JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 5, 441-446 (1966)

The Information Value of Initial Letters in the Identification of Words

A. C. BROERSE* AND E. J. ZWAAN†

Psychological Laboratory, University of Utrecht, The Netherlands

In most experiments dealing with the relative effectiveness of different word parts in word identification, the greater importance of the word beginning has been ascribed to the sequential order of speech. However, differences in the amount of information must also be taken into account: initial letters contain more information than final letters. In order to determine whether both factors have an effect, an experiment was carried out in which 48 *S*s had to guess Dutch 7-letter nouns from a varying number of letters which constituted either the initial or the final word part. For these nouns as a group, beginnings and endings carried equal amounts of information.

The results indicated that both information and serial order in speech were effective. The time required for identification was dependent on the amount of information of the n-gram presented. The *S*s also enumerated more 7-letter nouns if the initial letters were available, and as a result identification took less time. In addition, the enumerated nouns were found to be relatively frequent words, and speed of solution was directly related to frequency of occurrence in the language.

Tachistoscopically presented words yield better recognition in the field at the right side of the fixation point than in the left peripheral field, as is shown in several experiments in which the recognition threshold for words was investigated (Mishkin and Forgays, 1952; Forgays, 1953). Melville (1957) showed that this difference in recognition time is larger for 7-letter words than for 3-letter words. In these experiments, the beginning of the words on the right side of the fixation point is more central in the field of vision than the initial letters of the words on the left side, and the longer the words, the larger this difference. These results indicate the relatively larger importance of the *beginning* of the word. The same conclusion was reached by Bruner and O'Dowd (1958) who showed that tachistoscopic recognition of words was faster when the final letters were reversed than when the initial letters were reversed.

This difference might have two reasons: (a) habits in speech, favoring the beginning as a starting point for producing and recognizing words, and (b) the differential amount of information in initial and final letters. Miller and Friedman (1957), in their explanation of the same asymmetrical effect for text passages, stress the former point. In word recognition, however, the amount of information has to be considered as well (Aborn and Rubenstein, 1960). The information, depending on the number of alternatives and the probabilities of these alternatives, can be measured in two ways: (a) the number of different words in which an n-gram (i.e., n sequential letters) occurs, and (b) the relative frequency of occurrence of the n-gram in the language. The correlation between these two measures is high, and both can be used as information indicators for the n-grams in relation to words.

In the English language the n-grams at

* Present address: Psychological Laboratory, University of Amsterdam, Keizersgracht 611, Amsterdam, The Netherlands.

† Present address: Psychiatric University Clinic, Research Department, Nic. Beetsstraat 24, Utrecht, The Netherlands.

the end of words are more frequent than the n-grams at the beginning, i.e., contain less information (Miller, 1951; Aborn and Rubenstein, 1960). In order to investigate whether the same holds for the Dutch language, Van Schooneveld's digram frequency data (Report RVO/TNO PL 216) were used, which are based on a text consisting of 117,000 letters, the space between words (*) being counted as a letter. The conditional uncertainty of the first letters (given *) is:

$$p_{(\mathrm{digr}\ *1)}|p_{(*)} = 4.075 \text{ bits of letter.}$$

The conditional uncertainty of the last letters (given *) is:

$$p_{(\mathrm{digr}\ 1*)}|p_{(*)} = 3.35 \text{ bits per letter.}$$

The mean conditional uncertainty of second letters in digrams is:

$$p_{(\mathrm{digr} 1_1 1_2)}|p_{(1_1)} = 3.07 \text{ bits per letter.}$$

Initial and final letters in a word both contain more information than the other letters, but, as in English, compared to each other, the final letters are more redundant.

Verbal Frequency and Word Recognition.

Under impoverished stimulus conditions, words that occur relatively often in the language are recognized faster than less frequent words (Howes and Solomon, 1951; Solomon and Howes, 1951; Postman and Schneider, 1951; Solomon and Postman, 1952; Postman and Rosenzweig, 1956; Riegel and Riegel, 1961). On the other hand, redundant prefixes and suffixes delay word recognition, because they tend to increase the length of a word without giving much additional information (Riegel and Riegel, 1961). Recognition of a word is, therefore, expected to be easier the less frequent (the more informative) the parts of it that are presented.

In the experiments dealing with the importance of different word parts for word recognition, the frequencies of the parts have not been taken into account. Oléron (1961) used nine French 9-letter words in an experiment resembling Bruner and O'Dowd's. He too presented his *S*s with partially anagrammed words, but they were allowed 5 min per word to find the solution. The number of words solved was not different for the cases in which the initial letters and in which the final letters were in the correct order.

Oléron questions whether the factors influencing the solution process for anagrams are the same as those which affect the perception of words. In solving anagrams the *S* has sufficient time to try systematically all cues available. This is not the case when he has to perceive words. Oléron then suggests a second explanation for the equivalence of beginnings and endings. He considers the possibility that, for the French language, word endings are at least as informative as the beginnings of words. Both suggested explanations are again considered as hypotheses in a later article by Oléron and Danset (1963). In order to decide between them, the same nine French words were presented tachistoscopically, one-third of the word (beginning, middle or end) distinctly and two-thirds in a blurred fashion. Recognition was easiest when the initial letters were presented clearly. The beginning of the word seemed to be more important for word recognition than the other parts of the word.

Discussing their results, Oléron and Danset suggest the sequential order of speaking and hearing as a possible reason for the superiority of the beginning. They also mention the probability with which word parts occur in the language, but they remain inclined to credit word endings with a large amount of information and drop the issue as being insoluble for the present. Pratt (1939) and Sacco (1951), however, have published tables containing n-gram frequencies for the French language. According to these lists, final digram frequency is higher than initial digram frequency for 8 out of the 9 words used. Moreover, 5 out of the 9 final trigrams versus 0 out of the 9 initial trigrams occur in the list in which Pratt and Sacco compiled the 24 most frequent trigrams in the French language. On the basis of these data, there is serious reason to doubt the informational equivalence of initial and final letters, at least for the nine words used by Oléron and Danset. This points to the necessity of controlling frequency of word parts in experiments of this kind.

Method

In order to investigate whether, apart from the larger amount of information, initial letters contribute more to the identification of words, the following experiment was carried out.[1]

Material. All 7-letter nouns were selected from a Dutch word-frequency list (Linschoten, 1963). For these words the initial and final tri-, tetra- and pentagram frequencies were defined, the space being included as a letter. The n-gram frequencies based on a total of 450,000 letters from ten Dutch texts were available up to pentagram level. For each noun the frequency of the beginning and ending have been compared. The results are given in Table 1. Again, for this special group of nouns, it turns out that the initial n-grams are less frequent than the final n-grams.

Out of these nouns two groups of 12 nouns each were selected with initial and final n-gram frequencies balanced, i.e., for tri-, tetra- as well as pentagram level the frequency distributions of initial and final n-grams within a group of 12 nouns were the same. Word frequencies ranged from 1 to 67 per million words.

Procedure. On a card only the first or the last two letters were given, while the letters to be guessed were indicated by dots. The *S* was to guess which 7-letter noun was the intended one. If he did not succeed in finding this word after 1 min, he was shown another card with the next letter added. If he did not find the solution during the second minute, he was given the fourth letter, and so on. Any 7-letter noun based on the given n-gram mentioned by the *S* was recorded. When the *S* hit the noun intended, the time required for solving was noted and the next noun was presented in the form of two letters and five dots.

Each *S* was given 12 words to solve, six with initial letters and six with final letters. A group of 48 *S*s, subdivided into groups A, B, C, and D, took part in the experiment. One set of 12 nouns was used with Groups A and B, and another set of 12 nouns with Groups C and D. The presentation was so balanced that half the items in each set were presented with the initial letter and half the items with the final letters.

Table 1
Comparison of n-Gram Frequencies of Beginnings and Endings for 1050 Dutch 7-Letter Nouns

	Ending more frequent	Beginning more frequent	p (bin.)
Trigram	613	416	$< .0001$
Tetragram	594	432	$< .0001$
Pentagram	405	376	.31

Results

Forty-one of the 48 *S*s needed less time to identify the words with the initial letters than the words with the final letters (bin. $p < .001$). Before it is concluded that the verbal habit of identifying words from the beginning was responsible for this result, one should consider the following. The selection of the 24 nouns took place on the basis of n-gram frequencies in running texts. Possibly the n-gram frequencies for 7-letter nouns are different. If, for instance, these nouns as a special group often share the same ending, but rarely an identical beginning, the *S*s would be able to mention more nouns starting from the end and it might then take more time for the intended noun to turn up. Analysis showed that more nouns were indeed possible for a given ending than a given beginning, while the frequency distributions of both word groups did not differ significantly, as tested by Kolmogorov-Smirnov's two-sample test (Siegel, 1956). This could have postponed identification when the final n-grams were given.

A better measure is, therefore, provided by the total number of nouns mentioned for a given beginning as compared with a given ending. Table 2 shows the mean number of words mentioned by the *S*s per given n-gram in the first, second and third minute. In addition, the mean number of possible nouns

[1] The authors are much obliged to Mr. A. X. van Naerssen for his help in carrying out this experiment.

at the corresponding n-gram level is indicated, corrected for words solved. The correction takes account of *S*'s inability to mention more possible nouns on the basis of the n-gram in question, once he has found the intended one. For the minute in which the solution was found, half the number of possible nouns was subtracted, for the next minute(s) the total number was subtracted.

In spite of the larger number of nouns possible on the basis of final n-grams, more words were enumerated when the initial letters were given. This result shows clearly that in finding words the beginning of the word as such is of paramount importance.

The Information Factor

With regard to the effect of information, a shorter solution time was expected in the case of (a) few alternative solutions, (b) high frequency of the solution words, (c) high frequency (i.e., low information value) of the n-grams to be guessed, and (d) low frequency (high information value) of the n-grams given.

For each *S*, the total solution time required for the 6 nouns with the characteristics indicated above, was compared with the time needed for the remaining 6 nouns. The four hypotheses were tested by Wilcoxon's matched-pairs signed-ranks test for 48 *S*s. Results were all in accordance with the predictions made. The T-values were 538, 511, 738, and 763 respectively. All are significant at the .01 level. One or more of these effects, however, could be artifacts because of interdependency of the variables concerned. As is shown in Table 3, Spearman rank correlation over the 24 solution words only indicated an interrelation between factors (a) and (d).

In summary it may be concluded that high frequency of the solution word and high redundancy of the missing word part both facilitate the identification of the word. A large number of alternatives provided by the given n-gram and high redundancy of this word part lengthen solution time, although it cannot be decided whether these factors are to be distinguished or not.

So far, the results for initial and final n-grams have been considered together. A breakdown of these data shows that the tendencies found are due largely to the beginning of the word. With regard to the four variables in question, the final part of the word is less influential, whether it is given or missing.

All nouns produced by the *S* were checked for effects of word and n-gram frequencies. Regarding word frequency, it was hypothesized that the frequencies of words produced would be relatively high as compared with the frequencies of all possible words beginning with the same n-grams. The frequencies of all words produced by one or more *S*s during the first minute were listed, and the distribution of these frequencies was compared with the distribution of the frequencies of all possible nouns. In accordance with the hypothesis the distribution of the nouns mentioned appears to be displaced towards the higher frequencies. The effect is significant at the .01 level by Kolmogorov-Smirnov's one-sample test (Siegel, 1956).

A second question is whether, apart from word frequency, *S*s tend to add the most probable sequential letters to the given n-grams. The group of mentioned nouns was

TABLE 2
MEAN NUMBER OF NOUNS POSSIBLE AND ENUMERATED PER GIVEN n-GRAM PER *S*

	Initial letters given		Final letters given	
	Possible	Enumerated	Possible	Enumerated
Trigram (1st min)	19.7	1.28	31.1	0.99
Tetragram (2nd min)	3.8	0.88	5.0	0.62
Pentagram (3rd min)	1.1	0.66	1.4	0.50

Table 3

Spearman Rank Correlations between (1) Number of Alternative Solutions on the Basis of the Trigram Given, (2) Frequency of Solution Word, (3) Frequency of Missing Trigram, and (4) Frequency of Given Trigram

Correlations between	For initial trigrams	For final trigrams
(1) and (2)	—.029	.098
(1) and (3)[a]	.133	—.177
(1) and (4)	.765*	.614*
(2) and (3)	—.070	.098
(2) and (4)	—.070	.098
(3) and (4)	.021	.021

[a] In this case 'For initial trigrams' refers to the given trigrams for (1) as against the missing final trigrams for (3); 'For final trigrams' refers to the given trigrams for (1) as against the missing initial trigrams for (3).

* $p < .01$.

compared with an otherwise randomly selected group of not-mentioned possible nouns that had the same word-frequency distribution, and it was determined whether the third letters in the mentioned nouns have higher conditional probabilities. Comparison of the distributions by means of Kolmogorov-Smirnov's two-sample test yielded no significant results here. By a one-tailed χ^2 test (Siegel, 1956), a significance level between .05 and .10 was attained.

The conclusion is that, where two letters are given, *S*s tend to enumerate relatively frequent words. Apart from word frequency, no definite effect could be demonstrated in these words with respect to the conditional probabilities of third letters.

Discussion

The importance of initial letters in the identification of words is based on two factors: (a) In general, the initial part of the word contains more information; the more information is given in a word part, the more easily the word is identified. (b) Even with the same amount of information, production of nouns and, eventually, identification of the intended noun, are easier when the initial letters are given than when only the final letters are available.

Words appear to be retrieved from memory in sequential patterns, the initial letters being the obvious starting point. This may either result from the way in which verbal information is stored in memory, or be a characteristic of the retrieval mechanism. In addition, habituation to a larger amount of information in the beginning of words may have sensitized the *S* to any information given in that part.

The fact that, in the language, word beginnings are more informative than endings might be a consequence of a general psychological rule. In terms of information theory, this rule could be as follows: Taking away uncertainty in the beginning leaves the organism with relatively fewer possibilities to be kept in mind during the whole course of transmission, and for that reason improves the transmission of messages. A 'principle of least effort' (Zipf, 1949) applied to the language as a temporal sequence of speaking and reading, would in fact imply that information should be provided by the part of the sequence where its function is optimal.

References

Aborn, M., and Rubenstein, H. Psycholinguistics. *Ann. Rev. Psychol.*, 1960, **11**, 291-322.

Bruner, J. S., and O'Dowd, D. A note on the informativeness of parts of words. *Language and Speech*, 1958, **1**, 98-101.

Forgays, D. G. The development of differential word rceognition. *J. exp. Psychol.*, 1953, **45**, 165-168.

Howes, D. H., and Solomon, R. L. Visual duration threshold as a function of word probability. *J. exp. Psychol.*, 1951, **41**, 401-410.

Linschoten, J. *De la Court's frekwentietelling van Nederlandse woorden.* Psychol. Lab. R.U. Utrecht. Rapport nr. 6301, 1963.

Melville, J. R. Word length as a factor in differential recognition. *Amer. J. Psychol.*, 1957, **70**, 316-318.

Miller, G. A. *Language and communication.* New York: McGraw-Hill, 1951.

Miller, G. A., and Friedman, E. A. The reconstruction of mutilated English texts. *Information and Control*, 1957, **1**, 38-55.

MISHKIN, M., AND FORGAYS, D. G. Word recognition as a function of retinal locus. *J. exp. Psychol.*, 1952, **43**, 43-48.

OLÉRON, P. Étude sur l'apprćhension des mots. *Psychol. Française*, 1961, **6**, 21-31.

OLÉRON, P., ET DANSET, A. Données sur l'appréhension des mots. *Psychol. Française,* 1963, **8**, 28-35.

POSTMAN, L., AND ROSENZWEIG, M. R. Practice and transfer in the visual and auditory recognition of verbal stimuli. *Amer. J. Psychol.*, 1956, **69**, 209-226.

POSTMAN, L., AND SCHNEIDER, B. H. Personal values, visual recognition, and recall. *Psychol. Rev.*, 1951, **58**, 271-284.

PRATT, F. *Secret and urgent.* Indianapolis: Bobbs-Merrill, 1939.

RIEGEL, K. F., AND RIEGEL, R. M. Prediction of word-recognition thresholds on the basis of stimulus-parameters. *Language and Speech*, 1961, **4**, 157-170.

SACCO, L. *Manuel de cryptographie.* Paris, 1951.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

SOLOMON, R. L., AND HOWES, D. H. Word frequency, personal values, and visual duration thresholds. *Psychol. Rev.*, 1951, **58**, 256-270.

SOLOMON, R. L., AND POSTMAN, L. Frequency of usage as a determinant of recognition thresholds for words. *J. exp. Psychol.*, 1952, **43**, 195-201.

VAN SCHOONEVELD, C. *Bigramfrekwenties van de Nederlandse schrijftaal.* Rapport PL 206, Physisch Lab. RVO-TNO.

ZIPF, G. K. *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley, 1949.

(Received January 25, 1965)