An information-theory measurement of CVC trigram meaningfulness*

KENT B. TAYLOR†
The George Washington University, Washington, D.C. 20006

The meaningfulness of nonsense syllables has generally been considered in terms of Ss' ratings and use of associations and pronounceability. The redundancy of nonsense syllables was quantified by means of their component transitional probabilities, using information-theory measurements. These mathematically derived ratings were in agreement with Ss' ratings of association value and pronounceability used by previous investigators to identify the relative meaningfulness of CVC trigrams. It is hypothesized that the redundancy measures, by measuring the amount of structure in trigrams, are indicative of the potentiality for yielding signification meaning in short verbal units.

One of the properties of learning materials deemed most necessary by Ebbinghaus in his classical studies on verbal behavior and memory was that of homogeneity. It soon became evident that his nonsense materials never completely met this criterion and that systematic attempts to order syllables along a dimension of meaningfulness was required for analyzing sources of variance among verbal stimuli. The nonsense syllable, or CVC trigram, has since been the object of numerous studies, each with its unique methodology and share of problems. Scale values for verbal units have been derived from a variety of different operational procedures. Among these intercorrelated operations are the presence or absence of associations, the number of associations elicited during different time intervals, rated associations, familiarity ratings, and pronounceability ratings.

The intent of this study is to measure the meaningfulness of CVC trigrams from an information-theory standpoint and to evaluate this unique approach in the context of existing research employing the aforementioned more traditional approaches. Information theory provides a theoretical framework within which data about interdependencies between letters of English can be evaluated. It provides a yardstick for measuring the structure or amount of sequential constraint occurring in short letter sequences such as nonsense syllables. Information theory is based on probability measures. For this study a word count was undertaken and tables of transitional letter probabilities and conditional probabilities of trigrams were generated by computer to quantify the sequential constraint operating in short letter sequences.

*The author expresses his appreciation to the George Washington University Computer Center for the use of its facilities.

†Please address reprint requests to 1229 Forest Ave., Palo Alto, Calif. 94301.

As Garner (1962) has pointed out, the total uncertainty present in a sequence of N letters may be established by totaling the uncertainty of the first state or letter together with the conditional uncertainty of successive letters, the immediately preceding letter being held constant. For example, the information or uncertainty associated with the sequence -abc-, where - denotes a blank, is derived from the uncertainty of the letter a preceded by a blank, b preceded by an a, and so forth. Probability measures for each of these units are obtained from a matrix of digram-transition probabilities. More inclusive measures are obtained from matrices of trigram probabilities which yield data on the conditional probabilities of three-letter sequences with the preceding first or first two symbols held constant. The redundancy of nonsense syllables is then determined by the difference between the uncertainty of the actual distribution of letter combinations and the total possible uncertainty of trigrams.

The relationship between "information" and "meaning" is not altogether manifest and demands discussion. The role of meaning in human behavior is difficult to define, primarily because of the many connotations associated with the term. However, if one approaches meaning in terms of structure and signification, as Garner (1962) has, the problem becomes easier to deal with. Signification refers to the dictionary type of definition of words or events and the associations connected with them. Structure, on the other hand, refers to the totality of relations between events and can be dealt with in terms of information theory concepts and methods of measurement. For example, a completely random assortment of dots and lines may be said to lack meaning, whereas if the same elements are structured to form a picture, this collection of lines and dots may become meaningful. In this sense meaningfulness is related to the amount of structure, and the quantification of structure is possible through information theory. Information-theory approaches to the probabilities associated with trigrams allow quantification of the degree of structure in trigrams. However, since structural meaning is a prerequisite to signification meaning, the information-theory approach measures also the capacity for signification meaning.

A unique feature of the information-measurement approach to the meaningfulness of verbal units is that, unlike most previous investigations, this method does not involve human Ss' responses or ratings as indices of meaningfulness. By comparing the information-theory ratings with those generated by a variety of established techniques, it is anticipated that the validity of this approach may be ascertained. The use of two information-theory measures, those incorporating digram and trigram conditional probabilities, permits further evaluation of their relative predictive powers.

MATERIALS AND PROCEDURES

The materials used in this study were a sample of 210 trigrams, meaningfulness ratings of these trigrams by other studies, and a 20,000-word sample from which letter transition probabilities were obtained.

As a representative sample of CVC trigrams, 210 trigrams were chosen randomly according to the following restrictions: (1) The letter y was not considered as a vowel, (2) 10 trigrams were chosen for each of the 21 possible initial consonants, and (3) each vowel was represented twice in the sample of 10 trigrams. In order to compare the information-theory approach to trigram meaningfulness with the approaches of previous investigators, ratings of associations, meaningfulness, frequency of occurrence, and pronounceability obtained by various studies were recorded wherever possible for the sample of 210 trigrams. Measures from the following studies were used: (1) Krueger's (1934) percentages of association value, (2) Underwood & Schulz's (1960) measurements of pronunciation difficulty according to scale ratings, and (3) Noble's (1961) measurements of association value (a), rated associations (a′), and scaled meaningfulness (m′) for the 2,100 CVC combinations of the English alphabet. To study the interrelationships among meaningfulness, information-theory values, and values derived from a straightforward frequency count of trigrams in samples of English, trigram frequency counts by

Table 1
Intercorrelations Among Trigram Measures

	A′	A	M′	Krueger	Frequency (Underwood)	Frequency (Mayzner)	Frequency	Digram H	Conditional H
Pronounceability	(116) −.78	(116) −.55	(115) −.78	(97) −.67	(63) −.33	(60) −.33	(66) −.34	(116) .76	(116) .77
A′		(210) .63	(209) .97	(171) .83	(106) .34	(101) .31	(110) .36	(210) −.63	(210) −.70
A			(209) .91	(171) .45	(106) .10	(101) .20	(110) .13	(210) −.58	(210) −.61
M′				(170) .82	(105) .34	(101) .30	(109) .36	(209) −.66	(209) −.73
Krueger					(74) .20	(70) .14	(77) .10	(171) −.55	(171) −.58
Frequency (Underwood)						(89) .94	(94) .95	(106) −.40	(106) −.39
Frequency (Mayzner)							(96) .92	(101) −.35	(101) −.34
Frequency								(110) −.39	(110) −.34
Digram H									(210) .81

Mayzner, Tresselt, & Wolin (1965) and by Underwood & Schulz (1960) were also used along with a count generated in this study.

The data base from which letter counts were derived and probabilities of symbol interdependencies generated was obtained from a sample of 20,000 words comprising two-thirds of a recent issue of a popular newsmagazine. The word count was keypunched for storage on magnetic tape. The language statistics regarding the data base were in close agreement with those of other studies with rank correlations of frequency of individual letter occurrence in this sample and four others all .97 or higher. Product-moment correlations for frequency of occurrence of trigrams in this count and in counts by Underwood & Schulz (1960) and by Mayzner, Tresselt, & Wolin (1965) were .95 and .92, respectively.

The nature of this experiment is computer-oriented in that all data used were either computer-stored or generated, and all computations were carried out by computer. Programs for the analysis were written in FORTRAN IV, G level, and were run on an IBM 360, Model 50. The principal data-manipulation program processed the 20,000 word-data base, eliminating single-letter words, punctuation, and nonalphabetic characters, leaving a single space between all words. It then processed the remaining string of approximately 105,000 characters and blanks, recording each consecutive occurrence of digrams in the form of a 27 by 27 matrix with the states representing each of the 26 letters of the alphabet and a blank. For example, if the sequence CL were found 85 times in the 20,000 words of text, 85 was entered at the intersection of Row C and Column L. This sum matrix was then converted to a stochastic transition matrix by dividing each of the row entries by their respective row sums, yielding digram-transition probabilities for the information-theory approach.

Another program processed the string of 105,000 characters and recorded the frequency of occurrence of trigrams in the form of 27 27 by 27 matrices. For example, if the three-letter sequence CAT appeared 66 times, the number 66 was entered at the intersection of Row A and Column T of the third, or C, matrix. To obtain the probability of occurrence of any trigram, the frequency of occurrence of that trigram was divided by the total frequency of occurrences for all trigrams. To obtain the conditional probability for the last two letters of a trigram given the first letter, the frequency of occurrence for these two letters in the 27 by 27 matrix representing the first letter was divided by the total frequency of occurrence of all final two-letter combinations for that matrix. To obtain the conditional probability of the last letter of a trigram given the first two letters, the frequency of the last two letters in the matrix representing the first letter was divided by the total frequency of occurrence of all final two-letter combinations for the row of the second letter of that matrix.

From the data generated by these programs, two information measures specific to the sample of 210 CVC trigrams were obtained. The first measure made use of only the digram probabilities and represented the amount of information associated with the four digram sequences present in the blank-consonant-vowel-consonant-blank series. The formula may be depicted as: $H_{\text{-abc-}} = \log_2 1/P(\text{probability})_{\text{-a}} + \log_2 1/P_{ab} + \log_2 1/P_{bc} + \log_2 1/P_{c\text{-}}$, where - represents a blank and a, b, and c represent three letters of a trigram. The second information measure takes into consideration three sets of three probabilities associated with the -abc- sequence. The first three probabilities are those of the three combinations of three consecutive symbols: -ab, abc, and bc-. The second three probabilities are those of the conditional probabilities of two consecutive symbols given the immediately preceding one: ab/-, bc/a, and c-/b, where / denotes "given." The final probabilities are those of the conditional probabilities of a symbol given the immediately preceding two: b/-a, c/ab, and -/bc. The formula for the average information of these events is: $H_{\text{-abc-}} = \log_2 1/P_{\text{-ab}} + \log_2 1/P_{abc} + \log_2 1/P_{bc\text{-}} + \log_2 1/P_{ab/\text{-}} + \log_2 1/P_{bc/a} + \log_2 1/P_{c\text{-}/b} + \log_2 1/P_{b/\text{-}a} + \log_2 1/P_{c/ab} + \log_2 1/P_{\text{-}/bc}$.

A final program was written to compute the correlations among the association, meaningfulness, pronounceability, frequency, and the two information measures associated with the sample of 210 CVC trigrams.

RESULTS AND DISCUSSION

The correlations among the ratings of CVCs along several parameters are given in Table 1. The N used in each correlation is given in parenthesis.

It is clear that the trigram frequency counts of Underwood & Schulz (1960) and of Mayzner, Tresselt, & Wolin (1965) and those of the present experiment correlate far more highly among themselves than they do with any other measurements of CVC trigrams. Their low positive correlations with meaningfulness scales fail to support or refute strongly Underwood & Schulz's (1960) conceptualization of the positive relation between trigram frequency and meaningfulness. Their high intercorrelations, however, lend support to the representativeness of their samples and to the reliability of their procedures. In consideration of the three scales of Noble (1961), it should be noted that he recommended the m′ scale derived from a and a′ values as the most appropriate for all indices of associative frequency for CVC material in terms of rationality, reliability, validity, and metrical properties.

The two information scales are highly correlated (r = .81), and their correlations with other scales of CVC meaningfulness are similar, with the conditional probability scale constituting the better

predictor of the two. This is to be expected, since the conditional probability measure makes use of a greater number of transition probabilities. The correlations between association scales and those of pronounceability and information are all negative. Sequences that are difficult to pronounce are more unfamiliar and contain more information and consequently elicit fewer associations. In some instances the information values correlate more highly with some association scales than do other association measures. However, the m′ association measure would seem to be a better measure of the rather elusive concept of meaningfulness.

Information-theory ratings, while not measuring the identical qualities of meaningfulness dealt with in other investigations, can be said to measure a different aspect of meaningfulness, the capacity for signification meaning. Trigrams with very little structure are far less likely to elicit associations or signification than those with more structure. In this context it is significant that a great deal of what we call meaningfulness can be predicted on the basis of information measurement of digram and trigram letter dependencies alone, without recourse to ratings or responses by human Ss.

REFERENCES

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

KRUEGER, W. C. F. The relative difficulty of nonsense syllables. Journal of Experimental Psychology, 1934, 17, 145-153.

MAYZNER, M. S., TRESSELT, M. E., & WOLIN, B. R. Tables of frequency counts for various word-length and letter position combinations. Psychonomic Monograph Supplements, Vol. 1, No. 3, 1965.

NOBLE, C. E. Measurements of association value (a), rated associations (a′), and scaled meaningfulness (m′) for 2100 CVC combinations of the English alphabet. Psychological Reports, 1961, 8, 487-521.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* Chicago: Lippincott, 1960.

Transgression, fate control, and compliant behavior

DAVID L. McMILLEN
Mississippi State University, State College, Miss. 39762

A hypothesis was tested that increased "fate control" will lead to increased compliance. Fate control was manipulated by inducing S to tell a lie about information he had received and then make use of the information. Other Ss were induced to lie but could not make use of the information. Control Ss who told no lie were included in the design. Significant compliance was observed in the "fate control" group only.

Several studies have demonstrated that compliant behavior increases following transgression. Different kinds of transgression have been employed, including destroying a machine (Brock & Becker, 1966; Wallace & Sadalla, 1966), upsetting the order of a graduate student's index cards (Freedman, Wallington, & Bless, 1967), costing another person green stamps (Berscheid & Walster, 1967), lying (Freedman, Wallington, & Bless, 1967), and administration of electric shocks (Carlsmith & Gross, 1969), and the results have been essentially the same. However, the theoretical explanations of the data are not uniform. Some investigators (e.g., Freedman et al, Carlsmith & Gross) employ a "guilt" interpretation to these findings, while others (e.g., Brock, 1969) state that such an interpretation is not warranted on the basis of the existing data. Brock expresses concern over the imprecise conceptual status of guilt and the lack of manipulation checks. In addition, he suggests that the body of data can be explained in terms of fate control and maintenance of social consistency. According to Brock: "An individual who has affected the fate of another person in a certain magnitude will repeat that magnitude of control over the other person (or a person in a similar role) if an opportunity to do so presents itself [p. 143]."

The present study was designed to test the viability of Brock's fate-control hypothesis. The basic design of this experiment was similar to the one employed by Freedman et al (1967). S was induced to tell a lie to E and later given an opportunity to comply with a request made by E. However, in this experiment half of the Ss gained fate control over E by lying; the other half (who also lied) did not. The design also consisted of two control groups composed of Ss who did not lie to E. If Brock's hypothesis is correct, then the "fate control" group should exhibit greater compliance than the "no-fate control" group, even though Ss in both groups tell a lie.

SUBJECTS

Thirty-two males were recruited from introductory psychology classes during the summer session at Mississippi State University. At the time they signed up, Ss were informed that they would be participating in a study of the psychology of education. Eight Ss were assigned randomly to each of four groups.

The Ss were scheduled in pairs. When they arrived, they were seated in a waiting room and told that the previous Ss had not completed the experiment as yet. E left the room; 1 min later the confederate entered saying that he was looking for a book he had forgotten when he was in an experiment earlier in the day. The confederate engaged the Ss in brief informal discussion. During the course of this discussion, the confederate furnished that half of the Ss with information about the experiment. Specifically, he said that he had taken a multiple-choice psychology test. He further stated that after he had completed the test, E had shown him the answer key and he had been surprised to learn that the majority of the correct answers were "B." In the control condition the confederate did not discuss the experiment. If S asked about the experiment, the confederate claimed he