facilitated more by long PFIs after errors than by long PFIs after correct responses. As was mentioned previously, Bourne et al found that any increase in PFI, regardless of whether it followed a correct or incorrect response, facilitated performance, while Wells and Laestadius found that long PFIs after correct responses facilitated performance more than long PFIs after incorrect responses. The discrepancy between the E-controlled and S-controlled studies may be a result of the lengths of the PFIs employed. In the S-controlled studies Ss take an average of 2 to 3 sec after correct responses and an average of 4 to 5 sec after errors. If intervals of these lengths were used in an E-controlled study, then the Ss might do better when given the long interval after errors.

In the latency analysis, as in the PFI analysis, the only reliable effects that appear consistently across instruction groups were those of trial block and type of response; latencies were longer during the first half of the session and on error trials. Assuming that during the latency period the S is trying to decide upon the appropriate classification for the stimulus pattern, one possible explanation for the shorter latencies on the correct response trials and during the second half of the session is that the S is more likely to know the correct response to the patterns appearing on these trials and, therefore, does not have to spend time deciding what

type of response to make. This explanation is, of course, based on the assumption that some stimulus patterns are learned before others, and there is some evidence which suggests that this is the case (e.g., King, 1966).

REFERENCES

- BOURNE, L. E., JR., & BUNDERSON, C. V. Effects of delay of information feedback and length of postfeedback interval on concept identification. *Journal of Experimental Psychology*, 1963, 65, 1-5.
- BOURNE, L. E., JR., DODD, D. H., GUY, D. E., & JUSTESEN, D. R. Response-contingent intertrial intervals in concept identification. *Journal of Experimental Psychology*, 1968, 76, 601-608.
- BOURNE, L. E., JR., & GUY, D. E. Learning conceptual rules. II: The role of positive and negative instances. *Journal of Experimental Psychology*, 1968, 77, 488-494.
- CONANT, M. B., & TRABASSO, T. Conjunctive and disjunctive concept formation under equal-information conditions. *Journal of Experimental Psychology*, 1964, 67, 250-255.
- DENNEY, N. W. The effect of varying percentage of positive and negative instances and instructions on conjunctive and disjunctive learning. *Psychonomic Science*, 1969, 17, 193-194.
- KING, W. L. A developmental study of rule learning. *Journal of Experimental Child Psychology*, 1966, 4, 217-221.
- LEVINE, M. Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology*, 1966, 71, 331-338.
- WELLS, H., & LAESTADIUS, J. Error-contingent and self-paced post-IF intervals in concept learning. Paper presented at the meeting of the Western Psychological Association, 1968.

A Markovian investigation of English language redundancy

KENT B. TAYLOR†

The George Washington University, Washington, D.C. 20006

To test the assumptions of ergodicity and regularity underlying the application of information-theory measures to language-redundancy studies, digram letter transition probabilities were represented in the form of a Markov source. The source was determined to be ergodic and regular, and sequential letter dependencies were found to be relatively short run. Trigram probabilities predicted by the Markov model were consistently slightly higher than empirical probabilities, demonstrating limitation of such an approach to statistical language analysis.

A direct measurement of the redundancy of printed English is nearly impossible since the number of possible letter sequences is prohibitively large for

any direct count. For this reason, redundancy has been investigated through methods of estimating the constraint operating on single letters at the end of long sequences of letters, yielding a measure of average redundancy. The two most widely used methods are Shannon's (1951) "guessing-game" technique, in which the last letter of a sequence is produced from preceding letters, and the multivariate-analysis approach (e.g.,

Garner, 1962) to terms involved in prediction of letters at various positions in sequences.

The present study is intended to demonstrate a Markovian approach to the analysis of the statistical properties of English language structure and to cite the implications of this new approach in terms of current research in language redundancy. While information-theory measures have been extensively used to identify the redundancy of language, the basic assumptions underlying the use of such measures, namely the qualities of ergodicity and regularity, have not been demonstrated. The information or entropy in nonergodic or nonregular sources is not determinable since statistical equilibrium is never reached.

Among the possible discrete Markov processes there exists a group with special properties of significance in communication theory. This special class of regular ergodic sources has the qualities that some power of the exponentiated matrix has only positive components, and there exists a state, or limit, from which further exponentiations will not change the matrix. Such ergodic sources reach a point of statistical equilibrium at which point the information of the process is defined. A Markovian approach permits validation of the crucial assumptions of ergodicity and regularity and yields additional data about the nature of sequential constraint operating along letter sequences.

Markov processes possess other mathematical properties that enhance their use to the mathematical language theorist. Among these are their ability to determine the probability of being in certain states beyond the succeeding one. Given the transition matrix P , and given a starting state i , it is possible to determine the probability that after N steps the process will be in state j . This probability is denoted by

$$P_{ij}^{(n)}$$

If one were to view the states of a Markov process defined by digram probabilities as letters of the alphabet and a blank, thus forming a 27×27 matrix, it would be possible to determine the probability of a given letter a being followed by a letter b after N transitions or N subsequent letters. This calculation can be accomplished by raising the transition matrix to the N th power and reading the probability at the intersection of Row a and Column b . Theoretically, by applying such methods with the square of a 27×27 digram transition matrix, it should be possible to describe the probabilities of occurrence of trigrams. That this does not occur,

*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.

however, demonstrates a limitation in defining the statistical properties of English in terms of digram or Markovian characteristics.

Another effect of employing a Markovian representation of English language structure is that the number of exponentiations required for equilibrium to be attained yields knowledge about the extent of sequential dependencies along letters. If equilibrium were rapidly achieved, for example, after two exponentiations, this would indicate that the probability of any letter following any other after two interceding letters was a function only of the probability of that letter's independent occurrence. That is, such rapidly achieved equilibrium would indicate that sequential dependencies are rather short run and that those studies whose data base included only short sequences might be measuring redundancy more accurately than their lower bound cautions would have one believe. On the other hand, if many exponentiations are required to reach equilibrium, sequential interdependencies along letters would be more extensive, and more comprehensive sampling would be indicated.

MATERIALS AND PROCEDURES

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 news magazine. The word count was keypunched for subsequent storage on magnetic tape. The language statistics regarding the data base were in close agreement with those of other studies. Rank correlations for frequency of occurrences of individual letters in this sample and of four others (e.g., Lysing, 1936; Pratt, 1939; Attneave, 1953; Encyclopedia Americana, 1956) were all .97 or higher. Product-moment correlations for frequency of occurrences of three-letter sequences in the present 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 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 blank between all words. It then processed the remaining string of

approximately 105,000 letters and blanks, recording each consecutive occurrence of digrams in the form of a 27 x 27 matrix with the states representing each of the 26 letters of the alphabet and a blank. For example, if the sequence "cl" was 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 representing a Markov source by dividing each of the row entries by their respective row sums. The transition matrix was then exponentiated by a factor of 50 in order to determine if the Markov source representation was ergodic and regular and to investigate its equilibrium properties.

Other programs were written to tabulate the probability of trigram occurrences and to determine the entropy of the Markov source. The average information per symbol was derived by obtaining the information average taken over all states of the equilibrium vector and weighted according to their probability of occurrence according to the following formula: $H = P_i \log_2 1/P_i$.

RESULTS AND DISCUSSION

Examination of the exponentiations of the original digram transition matrix reveals that the Markov process represented by the matrix is regular and ergodic. Therefore, information theory measurements are appropriate for language-redundancy estimates.

In the equilibrium matrix the 27 column values associated with each given state are identical. These values demonstrate that at equilibrium the probability of going from any state i to any state j is the probability of the independent occurrence of state j . The 27 row values associated with each of the 26 letters and a blank are identical with those values obtained by summing the occurrences of each symbol in the sample and dividing this sum by the sum of occurrences for all symbols. The information in this source, measured from the equilibrium vector, is 4.12 bits. This is in accordance with the measure of the actual uncertainty of letters and a space of 4.10 bits and 4.08 bits given by Garner (1962) and by Newman & Gerstman (1952), respectively.

Examination of the exponentiations also reveals that equilibrium was reached quite rapidly, with a quite close approximation to stability appearing at the fourth power of the original matrix and near perfect equilibrium appearing by the sixth power. This may be interpreted as denoting that, given an initial state or symbol, i , the probability of any other symbol occurring after six or more transitions, or six or more

symbols away from i , is the probability of that letter's independent occurrence. In other words, digram dependencies regarding the initial letter do not hold over more than six consecutive letters. This finding, that long context is not highly important, indicates that the redundancy figures obtained by studies using empirical counts of moderate sequence length are not necessarily lower-bound estimates.

The essential limitation in the use of Markovian procedures to study the statistical structure of printed English is best illustrated by comparison of actual trigram probabilities derived from the count and those trigram probabilities predicted by the original digram transition matrix raised to the second power. Although there is no method for directly measuring the degree of similarity between the predicted and empirical probabilities, impressionistically, the predicted probabilities are consistently slightly higher. The Markovian-predicted probabilities are computed by taking into consideration the probabilities associated with all paths from the initial symbol to all possible intermediate symbols and from these to the final symbol. When trigrams are considered as units of three instead of two sets of length two, far fewer paths are actually possible. This overestimation demonstrates the limitation in the predictive power of the Markov model. It is a productive approach but limited by the fact that sequential letter dependencies in printed English exist to some extent over longer sequences than length two.

REFERENCES

- ATTNEAVE, F. Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, 1953, 46, 81-86.
- GARNER, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.
- LYSING, H. *Secret writing: An introduction to cryptograms, ciphers, and codes*. New York: D. Kemp, 1936.
- MAYZNER, M. S., TRESSELT, M. E., & WOLIN, B. R. Tables of frequency counts for various word-length and letter position combinations. *Psychonomic Monograph Supplements*, 1965, 1(3, Whole No. 3).
- NEWMAN, E. B., & GERSTMAN, L. S. A new method for analyzing printed English. *Journal of Experimental Psychology*, 1952, 44, 114-125.
- PRATT, F. *Secret and urgent*. Indianapolis: Bobbs-Merrill, 1939.
- SHANNON, C. E. Prediction and entropy of printed English. *Bell System Technical Journal*, 1951, 30, 50-64.
- UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning*. Chicago: Lippincott, 1960.