DOCUMENT RESUME

ED 283 374 FL 016 763

AUTHOR	Moore, Michael; Goldstein, Zahava
TITLE	Predicting the Vocabulary of Children from Written or Spoken Texts.
PUB DATE	Aug 86
NOTE	8p.; Paper presented at the Annual Meeting of the American Psychological Association (94th, Washington, DC, August 1986).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Child Language; Foreign Countries; Grade 6; Grade 8; Hebrew; Junior High Schools; Junior High School Students; *Mathematical Models; *Oral Language; Preadolescents; *Predictor Variables; Reliability; Socioeconomic Status; Uncommonly Taught Languages; Validity; *Vocabulary; Word Frequency; *Written Language
IDENTIFIERS	*Israel

ABSTRACT

A study investigated the use of a mathematical model to predict individuals' total active Hebrew vocabulary from samples of their written and spoken language. The model is based on a generalized inverse Gaussian distribution. The subjects were Israeli junior high school students from both high and low socioeconomic groups. Hebrew language samples of at least 2,000 words were obtained from each child. In 60 of 70 cases, the fit between empirical data and the theoretical model was acceptable. Several hypotheses relating to the model's construct and concurrent validity and its reliability and objectivity were tested. In addition, a hypothesis about the significantly richer vocabulary of higher socioeconomic status children was confirmed. The model can serve as a viable basis for further extensive inquiry into the vocabulary of different samples. (Author/MSE)

PREDICTING THE VOCABULARY OF CHILDREN FROM WRITTEN OR SPOKEN TEXTS

Michael Moore and Zahava Goldstein

Technion - Israel Institute of Technology, Haifa, Israel

ABSTRACT

A mathematical model was used to predict the total active Hebrew vocabulary of Israeli children, given a written or spoken Hebrew text sample of at least 2000 words. The model is based on a generalized inverse Gaussian distribution, suggested by Sichel. In 60 out of 70 cases analyzed, the fit between empirical data and the theoretical model was acceptable. Several hypotheses relating to the model's construct and concurrent validity, as well as its reliability and objectivity, were tested. It is concluded that the model proposed can serve as a viable basis for further extensive inquiry into the vocabulary of different samples.

Paper presented at the Annual Meeting of the American Psychological Association (94th, Washington, DC, August 1986).

PREDICTING THE VOCABULARY OF CHILDREN FROM WRITTEN OR SPOKEN TEXTS

Michael Moore and Zahava Goldstein
Technion - Israel Institute of Technology, Haifa, Israel

Language is a set of phonetic and graphic symbols that serves mainly as a means of communication which enables the user to enhance understanding. The vocabulary of an individual, i.e. the number of words one has at one's disposal, is necessarily finite. When speaking or writing, we sample from our personal store according to the subject, the circumstances, education and ability. All of these factors interact in creating the final product which is a spoken or written text with characteristics that can be analyzed - at least partially - by means of objective statistical methods. Even though the length of a given text is the same, it may contain a greater or smaller number of different words. There may be differences in the growth curve - the rate of increase of vocabulary with increase in the sample size. There may be differences in the appearance of the rare and common words, and so on. In the last hundred years several attempts hwe been made by statisticians and mathematicians to fit the best word usage model, the one which will optimally measure the writer's language ability. In this research a mathematical model was applied which had been formulated by Sichel (1973, 1974, 1975). This model is based on a Generalizec Inverse Gaussian Distribution, which includes a Modified Bessel Function of the 2nd kind. The model predicts the total active vocabulary of a subject, given a written or spoken text sample. This model has been chosen because of the excellent fit it has to empirical data in English - better than other existing models.

The presentation deals with two aspects:

- The analysis that tests whether the model has the required standards of validity, reliability, objectivity and utility, when applied to Hebrew. If the answer to these questions is affirmative, then the model can be used as a powerful psychometric tool which measures richness of vocabulary.
- The mathematical analysis of the parameter estimates of the distribution by means of simulation in order to derive new analytic formulas.

To accomplish the above, it was decided to sample the written and spoken language of children in Israeli junior highschool, from high and low socioeconomic background. The main reason for selecting this target population lies in the fact that there is a large amount of literature that relates to the failure of socially deprived children in school due to their so-called language barrier. An additional reason for this particular sample is that even though the Israeli Ministry of Education invests a considerable amount of its resources in advancing socially deprived children, there are relatively few research attempts in Israel that directly analyze the language of these children, or of children in general. Our attempt can provide a valuable contribution, because it is the first in Hebrew to deal with vocabulary as a predicted total of the amount of words known. The reason for selecting children 11 - 13 years old, is mainly technical: in order to carry out the proper analysis, a text of at least 2000 words is needed - an amount which is impossible to obtain in written language with younger children.
In the process of this research the vocabulary of three textbooks and of 67 children was analyzed - 37 children in written language, all from the same 8th grade class, 20 from high and 17 from low socioeconomical levels.

In spoken language 30 sixth graders were sampled - 15 from a well established urban highschool, while the other 15, representing the lower class, were taken from development towns in Israel.

From all these children language samples of at least 2000 words were obtained. Each word was classified as a lexeme, i.e. as a vocabulary item without its grammatical form. The text thus edited entered the computer. The program for each subject counted their word (lexeme) frequency distribution which was used in estimating the parameters that resulted finally in the predicted value of the subject's total active vocabulary. Simultaneously, the program also carried out a chi-square test for goodness of fit that checked the difference between the observed word frequencies and those expected according to the theoretical model.

The results were encouraging. The validity of the model was tested by hypotheses referring to both construct and concurrent validity. The first hypothesis tested the fit to the theoretical model. It was found that out of the 70 texts analyzed (67 children and three textbooks), in 85.7% of the cases the fit was acceptable (the criterion was $p>0.05$). By simulation, as well, a good fit was found between the observed and the expected values of the parameter estimators. Another hypothesis about the significantly richer vocabulary of high SES children as compared to the lower ones, was also confirmed in both written and spoken language. The results were:

Written Language:

High SES children N=20

Mean predicted vocabulary = 7104.35 words

Standard deviation of predicted value = 3857.91 words

Low SES children N=17

Mean preducted vocabuary = 4684.29 words

Standard deviation = 3140.97 words

Spoken Language:

High SES children N=15

Mean predicted vocabulary = 1538.21 words

Standard deviation = 454.85 words

Low SES children N=15

Mean predicted vocabulary = 1085.20 words

Standard deviation = 495.10 words

The hypothesis that written language has a significantly larger vocabulary than spoken language does, was supported by the following results:

Written Language N=37

Mean predicted vocabulary = 5992.43 words

Standard deviation = 3706.06 words

Spoken Language N=30

Mean predicted vocabulary = 1424.80 words

Standard deviation = 827.22 words

A further hypothesis, according to which the vocabulary of a textbook is significantly larger than that of the children studying it, was largely confirmed by the results. it was found that textbooks consisting of selected texts in literature had a predicted value of vocabulary size between 20,000 and 24,000 for the sixth and eighth grades, respectively. A biology textbook for the eighth grade children had a predicted value of 6,000 words. In this subject, unlike in the literature textbooks, there was no siqnificant difference between the vocabulary size of the text and that of the children studying from it.

Concurrent validity was tested by computing the correlation coefficient between the predicted vocaublary size and several independent criteria: (1) a standardized Hebrew vocabulary test (MILTA) (r = 0.46); (2) a standardized Hebrew reading comprehension test (MILTA) (r = 0.44); (3) teacher's estimate of child's language skill (r = 0.54).

The reliability of the model was investigated by means of an internal consistency procedure. The text of each subject was split into two halves according to odd end even lines. Several measures such as (1) the number of different words, (2) the number cf words that appeared once only (hapax legomena), (3) the frequency of the most common word, were counted separately in each half, and a correlation coefficient was computed between the results of the odd and even parts. These coefficients were exceptionally high and ranged between 0.85 and 0.98, according to different measures.

In order to test to what extent the possible different classifications of words as lexemes affect the outcome of the predicted vocabulary size, the texts of six children were lemmatized by two judges whc worked independently of one another. For these six children the process of prediction was carried out twice. It was shown that the model is very stable (has similar predictions) when the vocabulary predicted is in the range of 2000 - 3000 words. For higher estimates the model is more sensitive to different judgments of classification. This aspect needs further investigation.

In spite of the cumbersome technical process which each text has to undergo prior to the use of the computer, the results of this research seem encouraging and open wide prospects for further extensive inquiry into the vocabulary of different groups in the population.

REFERENCES

Sichel, H.S. "On a family of Discrete Distributions Particularly Suited to Represent Long-Tailed Frequency Data," Proceedings of the Third Symposium on Mathematical Statistics, 1971, 51-97.

Sichel, H.S. "On a Distribution Representing Sentence-Length in Written Prose," Journal of the Royal Statistical Society Association, 1974, 25-34.

Sichel, H.S. "On a Distribution Law for Word Frequencies," Journal of the Royal Statistical Society Association, 1975, 542-547.