CALIFORNIA STATE SCIENCE FAIR
2010 PROJECT SUMMARY

Name(s)

Revanth S. Kosaraju

Project Number

S0314

Project Title

A Study in Understanding and Usage of the English Language through Probabilistic Modeling and Frequency

Abstract

Objectives/Goals

Theories of language acquisition differ significantly in their treatment of the extent of abstractness in language. Nativist theories emphasize language acquisition as an abstract, innate process, whereas others suggest it is based on prediction-and-error. This study explores the issue of abstractness versus predictability in English and proposes a new probabilistic model for clarifying the issue.

Methods/Materials

It was determined that abstractness could be equivalently studied using the concept of frequency, a term denoting how often a word or phrase occurs in the English language. Four probabilistic models for frequency were studied. Three widely used models (Markov, general construction, and independent probability) represented abstract learning. A new chunk model which determines frequency based on the occurrence of entire sequences of words was developed by the researcher and was used to represent prediction-and-error learning. To compare the theoretical predictions of each model with real-life language processes, 31 children between the ages of 3 and 4 were tested for proficiency of repetition and comprehension using 28 pairs of high and low frequency expressions. Length of the expressions, the grammatical structure of the expressions, and the individual words making up the expressions were controlled variables, and the frequency of the expression was the independent variable studied. The repetition/comprehension accuracies and delay times were recorded. This data was analyzed for statistical significance and was subsequently compared with each of the four probabilistic models individually to determine the most valid model.

Results

A standard t-test conducted on the data from the child study determined that repetition accuracy for the chunk model was the sole significant measure ($t\text{-stat}=2.18$, $p<0.05$).Proficiency of repetition was greater for high-frequency expressions than for their lower frequency counterparts, but comprehension was not affected by frequency.Through a formula of correspondence established by the researcher, the chunk model was deemed 73.5% correspondent with the data from the child study

Conclusions/Discussion

The validity of the chunk-based model showed that language is prediction-and-error based, rather than being completely abstract and suggested more effective ways for teaching methods of English, in accord with the prediction-and-error processes by which children learn.

Summary Statement

This project was an investigation into the mechanisms by which children acquire language; it revealed that prediction-and-error processes play a huge role in language acquisition as opposed to pure abstract learning.

Help Received

Used Stanford University's data base for recruiting children for the study;Dr.Ramscar introduced me to Probabilistic modeling concept and assisted in data analysis