DOCUMENT RESUME

ED 369 044 CS 011 662

AUTHOR	McEneaney, John E.
TITLE	Sources of Redundancy in Printed English.
PUB DATE	[94]
NOTE	9p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cloze Procedure; English; Higher Education; *Information Processing; *Models; Prediction; Reading Research; *Redundancy; Undergraduate Students
IDENTIFIERS	*Text Factors

ABSTRACT

A study examined the relative contributions of semantic and non-semantic sources of redundancy in printed English, which play a central role in information processing models. Subjects, 40 undergraduate college students, were divided into two groups. One group predicted missing characters using English text, and the second group was required to predict missing characters from English-like pseudo-text generated from the original text using a procedure that retains the statistical characteristics of the original text while draining it of semantic content. Results indicated significant differences in the performance of the two groups. Overall redundancy of the text was approximately 50% and non-semantic sources (orthography and syntax) accounted for approximately 70% of this total. (Contains 8 references.) (RS)

Sources of redundancy in printed English.

John E. McEneaney, Ph.D.
Indiana University South Bend
South Bend, IN
JMcEnean@Indiana.edu

Abstract

This study examines the relative contributions of semantic and non-semantic sources of redundancy in printed English. An experiment was carried out with forty undergraduate college students involving the prediction of upcoming characters in a modified cloze procedure. Subjects were divided into two groups. One group predicted missing characters using English text. The second group was required to predict missing characters from English-like pseudo-text generated from the original text using a procedure (Shannon, 1949, p. 44) that retains the statistical characteristics of the original text while draining it of semantic content. As expected, significant differences in the performance of the two groups were found. On the basis of these differences, it was determined that the overall redundancy of the text was approximately 50% and that non-semantic sources (orthography and syntax) accounted for approximately 70% of this total.

Sources of redundancy in printed English.

John E. McEneaney, Ph.D.
Indiana University South Bend
South Bend, IN
JMcEnean@Indiana.edu

The concept of redundancy has played a central role in information processing models of reading for many years. Briefly, redundancy refers to the "predictability" of language. This predictability arises from both non-semantic linguistic conventions and from constraints imposed by meaning. For the most part, however, the concept of redundancy has either been treated informally (e.g. Smith, 1971; Goodman, 1984) or as a single global measure (e.g. Edwards, 1964; Pierce, 1980; Shannon, 1951). This study defines an articulated theory of redundancy and presents preliminary findings that identify the relative contributions semantic and non-semantic (orthographic and syntactic) sources make to the redundancy of printed English.

The present work is based on a method (Shannon, 1951, p. 44) that provides for the quantitative assessment of the redundancy of English text. In this procedure, subjects are required to guess missing characters in text passages. The first part of this study is concerned with the determination of the overall redundancy for a number of randomly selected prose passages. This represents a replication of prior work by Shannon (1951). The second part of this study is concerned with the measurement of the specifically non-semantic (i.e. orthographic and syntactic) contribution to the redundancy of these passages. In order to assess non-semantic redundancy a special pseudo-text, which is a derivative of the original prose, was generated.

Using a method described by Shannon (1949), a list of English-like nonsense words was generated. These nonsense words were then substituted for all of the lexical words (Fries, 1953) in the original English passages. Words were substituted in such a manner as to ensure syntactic agreement with the original text (e.g. inflectional endings of verbs and adverbs were retained). The resulting pseudo-text is both orthographically (i.e. statistically), and syntactically "equivalent" to the original English text. The pseudo-text is, however, devoid of semantic content. Subjects, therefore, who are asked to predict missing letters must do so solely on the basis of orthographic and syntactic constraints. The value for redundancy so determined is, therefore, non-semantic. Given an assumption of simple additivity (that redundancy as a whole is a simple sum of semantic and non-semantic contributions), a value for semantic redundancy can be inferred.

Method

Subjects included 40 undergraduate college students with verbal Scholastic Aptitude Test (SAT) scores above 450 (in order to assure competence in reading). Subjects were randomly assigned to two groups composed of equal numbers of males and females. No significant difference was found between the verbal SAT scores for the two groups . Subjects were asked to predict every fifth character in three 100-character text passages. Group 1 predicted randomly selected English text passages from a novel, The Midwich Cuckoos (Wyndham, 1957). Group 2, however, predicted characters in pseudo-text passages prepared in the manner described above. All testing took place in a computer laboratory on Apple IIe microcomputers. When a subject arrived for testing s/he was taken to a computer and introduced to the computer-managed experimental task. The experimental task was explained to each subject by an assistant who then observed subsequent trial sessions in order to assure that each subject understood how to operate the computer and carry out the experimental task.

After reading through an introduction on the computer, subjects were provided with an opportunity to read them again or go on to a trial session. The trial session consisted of a 50-character English or pseudo-text passage. This passage was presented at the top of the screen. Initially, only the first four characters and a dash, representing the character to be guessed, were presented. As the subject progressed, however, more of the text was presented. At the bottom of the screen was a list of all possible character choices (26 letters and "/" to represent the space between words). When a subject guessed a character it was removed from the list of possible choices. If a subject chose an unacceptable character not on the list, s/he was informed of the error by the computer and instructed to guess again. Errors such as these were not included in the data used in this study. When the subject guessed a letter correctly, four more characters of text were provided and the choice list at the bottom of the screen was reset to include all 27 characters.

When the trial session was completed, the subject was given the opportunity of reading the introduction again or going on to the experimental sessions. Procedures in the experimental sessions were exactly as have been described for the trial session. As the subject progressed through the experimental sessions the computer recorded when the correct character was guessed at each position for the passages. When the subject completed the third experimental session, these data were stored in a file on the disk. Group files were generated by compiling individual files.

Results

Following the compilation of frequency tables for each group, average entropy values across all 20 positions were calculated. In addition, the total number of guesses required for each subject to complete the experimental sessions were also recorded and analyzed. The English text group generated an entropy of approximately 2.18 bits/symbol. The pseudo-text group generated a mean entropy of 2.94 bits/symbol. As expected, this difference was significant, $t(38) = 2.82$, $p < .05$. The mean number of guesses required by English text subjects was 228; for the pseudo-text group the mean total number of guesses was 294. As before, expected differences were apparent ($t(38) = 5.23$, $p < .05$).

An analysis was carried out to determine whether verbal SAT scores correlated significantly with total number of guesses required. No significant correlation was found, suggesting that the ability to anticipate text is well established in subjects whose verbal SAT scores are greater than 450. No significant sex differences were noted for any variable.

On the basis of the prediction distribution provided by group 1, and Shannon's (1951) upper-bound computational procedure an average lower-bound redundancy value of .54 was calculated (where Redundancy $= 1 - [\ (- \Sigma_{i=1}{}^{27}\ (\ q_i{}^n \log q_i{}^n\ ))/\ 4.75])$, where q refers to the relative frequency recorded for correct guesses on the i th guess with an n-1 gram text provided to the subject. This value is slightly lower than the average value indicated for Shannon's (1951) samples (.62), but this can reasonably be attributed to the fact that the estimate represents a lower-bound measure. The present research, therefore, appears to support Shannon's estimate.

Using the data from group 2, and the same computational procedure described above, a non-semantic redundancy of .38 was calculated. Based on these figures, the relative contribution non-semantic sources make to redundancy as a whole is given by: .38/.54, or .70. This result can be interpreted as an indication that 70% of the redundancy of the English text is attributable to non-semantic sources. Given the assumption of simple additivity (Total redundancy = a simple sum of semantic and non-semantic sources), a semantic contribution of 30% can be inferred.

Discussion

It appears, on the basis of these results, that non-semantic constraints having to do with syntax and orthography are more powerful determinants of the letter-by-letter predictability of text than are semantic constraints. It is certainly true that the task employed in this study differs substantially from cloze tasks typically used in educational assessment. Nevertheless, the remarkable performance of subjects predicting missing letters from pseudo-text cannot reasonably be dismissed as unimportant either from a theoretical or an applied educational perspective. Although it is true that letter-by-letter (as opposed to word-by-word) prediction dramatically simplifies the predictive task by reducing the range of possible responses, the task was equally simplified for both groups of subjects and word-by-word prediction would have excluded the application of orthographic knowledge in the task.

It should be noted that the actual redundancy of individual samples of printed English is subject to variability dependent upon the kind of text being examined. Prose was chosen for the present study on the assumption that it represents the "middle ground." Poetry could yield a lower non-semantic contribution. Legal text, on the other hand, would likely result in a larger non-semantic contribution. It is reasonable to expect that the ratio of semantic and non-semantic contributions to redundancy will vary continuously between two theoretically limiting values. What these two limiting values are is a question of considerable theoretical interest which remains unanswered. In addition, the relation this ratio has to certain cognitive processes involved in reading (comprehension, recall, inference, etc.) is worthy of further exploration.

References

Edwards, E. (1964). Information transmission. London: Chapman and Hall.

Fries, C. (1952). The Structure of English New York: Harcourt, Brace and Company.

Goodman, K. (1985). Unity in reading. In Harry Singer and Robert Ruddell (Eds.), Theoretical models and processes of reading (3rd. Ed.). Newark, DE: International Reading Association.

Pierce, J. (1980). An introduction to information theory. New York: Dover.

Shannon, C. (1951). Prediction and entropy of printed English. Bell System Technical Journal, 30, 50-64.

Shannon, C., & Weaver, W. (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Smith, F. (1971). Understanding reading. New York: Holt, Rinehart, & Winston.

Wyndham, J. (1957). The Midwich Cuckoos. New York: Penguin Books.