DOCUMENT RESUME

ED 052 898 RE 003 707

AUTHOR Rudolph, William B.
TITLE Measuring Reading Comprehensibility and Difficulty in Mathematical English Using Relative Sequential Constraint.
PUB DATE Feb 71
NOTE 21p.; Paper presented at the meeting of the American Educational Research Association, New York, N.Y., Feb. 4-7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Conference Reports, Information Theory, *Language Research, *Linguistics, Literature Reviews, *Mathematical Linguistics, *Mathematical Models, Models, Readability, *Reading Comprehension, Reading Research, Set Theory

ABSTRACT

Literature in the area of measurement of comprehensibility is reviewed as it relates to the utilization of mathematical models for English. A number of mathematical language models are presented and explained, and evidence of their usefulness is given where available. In nonmathematical research, studies concerned with measurement of letter redundancy and other textual constraints are considered as they relate to reading comprehension. Finally, relationships between redundancy and learning are discussed. It is suggested that studies be undertaken to further examine relationships between reading comprehension and mathematical English. References and a glossary are included. (MS)

Measuring Reading Comprehensibility and Difficulty in Mathematical English Using Relative Sequential Constraint

William B. Rudolph
Iowa State University

INTRODUCTION

Shannon's (1948) and Wiener's (1948) work on information theory opened new approaches to the study of human behavior. Their work enabled educators and psychologists to initiate information-theory-based studies applicable to the structural analysis of language.

Shannon (1948) defined (relative) redundancy as one minus the relative entropy where the relative entropy is the ratio of the entropy to the largest value it could have while still restricted to the same symbols. Thus the relative redundancy (R) is given by the formula:

$$R = 1 - \frac{H}{H_{nom}}$$

where the entropy $H = \lim_{N \to \infty} F_N$, $F_N = -\sum_{i,j} p(i)\ p_{ij} \log_2 p_{ij}$,

---

Throughout the paper the reader will encounter some unfamiliar words. Definitions of these technical terms appear after the references.

$p(i)$ is the probability of the $(N-1)$ gram i, p_{ij} is the probability of the symbol j given the $(N-1)$ gram i, and the nominal value of the entropy $H_{nom} = \log_2 n$ where n is the number of distinct symbols.

The pattern of language evolvement in textual material is sequential since an order of perception is established. In reading the reader progresses from left to right and, moreover, prior context influences the future appearance of letters. Thus the occurrence of a letter may depend not only on immediately adjacent letters, but constraints may extend over much of the prior context. A tool for measuring these constraints must therefore use the probabilistic statements inherent in evolving sequential data.

An appropriate mathematical model for analyzing data of a sequential nature is a Markov chain with a discrete time parameter (Binder and Wolin, 1964). A characteristic of Markov chains, making them especially appropriate for application to the entropy concept of information theory, is that the probability of occurrence of a state is contingent upon only the immediately preceding state and none before that. For example, the assumption is made that the probability of occurrence of a specific letter (state) depends only on the immediately preceding, say 20, letters and none before those (it is irrelevant to the prediction problem whether the probability of a specific letter or the probability of the state which is induced by the letter is determined). Assuming, for

this example, a 28-letter alphabet (26 letters, end of sentence, and space) there would be $(28)^{20}$ states with an accompanying probability value attached to each. The probability values when substituted in the appropriate formulas from information theory give an estimate of the 21-gram entropy. Since the determination of entropy is a limiting process, the above procedure would be repeated for dependencies extending over spans greater than the 20 preceding letters.

While it would be desirable to successively approximate the entropy by using longer and longer sequences for predicting the occurrence of a letter, the problem becomes insurmountable quickly. For example, with prediction depending on only the immediately two preceding letters there are $(28)^2$ possible states (assuming a 28 letter alphabet). Tabulation of the frequency of occurrence of each of these digrams is possible (Shannon, 1951) and the resulting estimate of entropy follows readily. However to estimate the entropy from sequences of the ten preceding letters is quite another matter. In this case there are $(28)^{10}$ distinct states. In order to estimate the entropy the length of the English passage from which the frequency distribution arouse would be prohibitively long. These examples illustrate the need for an alternate approach to entropy determination.

Garner and Carson (1960) separated redundancy into two parts, the distributional constraint and the sequential constraint. For example, the model for redundancy when only the

N - 1 preceding variables (each letter position is a variable) are considered is:

$$H_{nom} - F_N = (H_{nom} - H_{max}) + (H_{max} - F_N). \qquad [1]$$

In equation 1, H_{nom} is defined as before, H_{max} gives the uncertainty when the symbols are independent but not necessarily equally probable and is defined as $H_{max} = -\sum_i p(i) \log_2 p(i)$, $p(i)$ is the probability of the symbol i, and F_N is the N-gram entropy. The left side of equation 1 approximates the numerator of the Shannon formula for redundancy and, in fact, would be identical if statistical effects did not extend over sequences longer than N - 1 symbols in length, that is, if $F_N = H$. The expression within the first parenthesis on the right side of equation 1 is the distributional constraint. This expression gives the reduction in uncertainty attributable to unequal frequency of occurrence of symbols. To illustrate, if there were no syntax or spelling rules and if all letters occurred with almost the same frequency, then the reduction in uncertainty would be minuscule and prediction of any letter in a sequence would be little better than chance. The expression within the second parenthesis on the right side of equation 1 is the sequential constraint which gives the reduction in uncertainty due to statistical effects extending over sequences of length N - 1.

The distributional constraint is, of course, easily obtainable

and is independent of problems inherent in working with sequential dependencies. However, calculation of a value for the sequential constraint is more difficult since the N-gram entropy term appears. Binder and Wolin (1964) proved that the sequential constraint is equal to the multiple contingent uncertainty. Consequently, the problem of determining the sequential constraint reduces to finding the multiple contingent uncertainty.

A technique suggested by Newman and Gerstman (1952) has been adapted in recent research to the problem of estimating the multiple contingent uncertainty. The procedure consists of calculating the simple contingent uncertainties and summing these to estimate the multiple contingent uncertainty. The specific formulas utilized to estimate the multiple contingent uncertainties and subsequently the relative sequential constraints from the summation of simple contingencies follow.

An estimate C_n of the relative sequential constraint for a sequence of length n is given by:

$$C_n = \frac{\sum_{k=2}^{n} H(1:K)}{H(1)} \qquad [2]$$

In formula 2, $H(1:K) = H(1) - H_K(1)$, where $H(1) = -\sum_i p(i)\log_2 p(i)$, the summation being over the entire alphabet, and $H_K(1)$ is the uncertainty of the letter being predicted when only the $(K-1)$th preceding letter is used in the prediction. Formally, $H_K(1) =$

$-\sum_{ij} p(i)\ p_{ij} \log_2 p_{ij}$, where p_{ij} is the probability of symbol j given that i occurred K-1 letters before it. The indices on this summation sign range over the entire alphabet being considered. The expression appearing in the numerator, $\sum_{K=2}^{n} H(1:K)$, is the summation of the simple contingencies, $H(1) - H_K(1)$, culminating in an estimate of the multiple contingent uncertainty when only the $N-1$ preceding variables in the Markov chain are utilized.

THE PSYCHOMETRICS OF REDUNDANCY MEASUREMENT

The measurement of the letter redundancy of English was initiated by Shannon (1948). Shannon reasoned that the letter redundancy of a passage could be estimated by a letter deletion scheme. For example, if every fourth letter in a passage is deleted and the subject is able to correctly insert into each blank the missing letter then that letter supplied no new information to him. Therefore, the original passage was more than 25% redundant. On the basis of a limited experiment of this type Shannon predicted that the redundancy of English was about 50%.

Chapanis (1954) selected 13 passages of varying content and style, deleted letters from them in numerous patterns and amounts, and asked subjects to reconstruct the original passages. Results indicated that with increasing percentage of deletions

fewer letters of the original textual material could be restored. For example, with 10% of the textual material deleted 80% of the replacements were correct, but with 50% deletion only 10% of the replacements were correct.

The variation in redundancy between the above study and Shannon's (1948) is apparently the result of the supplementary knowledge that the subject has in Shannon's study, that is, knowledge of the amount of textual material to be deleted, deletion patterns, and type of textual material (Chapanis, 1954).

Miller and Friedman (1957) investigated the effect of a particular type and amount of error in passages on subjects' ability to replace accurately letters within those passages. Results indicated that people on the average are unable to replace accurately mutilated textual material in which more than 10% of the characters are missing, thus adding credence to the findings of Chapanis.

Shannon (1951) calculated the N-gram entropy for N = 0, 1, 2, 3, and 5 from tabulated frequencies for letters, digrams, trigrams, and words. To measure constraints extending over longer sequences of letters Shannon established an "information isomorphism" between a passage and its reduced form (a sequence of symbols, each of which indicates the number of guesses taken by the subject to obtain the letter in the original passage). Therefore, if the entropy can be determined for the reduced passage then the entropy for the original passage is uniquely determined and the ensuing redundancies are equal. Long range statistical effects (up to 100 letters) were manifest in

entropy of approximately 1.0 bit (binary digit) per letter with a corresponding redundancy of about 75%.

Hake and Hyman (1953) found that the prediction of any symbol in a binary sequence depended not only on the statistical effects extending over the two antecedent letters but also on what the subject had predicted for the previous two positions and the accuracy of these predictions.

Furthermore, it was noted that subjects tended to perceive structure in a random series of symbols. The investigators concluded that subjects will always perceive a random sequence of events as being more structured than it actually is.

The findings of Hake and Hyman's study cast doubt on the desideratum of using data from student performance to determine N-gram entropy.

Newman and Gerstman (1952) approached the measurement of N-gram entropy from a different viewpoint in an attempt to eliminate the subject as a possible confounding influence. The investigators proposed a new measure, the coefficient of constraint with values in the interval 0 to 1, which gives a relationship between each letter position (1, 2, ..., N - 1) and the N^{th} letter. For example, to determine the coefficient of constraint for letters separated by k (k a non-negative integer) intervening characters, contingency tables would be constructed from the textual material showing the number of times each letter followed every other letter after an interval of distance k. From this table a ratio between the Markovian information (for letters separated by k intervening characters)

and single letter uncertainty could be determined.

The complement of this ratio is the coefficient of constraint. Thus, if letters separated by k intervening characters are independent, the coefficient of constraint has value 0 while if the separated letters are completely dependent the coefficient of constraint has value 1. A combination of the coefficient of constraint values for letters separated by k intervening letters (k = 0, 1, ..., N - 2) leads to an estimate of the N-gram entropy.

If letter positions are considered to be variables, then N - 1 variables enter into the Newman and Gerstman estimation of the N-gram emtropy. However, interactions between these N - 1 variables do not enter the estimation. The insignificant contribution of these interaction terms to the Newman and Gerstman estimate of N-gram entropy is documented in Garner's (1962, pp. 239-242) work.

To test the efficacy of the method, a 10,000 symbol sample of English textual material from the Bible was analyzed. The results indicate that values for the N-gram entropy (N = 1, 2, ..., 10, 100) found by using Newman and Gerstman's method are in agreement with the comparable values obtained by Shannon (1951).

Newman and Waugh (1960) assessed the redundancy and N-gram entropy for three samples of English textual material (taken from the Bible, a work by William James, and the Atlantic Monthly) using an adaptation of the Newman and Gerstman

method. The two measures varied between samples indicating their dependence upon the particular textual material under examination.

Newman and Waugh next examined size of alphabet as a variable in information measurement. Identical Biblical passages in Samoan (16 letter alphabet), English, and Russian (35 letter alphabet) were selected. Less freedom of letter choice was found for Russian than either English or Samoan while English was more restrictive than Samoan. The N-gram entropies over a long passage (N = 12) for each sample were approximately equal. Newman and Waugh suggested that comparisons across languages should be sensitive to variations in textual material.

Paisley (1966) investigated the effects of authorship, time of composition, topic, and structure on letter redundancy of English textual material. Letter redundancy varied with each of th[illegible]ur factors.

Carterette and Jones (1963) measured the redundancy of children's books by an adaptation of the Newman and Gerstman method. Letter redundancy between readers (first, second, third, and fifth grade) decreased with increasing grade level. The Bible is an constrainted as a third grade reader while constraints on a fifth grade reader approach those of an adult book (a work by William James).

A perfect inverse correlation between redundancy and mean word length led the researchers to conclude that constraints are probably heavily determined by size of lexicon. Increase

in constraint was negligible after sequences of more than nine letters.

Jones and Carterette (1963) compared the redundancy of free-reading choices of children at the first, third, and fifth grades with first, third, and fifth grade readers. Free-reading choices were less redundant than the corresponding readers. The investigators concluded that children prefer reading materials which are less redundant than their readers.

Rudolph (1969) investigated whether constraint differs for modern and traditional mathematics books. At a fixed grade level textbooks illustrative of each approach were chosen. The topic was controlled between books, and 20,000 symbol passages were randomly drawn from each of the two textbooks. For each passage alphabet size and constraint were determined. Neither modern nor traditional textbooks consistently had greater sequential constraint although modern books used more symbols in six of the nine comparisons presented. Also alphabet size was directly related to relative sequential constraint, at least when topic was controlled, in seven of the nine comparisons.

Another aspect of the research was to determine whether relative sequential constraint fluctuates between topics within a book. Passages containing 20,000 symbols were randomly selected from each of two topics for each of four textbooks. Results indicated that sequential constraint varies between topics. The implication is that a unique value of constraint for ME, even within a given textbook, is nonexistent.

Another question was whether constraint varies with ascending grade level. To answer this question 20,000 symbol passages were randomly selected from textbooks at different grade levels, but with topic and authorship controlled. Measures of constraint on these passages revealed an inverse relationship between relative sequential constraint and grade level. Thus textual material at the third grade level was more constrained than that at the fourth grade level.

In mathematics textbooks some of the language is concerned with deductive reasoning. The comparison of constraint for this language style and less directive discourse was also investigated. A passage of each language style was selected from each of two mathematics books. Alphabet size and relative sequential constraint were determined for each of the ME passages. Results indicated that the deductive style of presentation was more constrained and had a greater number of symbols than the less directive discourse.

REDUNDANCY AND LEARNING

Miller and Selfridge (1950) examined the relationship between contextual constraint and recall. Passages of various approximations to English of different lengths constructed using words as the basic sampling unit were presented orally to subjects. After each passage a subject wrote down in any order the words in the textual material. Percentage of recall increased with increasing structure and decreased with increasing length of passage.

However, less meaningful 10 and 20 word passages were recalled about as well as textual meterials. With longer passages the relationship between degree of organization (a passage with a high degree of organization approximates textual material while a low degree of organization implies words for the passage were randomly selected) and recall was more evident.

Miller and Selfridge concluded that meaningful materials are more readily learned, not necessarily because the materials are more meaningful, but because short range contextual constraints are retained that facilitate learning.

With an improved design Marks and Jack (1952) replicated the work of Miller and Selfridge. Again degree of organization effected ability to recall but the researchers concluded that a prior relative meaningfulness of the materials and not necessarily short range contextual constraints accounted for the relationship.

Sharp (1958) constructed passages of varying contextual constraint by using modal responses of subjects to stimulus phrases. The mean number of words correctly recalled increased with increasing structure but decreased when meaningful textual material occurred. The author attributed this inconsistency to the method employed in the construction of the textual material (the sample of words found in the passage was selected from the word-distribution generated by the subjects). Increasing structure resulted in a large amount of relearning after a one week interval.

Aborn and Rubenstein (1952) attempted to determine the relationship between passage structure and learning (as measured by the ability of a subject to recall passages of approximately 30 three syllable nonsense words). Learning increased with increasing structure and information learned was constant up to a certain degree of organization but thereafter increasing organization resulted in a smaller amount of information. In addition, the amount of material the subject recalls can be predicted from knowledge of the average information per syllable and the subject's ability as measured on another passage having a different average rate of information.

In a variation of their 1952 study, Rubenstein and Aborn (1954) further investigated the influence of degree of organization on a subject's ability to recall passages of nonsense syllables immediately after learning. In contrast to the earlier study an intensive training session and varying study time were provided each subject. As before, an inverse relationship between degree of organization and information (bits per syllable) recalled was noted. Moreover, this relationship was independent of length of study period. Furthermore, recall scores were different for each degree of organization whereas in the prior study information recalled remained constant for lower degrees of organization.

Miller (1958) built statistical structure into a sequence of letters using a finite state generator operating on four randomly chosen consonants. Two lists, consisting of passages of varying but equivalent length, were constructed from the

resulting subset of well formed strings. Additionally, two other lists consisting of the same population of symbols were constructed randomly with the constraint that passage length between the four lists must be equivalent. A combination consisting of two of the four lists was presented to a subject in a specified order. Each passage within a list was placed on a 3 x 5 inch card and shown to the subject at five second intervals. After all cards in one list were seen the cards were reshuffled at which time the subject was asked to write down as many of the passages in the list as could be remembered. The procedure was repeated for 10 trials of each list. Although the subjects knew nothing of the rules of formation they were better able to reproduce the redundant strings. Apparently subjects are able to recode and group the redundant strings of letters, thereby aiding in the memorization. It was also found that while the amount of material learned increased with structuring the amount of information decreased. Thus redundancy in materials to be learned does not necessarily increase the efficiency of learning.

Ruddell (1965) explored the relationship between the redundancy of syntactical elements in written language and reading comprehension. Redundancy measures were determined for each of two passages of diverse syntactical structure. Scores from 131 fourth grade students on cloze tests over the two passages indicated a positive relationship between redundancy and reading comprehension. Ruddell concluded that reading comprehension is a function of the redundancy of the syntax of

written language.

Five passages over which reading comprehension tests had been administered were analyzed in Rudolph's (1969) study in an effort to investigate the relationship between reading comprehension of ME and relative sequential constraint. Measures of constraint on each of the five passages were determined. The correlation coefficient indicated an inverse relationship between relative sequential constraint and reading comprehension. Thus more constrained textual material seems to result in lower scores on reading comprehension tests, at least for ME. Herein may lie a distinction with OE where a direct relationship exists between reading comprehension and constraint. Possibly topics which have low constraint associated with them might be developed to a greater depth since low constraint is associated with higher scores on reading comprehension tests. That is, detailed discussion of peripheral areas related to topics having low constraint may be beneficial. Furthermore, greater emphasis in teaching should be placed on those topics having high constraint since such topics are associated with lower reading comprehension scores.

Similar studies should be undertaken to corroborate the relationship between reading comprehension and constraint for ME. Rudolph's study suggests that when the topic is controlled alphabet size and relative sequential constraint are directly related. However, the disparity in alphabet size is small. Greater variability in alphabet size with topic controlled may help to isolate the

effect of this variable on relative sequential constraint. Such diversity in alphabet size may be found when comparing ME in books, journals or other works where symbolism is restricted.

REFERENCES

Aborn, M., and Rubenstein, B. Information theory and immediate recall. Journal of experimental psychology, 1952, 44, 260-266.

Binder, A., and Wolin, B. R. Informational models and their uses. Psychometrika, 1964, 29, 29-54.

Carterette, B. C., and Jones, M. H. Redundancy in children's texts. Science, 1963, 140, 1309-1311.

Chapanis, A. The reconstruction of abbreviated printed messages. Journal of experimental psychology, 1954, 48, 496-510.

Garner, W. R., and Carson, D. H. A multivariate solution of the redundancy of printed English. Psychological reports, 1960, 6, 123-141.

Garner, W. R. Uncertainty and structure as psychological concepts. New York: Wiley, 1962.

Hake, H. W., and Hyman, R. Perception of the statistical structure of a random series of binary symbols. Journal of experimental psychology, 1953, 45, 64-74.

Jones, M. H., and Carterette, E. C. Redundancy in children's free-reading choices. Journal of verbal learning and verbal behavior, 1963, 2, 489-493.

Marks, M. R., and Jack, O. Verbal context and memory span for meaningful material. American journal of psychology. 1952, 65, 298-300.

Miller, G. A., and Selfridge, J. A. Verbal context and the recall of meaningful material. American journal of psychology, 1950, 63, 176-185.

Miller, G. A. Free recall of redundant strings of letters. Journal of experimental psychology, 1958, 56, 485-491.

Newman, E. G., and Gerstman, L. J. A new method for analyzing printed English. Journal of experimental psychology, 1952, 44, 114-125.

Newman, E. B., and Waugh, N. The redundancy of texts in three languages. Information and control, 1960, 3, 141-153.

Paisley, W. J. The effect of authorship, topic, structure, and time of composition on letter redundancy in English texts. Journal of verbal learning and verbal behavior, 1966, 5, 28-34.

Rubenstein, H. and Aborn, M. Immediate recall as a function of degree of organization and length of study period. Journal of experimental psychology, 1954, 48, 146-152.

Ruddell, R. B. Reading comprehension and structural redundancy in written material. International reading association conference proceedings, 1965, 10, 308-311.

Rudolph, W. B. Estimates of the relative sequential constraint for selected passages from mathematics books and the relationship of these measures to reading comprehension. Unpublished doctoral dissertation, Purdue University, 1969.

Shannon, C. E. A mathematical theory of communication. Bell system technical journal, 1948, 27, 379-423.

Shannon, C. E. Prediction and entropy of printed English. Bell system technical journal, 1951, 30, 50-64.

Sharp, H. C. Effect of contextual constraint upon recall of verbal passages. American journal of psychology, 1958, 71, 568-572.

Wiener, N. Cybernetics or control and communication in the animal and the machine, New York: Wiley, 1948.

DEFINITIONS

1. Deductive Textual Material. Material which results when an axiomatic system is applied over a sequence of steps leading a person from initial conditions to the conclusion, examples are proofs of theorems, lemmas, etcetera.
2. Entropy (H). The minimum average number of binary digits required to encode each character of textual material, formerly $H = \lim_{N \to \infty} F_N$ where F_N is the N-gram entropy; information; uncertainty.
3. Information. See entropy.
4. Letter Redundancy. Redundancy measurement in which the basic sampling units are letters.
5. Mathematical English (ME). The written language found in mathematics textual materials.
6. Multiple Contingent Uncertainty. The total amount of uncertainty in the criterion variable which can be predicted from simultaneous values of the preceding variables.
7. N-gram Entropy. Information when the $N - 1$ preceding letters are used in predicting the Nth letter of a sequence N letters long.
8. Redundancy. $1 - \frac{H}{H_{max}}$ where H is the entropy and H_{max} is the entropy which would result if all states were independent and equally probable. The redundancy is a measure of the constraint imposed on textual material due to its statistical structure, for example, in English the tendency of H to follow T.

9. Relative Sequential Constraint. A measure of redundancy computed from the summation of contingent uncertainties.
10. Simple Contingent Uncertainty. A measure of the amount of uncertainty reduction due to the contingencies between the initial predictor variable and the criterion variable.
11. Single Letter Uncertainty H(1). Uncertainty when each letter is independent of every other.
12. State. Some specific set of values of all the variables of concern.
13. Uncertainty. See entropy.