GAP ANALYSIS AND SYNTAX*

Victor H. Yngve
Department of Modern Languages and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

A statistical procedure has been tried as a method of investigating the structure of language with the aid of data processing machines. The frequency of gaps of various lengths between occurrences of two specified words is counted. The results are compared with what would be expected if the occurrences of the two words were statistically independent. Deviations from the expected number give clues to the constraints that operate between words in a language.

Introduction

Language is a very complex communication code. One of the tasks of the linguist is to discover the structure of languages, or the rules of the codes, and to state them in a simple and concise way. To do this, he collects for data actual samples of a language, looks for regularities in the data by applying various procedures of analysis and an appropriate amount of intuition, forms hypotheses, and tests them on more data. When he is finished, he has what he calls a description, or a grammar of a language.

Many of the difficulties that the linguist faces in his task of discovering and describing structure stem from the very complexity of the code and from the large amounts of data that must be examined. It has been suggested that modern data handling techniques using punched card machines or electronic digital computers might be able to overcome difficulties that arise from the sheer bulk of data. The purpose of this paper is to discuss a way in which this might be done.

Our procedure is a statistical one and makes use of the fact that order and disorder are in a sense complementary. Statistical independence implies lack of structure, and any deviations from randomness can be taken as an indication of structure. The procedure, therefore, has two parts: one part deals with the setting up of an appropriate statistical model of language; the other part deals with the deviations from randomness exhibited by language and their interpretation in terms of structure.

We assume that language can be represented as a sequence of symbols. These might be letters, phonemic characters, syllables, morphemes, or other elements. It is convenient if there is an operational procedure for segmenting the text into separate symbols. Such a procedure exists for words in conventional spelling; they are separated in a text by spaces or punctuation. For purposes of the example in this paper, we have adopted English words as the symbols.

The obvious first step in the statistical analysis of a text is to investigate the relative frequency of the different symbols. Counting the frequencies of words has been a favorite occupation for over 60 years and a considerable amount of data exists[1]. In 1928, E.U. Condon[2] observed that when words are ranked in order of decreasing frequency, the product of the frequency and the rank is approximately constant. Several explanations have been offered for this or other formulations of the word-frequency distribution law. Of these, two are especially interesting.

B. Mandelbrot[3] assumes that words are separated by spaces and are spelled with letters to which a cost function is attached. A message composed of words with the observed frequency distribution transmits the maximum amount of information in the sense of Shannon, compatible with a given average cost per word.

H.A. Simon[4] assumes a simple stochastic model. The probability that the next word to appear will be one of the words that has already appeared n times is set proportional to the total number of occurrences of words that have each appeared n times. There is a constant probability that the next word will be a word that has not already occurred. The observed frequency distribution agrees with the one that will keep the fraction of the words that occur n times approximately constant.

The observed distributions are nearly the same for all languages. If the assumptions on which these explanations are based are valid for one language, they are valid for all languages. The Mandelbrot explanation involves an economy argument; the Simon explanation follows if users of a language try to maintain the word frequencies that they observe.

After an investigation of the frequencies of individual symbols, the next step of a statistical analysis of a text is an examination of intersymbol constraints. These are of more direct concern to the linguist because they are different for different languages.

Intersymbol constraints have been investigated for various purposes. For cryptanalysis

* This work was supported in part by the Army (Signal Corps), the Air Force (Office of Scientific Research, Air Research and Development Command), and the Navy (Office of Naval Research); and in part by the National Science Foundation.

purposes, frequency tables of two-letter and three-letter sequences have been tabulated. For the purposes of estimating the entropy of printed English, Shannon[5] has used various methods of measuring the conditional probabilities that various letters will follow certain sequences of letters. The conditional probability concept has been used[6] as the basis for a model of a human being regarded as a talking animal. A grammar is conceived as an enormous array or matrix of the conditional probabilities that each morpheme in the language will be produced after a given sequence of morphemes. A scheme of this sort focuses attention on each position in a text and on the effect there of the immediately preceding one, two, three symbols, etc.

The method of investigating intersymbol constraints reported here is also concerned with the conditional probability of finding a given word at a certain position in the text. But instead of specifying the immediately preceding one, two, or more, text positions and investigating the effect on the probability of the words found there, we specify certain words or word combinations and investigate their effect as they are moved around in the vicinity of the given word. An advantage of this is that it allows more direct investigation of the effect of the occurrence of a word on the probabilities some distance away. It also allows easy and rather full investigation of the effects of the most frequent words first. Being most frequent, these words have an especially great influence on the grammar.

The Procedure of Gap Analysis

The statistical model that we use is a model for a text divided into symbols (words). We assume that the frequency f of each word V and the total number of words N, or the length of the text, are given as a result of direct measurement. We assume that the probability of occurrence of each word is equal to its relative frequency, p(W)=f(W)/N, and is therefore independent of its position in the text and of what words are nearby.

We look for deviations from the assumption that the probability of a word occurring is independent of the words in the neighborhood. To do this, we choose two different words and investigate their effect on each other's probability of occurrence. Or we can investigate the effect that a given word has on other occurrences of the same word. We define a gap of type A-B as the number of words intervening between an occurrence of A and a later occurrence of B. We can have gaps of type A-A between two occurrences of the same word. For each type of gap, we count the number of gaps of length 0, 1, 2, ... This can be done easily by machine by collecting one sample of text for each occurrence of A. All samples should be the same number of words in length and should have the occurrence of A at the same position. Then, for each of the other word positions, the number of occurrences of B in all the samples is counted. The results can be plotted as a histogram of the number of gaps against the gap length. Gaps of type A-B can be plotted on the right of the center of the histogram; gaps of type B-A on the left.

Several features of the histogram presentation of gap data should be noted.

1. If the probability of occurrence of B is independent of its position with respect to A, we expect the distribution of gaps to be flat except for statistical fluctuations.

2. The expected number of gaps of length n is then independent of n and can be calculated from the given frequencies:

$$f(G_n) = p(B)\ f\ (A) = \frac{f(A)\ f(B)}{N}$$

3. We ignore the effect that the ends of the text have in reducing the possible number of gaps. Such effects will be small if the gap lengths investigated are appreciably smaller than the length of the entire text.

4. A histogram with gaps of type A-B plotted on the left and gaps of type B-A plotted on the right is a mirror image of a histogram with B-A on the left and A-B on the right.

5. A histogram of gaps of type A-A is symmetrical about the center position.

6. If a gap had been defined as the number of words occurring between an occurrence of A and the first occurrence of B, the assumption of statistical independence would, of course, lead to an exponential distribution instead of a flat one, a fact that seems not to have been understood by various counters of gaps. We would have for the frequency of gaps of length n,

$$f(G_n) = f(A)[1-p(B)]^n p(B) = \frac{f(A)\ f(B)}{N} e^{-kn}$$

where $k = -\ln[1-p(B)] > 0$

For our purposes, gaps of the exponential type are not as convenient because they are harder to count by machine, require more calculating to obtain the expected number and the expected deviations, and because histograms with gaps of type A-B on the left and B-A on the right are not mirror images of those with B-A on the left and A-B on the right on account of the different exponentials.

Trial Application to English Structure

In order to be in a better position to assess the results in a first trial of the above procedure, we selected a small sample of a familiar language - English. An article of about ten thousand words from a popular magazine was chosen. Since this was a rather short article, only six of the most frequent words were investigated. The total number of words was counted as well as the frequency of each of these six words. These

numbers are tabulated below:

word	frequency
the	599
to	252
of	241
a	221
and	207
in	162
	1682
(number of words in article)	9490

It can be seen that these six words alone account for over 17.5 per cent of the occurrences of words in the text. Punctuation was ignored.

Using these six words, all fifteen of the type A-B gaps were counted, and the six of type A-A. The results of the gap counting are presented in Figs. 1 and 2. Along the abscissa of each histogram are plotted the various word positions. For example, in Fig. 1-g, the word "the" in all of the "the" samples of text is placed at the center position. The length of the bars of the histogram represents the number of times the word "a" appeared at the various text positions to the right or left of the "the". The numbers along the abscissa give the length of the gap or the number of words intervening between the occurrence of the word "the" and the word "a". The six histograms of the gaps between two occurrences of the same word are shown in Figs. 1-a to 1-f. In Fig. 1-f, the gaps were counted out to a length of 31 words, and since the histogram would be symmetrical anyway, only the right half is plotted.

The expected height of the histogram bars, under the assumption of the statistical independence of the two words, is given by the middle horizontal line. The upper and lower horizontal lines represent deviations amounting to plus or minus the square root of the height of the middle line.

Discussion of the Data

It can be seen that, in general, the histograms show considerable deviation from what would be expected on the assumption of statistical independence. These deviations can be attributed to syntactic structure. Since our aim is to develop techniques that can be used in discovering structure, it is of interest to see how the deviations from randomness shown by the data correlate with what is known about the structure of English.

It two words occur together in a structure, that particular combination of two words will probably occur more frequently than expected on the assumption of statistical independence. The greater frequency of the particular combinations representing structures reduces the probability of occurrence of other combinations that do not represent structures.

Figures 1-a to 1-f all show a depressed region near the center of the histogram. This is taken to mean that these words tend not to recur immediately. The device of reduplication has only a limited use in English; this is probably true of many other languages. The length of the depressed region gives an idea of the length of the structures that frequently occur with these words. Tor example, structures with "the" can be expected to have two or three words. This is indeed true. But in the case of "and", the depressed region extends over at least 15 gaps. The total number of gaps of length 15 or less between occurrences of "and" amounts to only 50 as compared to an expected 68. This long depressed region can be understood as attributable to the fact that "and" not only correlates words, but longer structures as well. One of the uses of "and" is to coordinate clauses. Two occurrences of "and" used for this purpose cannot be closer together than the length of a clause.

Figure 1-a also shows that the word "the" has a slight periodicity with a gap length of 2 to 6. Such a periodicity can result from structures like the following:*

the tasks of the (2)
the linguist is to discover the (4)
the structure of languages or the (4)
the rules of the (2)
the difficulties that the (2)
the very complexity of the (3)

The average gap length between nearest occurrences of "the" is about 15 words. It is true that the depression at 0 and 1 must be compensated for by an increase elsewhere, but in the absence of other constraints, this increase would not cause a peak, but would be spread evenly over all other gap lengths. There would be fewer positions available for "the", but they would all be equally probable.

The gaps between different occurrences of the some word give an exactly symmetrical histogram, because it is always possible to interchange the words without altering their roles. Whenever two different words give an approximately symmetrical histogram, it gives us a clue that it may be possible to interchange them without altering their roles, i.e., they often play the same role and can be classed together.

In Figs. 1-g to 1-l, we have collected all the rest of the histograms that might be considered symmetrical. There is little question about the first four, but perhaps the last two do deviate from symmetry by more than the statistical fluctuations. Fig. 1-k shows possible deviations at gap lengths of one and two; Fig. 1-l shows possible deviations at a gap length of one. Let us assume that the top four are symmetrical. On this basis we tentatively group together and name:

* Examples are taken from the first paragraphs of this paper.

"the" and "a" (the article group)
"of" and "to" and "in" (the preposition group)

keeping "and" separate from all the rest on the assumption that Figs. 1-k and 1-l are not symmetrical .

All of the histograms of Fig. 2 are unsymmetrical. For two words to have an unsymmetrical gap histogram, they must frequently play different roles with respect to each other, and therefore they should not be grouped together.

Figures 2-a to 2-f are the six histograms that relate an article and a preposition. Our tentative grouping is given additional weight because these six histograms show certain similarities that can be attributed to the nature of articles and prepositions: The pattern "preposition article" occurs, and often with high frequency, while there are no cases of "article preposition".

These six histograms also show differences between the two articles and between the three prepositions. "The" is different from "a" in that it is the preferred article after "of". This shows up when Fig. 2-a is compared with Fig. 2-d. "Of" is different from "to", and from "in", which is probably a typical English preposition, in that "of" frequently follows an article with a gap of one or two. This is due to the very frequent "genitive" construction:

the tasks of
the structures of
the rules of
a grammar of

"To" is different from "of" and "in" in that it has a relatively low and broad peak before "the". The lowness of the peak is probably caused by the competition between the prepositional use of "to" and the use of "to" before an infinitive. The broadness is probably caused by an infinitive interposed between "to" and "the":

to discover the

Figures 2-g, 2-h, and 2-i show that "and" is different from "the", "a", and "in". If we take Fig. 1-k and Fig. 1-l as being unsymmetrical, it is also differentiated from "of" and "to".

The outlying peaks in Figs. 2-b and 2-e have not been explained. Perhaps they are statistical fluctuations. The accuracy of the counting has been verified.

Conclusions

The first trial of the use of gap analysis for revealing certain aspects of the syntax of a language has been quite fruitful in exposing certain ways in which the technique can be improved, and has been quite suggestive of its potentialities, The technique is certainly not a purely mechanical way of investigating the structure of language. A considerable amount of insight into language structure is required in order to make best use of the gap histograms as a tool of analysis.

The success of the procedure depends largely on the skill with which the text has been segmented into symbols. It is felt that this particular experiment would have been more meaningful if morphemes had been used instead of words as they are spelled. By using words, however, we eliminated much preliminary work. If one wants to use morphemes, perhaps it would be appropriate to segment into phonemes and use statistical procedures directly on the phonemes. The frequent morphemes would soon appear as frequent patterns of phonemes. If one sticks to conventional spelling, it would probably be better to include punctuation on a par with words.

One of the most serious limitations of our application of the procedure to English, was the shortness of the text that we chose. Conclusions could have been drawn with much greater certainty if statistical fluctuations had been smaller. Instead of 10,000 words, perhaps 100,000 words should be the minimum length of text for gap analysis. With a longer text, one could include many more of the frequent words because the word frequency distribution function begins to level off. Also with a longer text, one could take the next step of treating frequent constructions in the same manner as words and examining their effect on words in the vicinity. The construction "of the" and "the - of" were particularly frequent. They are probably as frequent as the 10th or 15th ranking word!

A systematic order of procedure for another experiment would be to count the frequencies of the words; then to count the gaps, taking first those that have the highest product of the frequencies of the words involved; then to look for frequent constructions and collate them into the list of word frequencies so that they could be used along with the words for further gap counts. By comparing the behavior of certain words in the vicinity of a two-word construction with the behavior of the words in the vicinity of each individual word involved in that construction, light can be shed on the multiple functions of words.

There are a number of other things that can be done with gap histograms. Certain features that many histograms or groups of histograms have in common, such as the "prepositional peak", can be used as a basis for grouping words together. Then the whole group of words can be counted as one to increase the number of cases, and rarer words can then be examined. On the other hand, the less frequent words can be grouped together on the basis of their behavior in the vicinity of frequent constructions. For example, one could group together all words that occur in the position between "the" and "of". Then the statistical behavior of the group could be investigated as if it were a single word. There is a certain resem-

blance here to the use of the substitution frame. Perhaps gap analysis will be able to reveal the best substitution frames for use with the more standard methods of linguistics.

Since the methods of gap analysis are far from being highly developed at this early stage, it is rather difficult to draw much of a comparison with the more standard linguistic methods involving informant techniques with native speakers. It is particularly difficult, if not impossible, to get accurate quantitative information from an informant. For this reason, grammars have made no pretense of being quantitative, but have merely reported what can occur and what cannot occur. Occasionally, linguists add the comment that something is "rare" or "usual" or a "favorite" construction. Statements like this reveal that they think that a certain amount of quantitative infornation is relevant, and that they would probably give more if they had the technique.

Gap analysis provides a wealth of numerical information, perhaps more than is really relevant. The linguist who uses it may have to pick and choose. Some types of numerical information about a text are of little significance. For example, the word "police" might be frequent in a newspaper, but the word "circuit" might be frequent in an electrical engineering article.

Because gap analysis is a numerical technique, it focuses attention on frequencies and numerical results. There is the continual implication that these numbers are worth something, that they are a relevant part of a grammar. To a certain extent this is true. In general, sentence structure is carried by combinations of the frequent morphemes. Infrequent words cannot indicate by their form alone their role in the sentence unless they have included in them some frequent role-marking morpheme. For example, a nonsense word like "sklack" could be a noun, a verb, an adjective, or an adverb. But "the sklack" or "sklacked", would have definite roles indicated by the frequent "the" or "ed". A sentence, then, can be considered as a structure of frequent morphemes with various open positions in it where all the rest of the morphemes can be put, including the infrequent and the new words. The frequent morphemes and their combinations can be considered as role markers for the less frequent ones. One can conclude that the frequent morphemes are the important ones for stating syntactic patterns, and that a careful use of gap analysis should be able to reveal these patterns.

References

1. Guiraud, P., "Bibliographie Critique de la Statistique Linguistique," Spectrum, Utrecht-Anvers, 1954.

2. Condon, E.U., "Statistics of Vocabulary," Science, 67, 300, (1928).

3. Mandelbrot, B., "Simple Games of Strategy Occurring in Communication through Natural Languages," Trans. I.R.E., PGIT-3, (March 1954), P. 124.

4. Simon, H.A., "On a Class of Skew Distribution Functions," Biometrica, 42 (Dec. 1955), pp. 425-440.

5. Shannon, C.E., "Prediction and Entropy of Printed English," Bell Telephone System Monograph 1819 (1950).

6. Hockett, C.F., "A Manual of Phonology," Baltimore, Waverly Press, 1955, (Indiana University Publications in Anthropology and Linguistics, Memoir 11)

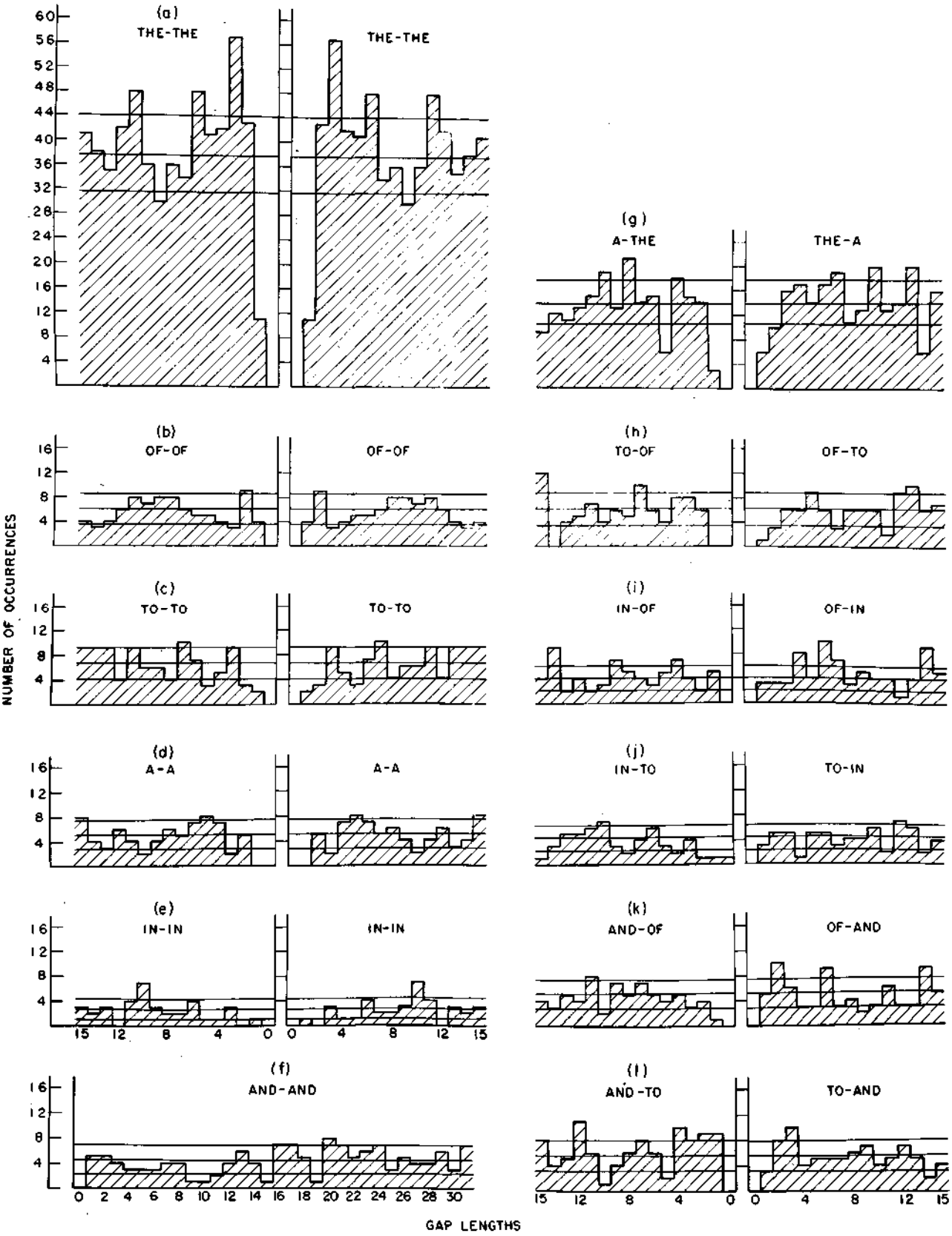

Fig. 1 - Gaps between occurrences of words.

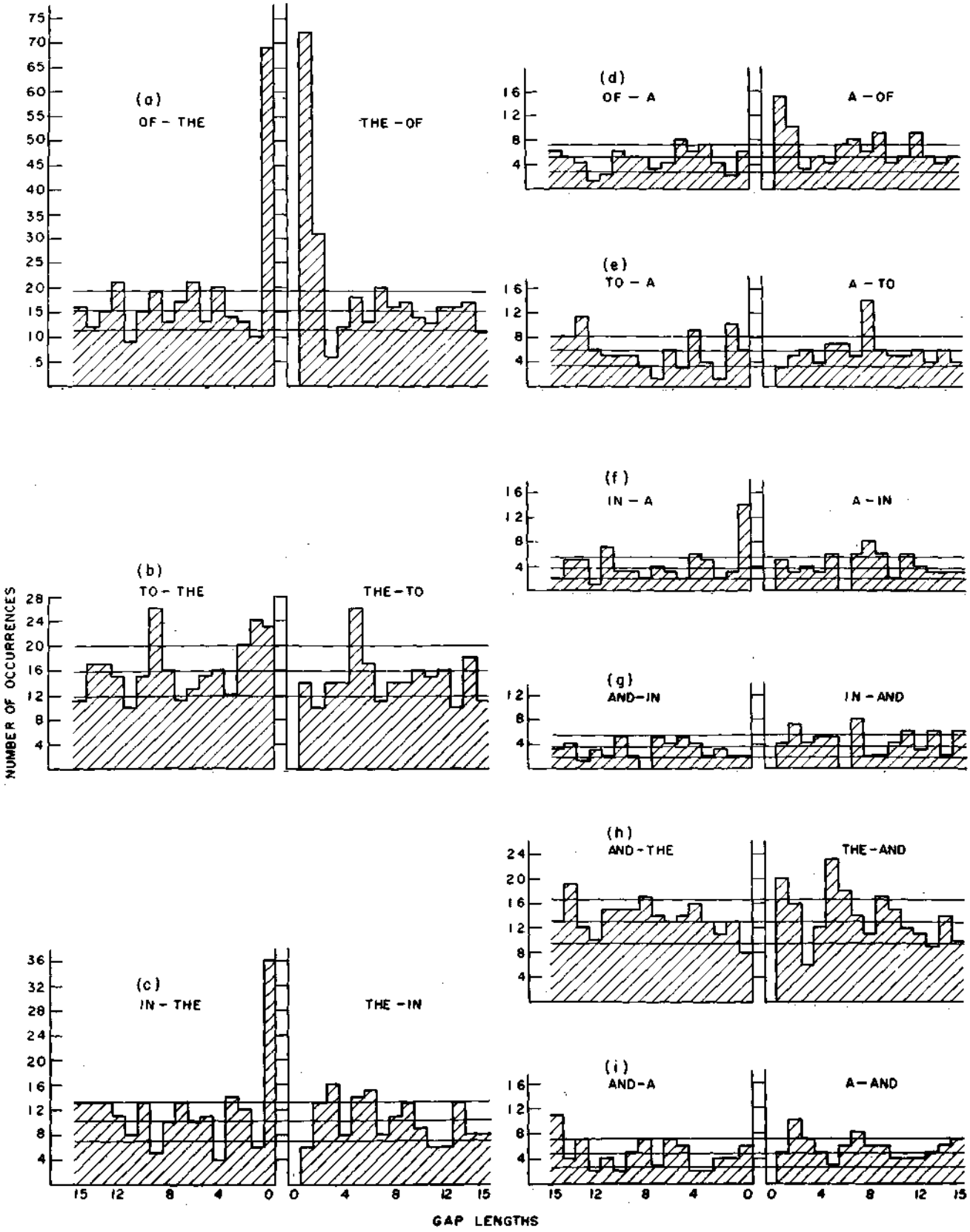

Fig. 2 - Gaps between occurrences of words.