authors to believe that using Google and COCA interchangeably should not have produced any significant inconsistencies in their modeling work.

After designing these expressions, several different modeling methods were used to determine the frequency of each of the expressions. This was done to determine how well these methods fit empirically with the findings.

B. Probabilistic Modeling

The question of how the frequency of a chunk should be determined can vary widely depending on the model being used and the theory that informs it. In recent years, three prominent probabilistic models that have been used to understand the relative frequencies of certain constructions include the independent probability model, the general construction 'syntactic' model, and the Markov model. In addition to these models, the authors devised a "distributional" model, elaborated below.

1) Independent Probability Model

The basic tenet of the independent probability model is that, given a group of words, the probability of any one arrangement of them occurring should be equal to the probability of any other arrangement of them occurring. The independent probability model is thus a unigram model of language and therefore models chunk probability 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]. Notably, because the individual frequencies of words were kept constant across expressions (e.g., the expression "filled a glass with milk" is matched for individual frequencies with "filled a cup with tea"—glass is approximately the same frequency as cup, and milk is approximately the same frequency as tea), this model generates equal probabilities for corresponding expressions.

2) Syntactic Model

A syntactic model expresses chunk probability from a grammatical standpoint, calculating the probability that a given part of speech ('grammatical class') will follow another part of speech. For example, for the expression "filled a glass with milk," one might calculate the number of occurrences of "filled + article" divided by the total occurrences of "filled." Again, it is important to note, that because the parts of speech that made up the 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. This means, that like the independent-probability model, it does not predict any differences in repetition across corresponding chunks.

3) The Markov Model

Markov models express chunk probability as the sum of the transitional probabilities of the bigrams across the chunk. In this case, 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. This model potentially predicts differences between the corresponding chunks we chose. For example, "juggle balls" has a much higher transitional probability than "juggle chairs," even though the independent frequencies of the individual words are similar and even though the same grammatical classes are present in both expressions.

4) Distributional Model and Log-Odds

Finally, a distributional model was devised to assess the probability of each expression as a function of its frequency as a whole unit in the distribution of the English language. For each expression, total frequency was established by determining the number of hits that appear on Google for that expression enclosed in quotes. The relative probability of the high and low frequency chunks was established using the following log-odds formula:

(Occurrences of high-frequency expressions/Total occurrences of high + low-frequency expression)*

Log (total occurrences of high + low frequency expressions)

The log-odds formula was used to combine actual frequency of occurrence with relative percentage of occurrence. The logarithmic part of the equation accounts for the absolute number of occurrences, while the fractional part accounts for the percentage of occurrence.

TABLE I
EXAMPLE OF LOG-ODDS CALCULATION

Expression	Number of Occurrences	Calculated Log-Odds	Log-Odds Formula
"filled a glass with milk"	33	1.08900212	Log(52)* 33/52

C. Assessing Model Correspondence

Using the logs-odd equation it was determined how well each other model—independent-probability, syntactic, and Markov—corresponded with the distributional model. While there was no apparent pattern of correspondence between the three original models themselves, all three had relatively similar correspondence with the distributional model (53.3% for the independent-probability, 53% for the syntactic, and 52.6% for the Markov). The correspondence of about 50% suggested that none of the models had a significant correlation with the distributional model or made similar predictions. The formula indicated in Fig. 1 was used to calculate fractional correspondence. The sum of the log-odds was taken across every expression for each model, with a positive log-odds used if the model predicted the same frequency (i.e. high or low) for an expression as the distributional model and a negative log-odds used if the converse was observed. This was divided by the total log-odds for the distributional model, with the addition of a scaling factor (the total log-odds for the distributional model) to both the numerator and the denominator to create a range for fractional correspondence from 0 to 1.

$$\frac{\sum\text{Log-odds for each model (+ value for match and – value for no match)} + (\sum\text{Log odds for Distributional model})}{\sum\text{Log-odds for Distributional model} + (\sum\text{Log odds for Distributional model})}$$

Fig. 1 Formula for Calculating Fractional Correspondence to Distributional Model.

D. Participants

25 children between the ages of 3 and 4 years of age participated in this study. They were living in the vicinity of Stanford University, primarily raised in the United States, and had the ability to speak English fluently. This is the same age group used in the Bannard and Matthews (2008) study.

E. Materials and Experimental Stimuli

A testing sheet was created with the 56 expressions listed in a randomized order. To further control for any possible ordering effects, a second "condition" in which each high-frequency expression swapped positions with its corresponding lower-frequency expression was created (e.g., "filled a glass with milk" was switched with "poured milk into a glass"). The authors ran roughly half of the subjects in each condition (12, 13).

F. Procedure

The experiment consisted of a repetition test, in which the individual repetitions were judged for accuracy and response-times. 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. Delay time was measured from the

time the child was asked to repeat the expression/point to the correct picture, to the time the child actually began to do so.

III. Results

Two main components were the focus of the results of this study: the comparison of the high and low-frequency expressions using the repetition test, and the evaluation of the distributional model to determine whether an adult corpus would produce similar results to the child-directed corpus used in the Bannard and Matthews study [2]. This would mean that children attend to the distributional properties of language, rather than memorizing specific sets of words and treating them as idiomatic constructions.

A. Comparison of High and Low-Frequency Expressions

Two measures were focused on for the purpose of this study: repetition delay and accuracy of repetition. It was proposed that the delays would be smaller for higher-frequency expressions and that the accuracies would be larger. T-tests were run with the preceding measures for all of the models to assist in data analysis.

It was found that repetition accuracy was significant for the distributional model according to the t-test, whereas for all other models no significant differences were observed between the high and low-frequency expressions. A value of $t(25)=2.18$ for repetition accuracy according to the distributional model and a p value of less than 0.05 indicated that the difference in the results for high and low-frequency expressions was statistically significant and that the null hypothesis should be rejected. Essentially, the high-frequency expressions were repeated with a much higher accuracy than that of the lower frequency expressions. These results based on the distributional model supported the hypothesis. The delays between expressions of differing frequency, however, were insignificant according to the t-test.

In total, the results served to support the hypothesis regarding the distributional model, and the replication of the significance of repetition accuracy validated the results by Bannard and Matthews in 2008.

B. How well did the Models do?

The repetition accuracy was then tested with the distributional model to determine the extent of the effect of distributional probabilities on children.

To compare the distributional model with the results, it was necessary to study the differences in repetition accuracy between high and low-frequency expressions.

Using the equation in Fig. 2, the correspondence between the results and the distributional model was obtained. The summation of d_n (difference between accurate number of high frequency repetitions and low frequency repetitions) was taken across the 25 children who participated in the repetition portion of the study. The numerator consists of the sum of d_n without absolute value to reflect the distributional model. The denominator consists of the sum of the absolute values of d_n across the 25 children to reflect the results. A scaling factor was added to both the numerator and denominator in order to create a percentage of correspondence (POC) from 0 to 100. Without the addition of this scaling factor, POC would range from -100 to 100 percent; the negative values are not realistic. It was discovered that the distributional model was 73.5% in accord with our results, a noteworthy finding.

$$\frac{\sum_{n=1}^{25} d_n(\text{+ value for match, - value for no match}) - \left(-\left(\sum_{n=1}^{25} |d_n|\right)\right)}{\sum_{n=1}^{25} |d_n| - \left(-\sum_{n=1}^{25} |d_n|\right)}$$

Fig. 2 Formula for Calculating Model Correspondence.

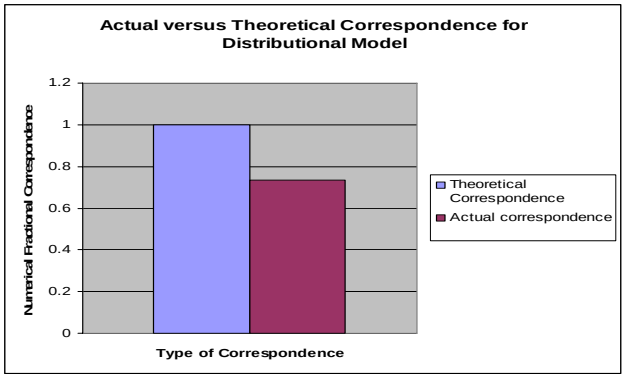

Fig. 3 Percentage of Correspondence between Distributional Model and Results

The observations derived from these results were supportive of the hypothesis and revealing of some extra findings. The results showed that the distributional model is the most accurate model for modeling frequency in terms of repetition accuracy, and validated the hypothesis regarding distributional probabilities affecting children, because children were affected by a sample of language taken from an adult corpus. The distributional model was found to be the most accurate model because none of the other models were observed to have significant differences between repetition accuracy for expressions of different frequency after a T-test was conducted, whereas the distributional model accounted for such differences.

IV. Discussion

One of the interesting aspects of the results was the reason behind the lack of significance in the repetition delay measure. One theory for this is that most children began repeating the expression right after being told to do so by the experimenter, creating an average delay of 1-2 seconds at maximum.

It was found that for the primary measure of repetition accuracy, differences between high and low frequency expressions were significant for the distributional model because 30-35% of the expressions were changed to varying degrees. After studying the various changes made by the children during repetition, it is believed that these differences can largely be accounted for by the children's level of comfort with the expression. Essentially, the children changed the expressions 30-35% of the time because they were trying to say a sentence that was more comfortable to them rather than the sentence that was presented. When using the distributional model, this occurred more with low-frequency expressions than high-frequency expressions, showing that children are in fact affected by the distributional probabilities of language because not only are they influenced by a corpus of child-directed speech, as shown by Bannard and Matthews [2], but also they are influenced by a corpus of adult speech as well.

In conclusion, the hypothesis was supported by the results gained from the experiment. Children were more likely to correctly repeat high-frequency expressions than low-frequency expressions according to the distributional model. The replication of results served to validate Bannard and Matthews' results from 2008 [2] and supported the hypothesis regarding the distributional properties of language. The observations also showed that children attend to distributional probabilities not only in a corpus of child-directed speech but also when being prompted with phrases from an adult corpus of language. This confirmation of the distributional properties of language ties into the abstract qualities of language, showing that language is more driven by prediction-and-error processes rather than abstract associations.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant Nos. 0547775 and 0624345 to Michael Ramscar. Correspondence concerning this article should be addressed to Michael Ramscar at the Department of Psychology, Jordan Hall, Stanford, CA 94305. Electronic mail may be sent to 12revanthk@students.harker.org.

References

[1] Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science, 9,* 321-324.

[2] 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. doi:10.1111/j.1467-9280.2008.02075.x

[3] Chomsky, N. (1967). A Review of B. F. Skinner's *Verbal Behavior*. In L. A. Jakobovits & M. S. Miron (Eds.), *Readings in the Psychology of Language*. Prentice-Hall. Retrieved from http://www.chomsky.info/articles/1967----.htm

[4] Davies, M. (n.d.). *Corpus of Contemporary American English*. Retrieved August 14, 2009, from Brigham Young University Web site: http://www.americancorpus.org/

[5] *Google*. (2009). Retrieved August 14, 2009, from http://www.google.com/

[6] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press. Retrieved August 14, 2009, from http://books.google.com/books

[7] McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgments of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 611-616).

[8] Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*: 45(2), 45-94.

[9] Ramscar, M., & Dye, M. (2009). Expectation and negative evidence in language learning: the curious absence of mouses in adult speech. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

[10] Ramscar, M., Dye, M., Witten, J., & Klein, J. (2009). Two routes to cognitive flexibility: Learning and response conflict resolution in the dimensional change card sort task. *Proceedings of the 31st Meeting of the Cognitive Science Society*, Amsterdam, Netherlands.

[11] Ramscar, M., Matlock, T., & Dye, M. (in press). Running down the clock: the role of expectation in our understanding of time and motion. *Language and Cognitive Processes*.

[12] Ramscar, M. & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*: 31, 927-960.

[13] Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (in press). Feature-Label-Order effects and their implications for symbolic learning. *Cognitive Science*.

[14] Rescorla, R.A., & Wagner, A.R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.

[15] Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

[16] Yarlett, D. (2008). *Language Learning Through Similarity-Based Generalization*. PhD Thesis, Stanford University.