and satisfactory definition exists, and some linguists deny any validity of the word, relegating it to folk linguistics.

Following Greenberg we take words as being composed of morphemes so that a word may be identified with a sequence of morphemes and no morpheme overlaps two words. From the distribution of the morphemes of a corpus we find clusters which approximate the words of the corpus. The approximating units are determined relative to the corpus from which the distribution is defined. The corpus may be either considered as a closed sublanguage in itself or as a sample from some larger corpus. We study the behavior of approximate units relative to longer and longer portions of the corpus, and also relative to the corpus considered as a statistical sample.

Assuming that a word may be represented as a sequence of morphemes, how should this sequence be distinguished? In the well-known paper of Togeby, (1949) there is a convenient summary of structural views of the word. In his discussion, the word is set forth as a morpheme sequence possessing properties classified under the headings of 1° Forme libre minimum, 2° Séparabilité, and 3° Permutabilité. In considering how a morpheme sequence should be distinguished as a word we will begin by examining Togeby's classifications.

In Togeby, under the discussion of a word as a forme libre minimum, reference is made to Bloomfield's (1933) statement about the word as a minimum free form and the smallest items which are spoken by themselves, in isolation.

The idea of minimum free form is actually found somewhat earlier in Bloomfield's (1926) Postulates.

> A minimum free form is a word. A word is thus a form which may be uttered alone (with meaning) but cannot be analyzed into parts that may (all of them) be uttered alone (with meaning). Thus the word quickly can be analyzed into quick and -ly, but the latter part cannot be uttered alone; the word writer can be analyzed into write and -er, but the latter cannot be uttered alone (the word err being by virtue of different meaning a different form) ...

Similar views are found in the older "universal grammars." They differ principally in taking the Aristotelian position that the word and not some smaller unit has meaning. For example, in Harris (1771) we find a concern with minimum units of meaning.

> But what shall we say? Have these parts (of a Quantity of Sound) again other parts, which are in like manner significant and may be pursued to Infinite? Can we suppose that all Meaning, like Body, to be divisible, and to include within itself other Meanings without end? If this be absurd, then must we necessarily admit, that there is such a thing as a Sound significant, of which no part is of itself significant. And this is what we call the proper character of a Word. For thus though the Words (Sun) and (shineth) have each a Meaning, yet is there certainly no Meaning in any of their Parts, neither in the syllables of one, nor the Letters of the other.

James Harris refers to Priscian's definition in which the word is defined as a minimum meaningful utterance in connected speech.

> Dictio est pars minima orationis constructae, id est, in ordine compositae. Pars autem, quantum ad totum intelligendum, id est, ad totius sensus intellectum. Hoc autem ideo dictum est, nequis conetur vires in duas partes dividere, hoc est, in vi et res; non enim ad totum intelligendum haec fit divisio.

For purposes of constructing our model we shall interpret minimum free form as follows:

A word is a sequence of subword units. If this sequence may be uttered alone, then it is to be expected that the sequence co-occurs freely with other sequences.

Under the classification of separabilite, Togeby places the requirement of Jakobson (1938) that words are the separable components of phrases: minimal actually separable components of the phrase. Conversely, the constituents of a word should not be separable.

The general requirement of séparabilité seems to be that a word is a morpheme sequence which may co-occur with other morpheme sequences to give grammatical utterances. If the sequence is a distinct word, then its morphemes must be contiguous, and the morphemes of a noncontiguous grammatical sequence cannot be identified with the same word.

Under permutabilité, Togeby quotes Hjelmslev(1943) "les mots pourront tout simplement être définis comme les signes minima dont l'espression, et de même le contenu, sont réciproquement permutables." According to Togeby, Hjelmslev means that "un changement de l'ordre des mots pourra entraîner un changement de sens, tandis qu'un changement de l'ordre des parties du mots n'en sera pas capable."

The requirement here is that if a sequence of morphemes is identified with a word, then the order of the sequence must be invariant.

In Greenberg (1957), the proposed definition of the word based on substitution and the recognition of grammatical sequences, we interpret as follows:

Let S be a sequence of linguistic units and G the class of grammatical sequences, in Greenberg's words the class of sequences which "exist as expressions in the language."

Suppose that S= X A B C D E$\in$ G is a morpheme sequence. We want to decide whether or not the boundary between B and C is a word boundary.

To each morpheme of S there corresponds a "nucleus." For the nucleus of B to be a word terminal it is necessary that "infinite insertion" of nuclei be possible between B and C, otherwise if there"is a maximum to the number of nuclei that can be inserted," the boundary is "intra-word boundary."

Nuclei are classes of morpheme sequences having strongly equivalent substitution properties. Some of the conditions for class membership are so strict that we would expect the defined classes to be empty for the language taken as a whole. Perhaps as Chomsky conjectures in a review of Greenberg's essay: "It might be that the notion of word may be defined relative to a particularly simple set of sentences. (1958)

In practice, Greenberg's conditions might be interpreted as follows: S= X A B C D E occurs in the language. The subsequence BC may belong to a single word if it is replaceable by a single morpheme and grammaticality is preserved. If for a small number of morpheme sequences S_i, the sequences X A B S_i C D E are grammatical, then the subsequence BC belongs to the same word. If the sequences X A B S_i C D E are grammatical for a large number of S_i, then the subsequence BC probably does not belong to the same word.

In an unpublished MS, Juilland develops a constructive definition of the word which requires the recognition of grammaticality. If S = X A B C D E $\in$ G is a morpheme sequence, the boundary between B and C is classified according to the potential sentence occurrences of B and B. Boundaries are classified as "conjunctive"or "disjunctive." Disjunctive boundaries isolate potential words called "functional units." Conjunctive boundaries occur potentially within words but must be tested by an

"insertion criterion." Thus if BC spans a conjunctive boundary, then B is a word boundary if there exists a morpheme sequence S_i such that X A B S_i C D E $\in$ G.

The Use of Numerical Linguistic Data

Our object now is to define a quantitative procedure for approximating words. The procedure attempts to meet the various requirements summarized in the last section. Since our interest is in distributional methods, we do not want the procedure to include an independent test for grammaticality.

The requirements that we attempt to fulfill are summarized by Julliand as adhesion and separability. These are realized as a common characteristic in the procedures of Greenberg and Juilland: A potential word is isolated as a sequence of morphemes which are associated in some special way, then the potential word is tested for its function as a word, according to some test of insertion.

Let us imagine a linguist confronted by the following data. Frequency refers to text frequency. Let X A B C D E be a sequence of morphemes to be segmented. Consider the boundary between B and C. Is this boundary a word boundary? Assume first that B occurs only with A, E, C and G, as indicated in Case 1.

Morpheme Pair	Frequency	Morpheme Pair	Frequency
AB	4	BC	3
EB	6	BF	7

Case 1

With no further information, we might observe that B occurs more frequently with A than with C, and segment as AB CD. Under this condition the requirement of adhesion may be met, but a simple consideration of frequencies is not sufficient to meet the requirement of separability. This is illustrated by the hypothetical set of data of Case 2.

Morpheme Pair	Frequency	Morpheme Pair	Frequency
AB	4	BC	3
EB	1	BF	7
GB	1		
HB	1		
IB	1		
JB	1		
KB	1		

Case 2

In this case the frequency of AB also exceeds the frequency of BC, but the segmentation AB CD would not agree with linguistic intuition at all. In Case 2, B has much greater freedom of combination on the left than on the right, and to satisfy the condition of separability, at least approximately, we would segment as A BCD.

In formalizing these intuitions, we refer to the procedure of Harris (1955) for grouping phonemes into morphs. Harris assumes that an utterance U may be represented as a sequence of phonemes $a_1\ a_2 \ldots a_n$. Let $R(a_1)$ be the number of different phonemes which may follow the phoneme a_1 in the total language. Similarly, let $R(a_1, a_2)$ be the number

of different phonemes which may follow a_1, a_2 and so on. Likewise, let $L(a_n)$ be the number of different phonemes which may precede a_n, $L(a_{n-1}\ a_n)$ the number which may precede $a_{n-1}\ a_n$, and so on. Then the sequence

$$SR = R(a_1)\ R(a_1a_2)\ R(a_1a_2a_3)\ \ldots\ R(a_1\ a_2\ \ldots\ a_n)$$

describes the freedom of co-occurrence on the right at each phoneme of U, and the sequence

$$SL = L(a_1\ a_2\ \ldots a_n)\ L(a_2\ \ldots\ a_n)\ \ldots\ L(a_n)$$

describes the freedom of co-occurrence on the left at each phoneme of U.

Harris observes that morpheme boundaries tend to occur at positions in U where the corresponding values of R and L are large or attain their relative maxima. Thus if $R(a_1\ \ldots a_k)$ is a relative maximum in the sequence SR, then a_k is a morpheme terminal. Likewise a_k is a morpheme terminal if $R(a_1\ \ldots\ a_k)$ exceeds a value comparable to the total number of different phonemes in the language. Under similar conditions for $L(a_j\ \ldots\ a_n)$, a_j is a morpheme initial.

Applied to sequences of morphemes with uncontrolled diversity, Harris's procedure becomes particularly unwieldy. We suggest that we might achieve the same results as Harris by using fixed-length subsequences rather than some higher-level syntactic unit. Thus for some fixed k, the co-occurrence measures

$$R_k(a_1\ldots a_k)\ R_k\ (a_2\ldots a_{k+1})\ \ldots\ R_k(a_{n-k+1}\ldots a_n)$$

might yield the same segments as the sequence

$$R(a_1)R(a_1a_2)\ \ldots\ R(a_1\ \ldots a_n)\ .$$

A Segmentation Procedure

The placing of segment boundaries at positions of maximum freedom of combination realizes separability, but the requirement that a word should be a morpheme sequence showing strong internal association is accounted for only in a negative way--we do not place boundaries at positions of low freedom of combination. We propose another procedure for grouping morphemes by combining both left and right freedom of co-occurrence. As a result we derive a scale of degrees of distributional separation.

In Harris's procedure there is sufficient information to form a ranking of boundaries. If $a_1 \ldots a_n$ is the sequence to be segmented then we place a boundary between a_k and a_{k+1} if one or more of the following conditions is met.

1. $R(a_1 \ldots a_k)$ is a relative maximum in SR.
2. $L(a_{k+1} \ldots a_n)$ is a relative maximum in SL.
3. R or L are large in comparison with the number of different phonemes.

If any two of these conditions are satisfied, we have stronger distributional evidence for segmentation than in the case of just one alone. Likewise, if all three conditions are fulfilled, then we would expect that a_k would be a morpheme terminal more often than if just two of the conditions are fulfilled. We shall adopt a similar line of reasoning to segmentations based on the distributions of fixed-length sequences.

For convenience we introduce some notation. Let

A B) C D indicate a right-hand boundary after B,

following from the distribution of B,

and

A B (C D indicate a left-hand boundary before C,

following from the distribution of C.

In a "first-order segmentation" of the sequence XABCDE, we will use only the distributional properties of single morphemes. Thus, in our hypothetical Case 2, we refer only to the distributional properties of B.

Morpheme Pair	Frequency	Morpheme Pair	Frequency
AB	4	BC	3
EB	1	BF	7
GB	1		
HB	1		
IB	1		
JB	1		
KB	1		

In this case the text frequencies indicate that B has much greater freedom of combination on the left than on the right. Given no further information, we segment as A (B C D. We formalize this decision in the following "Cutting Rule."

If R(B) > L(B) cut as X A B) C D E.

If R(B) < L(B) cut as X A (B C D E.

If R(B) = L(B) cut either as X A B) C D E or as

X A (B C D E.

Let us insert right- or left-hand boundaries at C by use of the cutting rule, as we did with B. The strongest evidence for segmentation (separability) is in the case where R(C)>L(C), so that we place a left-hand boundary before C; and at the same time R(B)>L(B), so that we place a right-hand boundary after B. The result is indicated as A B) (C D. The weakest evidence for segmentation (adhesion) is where R(B)<L(B), and at the same time R(C)>L(B). The result is indicated as A (B C) D.

There are nine possible combinations according to the distributional properties of B and C. These are shown in Figure 1 , which we refer to as a "Segmentation Rule." The number of slashes--the "degree" of the boundary--indicates the relative evidence for segmentation.

R(B) - L(B) \ R(C) - L(C)	>0	=0	<0
>0	B//C	B///C	B////C
=0	B/C	B//C	B///C
<0	BC	B/C	B//C

Figure 1. Segmentation Rule

The first sample which we will consider for purposes of illustration is from the primer Ted and Sally. This text contains 121 different printers' words in all. As in other deliberately morphemically closed

texts, Zipf's law does not operate so we have a large variety of contextual combinations with many repetitions. The sample consists of the first 4,670 morphemes and forms the main narrative. We obtain the segments:

come//Boots//sai d//Ted////

come and//ride////

come and//ride///in///my wagon////

jump/in///Boots//sai d//Ted////

ride///in///my//wagon//Boots//

jump/in/and//ride////

here//we//go//sai d//Ted

The foregoing segmentation is first order in that inference is made using only the distribution properties of single morphemes. The procedure may be extended to consider n-tuples of units for "n-th order" rules. However rules using extended context have two difficulties. One is the simple difficulty of finding enough context in a short text. A second, more interesting restriction is that certain boundaries may not follow each other, depending on the order of the segmentation. For example, two zero-degree boundaries may not follow each other under a rule of any order.

The simple type counts, as measures of freedom of co-occurrence may be replaced by other more general measures, for example the entropy E of the type-frequency distribution. See, for example, Khinchin (1957). Entropy has the desirable property that it may be used to estimate the average number of morphemes that may co-occur with a given unit. For example, if the unit U has entropy $E_R(U)$ of successors, then the "diversity"

of successors is $2^{E(U)}$. The entropy would be the same if all the $2^{E(U)}$ successors were equally likely.

Evaluation Procedures

Applied to real data, the constructive procedures of Greenberg and Juilland are developed with the aid of many illustrative examples, but are still programmatic and have not been applied to large linguistic samples. Likewise, Harris gives the morphemic segmentation of many sentences but does not give a numerical evaluation of his results for a large text.

In evaluating our approximation procedures, we will be concerned with degrees of adequacy. The results presented so far suggest that there is a strong correspondence between the degree of a segment boundary and the corresponding syntactic boundary. It appears that segment boundaries of zero and first degrees correspond to intra-word boundaries, second-degree segment boundaries to word boundaries, and third-and fourth-degree boundaries to phrase and sentence boundaries.

To determine the correspondence, we give a more precise formulation. In the morpheme sequence X A B C D E let B^1 and C^1 be the lowest level constituents containing B and C respectively. It may happen that $B^1 = B$ and $C^1 = C$. If B^1 and C^1 belong to the same printers' word, then the syntactic boundary between B and C is a morpheme boundary. If B^1 and C^1 do not belong to the same printers' word, then the syntactic boundary between B and C is labeled according to the highest syntactic level of B^1 or C^1.

Thus in the sequence un gentlemanly the space marks a morpheme boundary, since ungentlemanly is a printers' word. However in the king of England's, where B = England and C = 's, B^1 = the king of England and C^1 = 's. Consequently we take the boundary between England and 's as a phrase boundary. In the two word sequences, the man and he went, the spaces mark word and phrase boundaries respectively.

Between any two morphemes we have 20 possible combinations of syntactic and segment boundaries. The correspondence may be evaluated by the χ^2 statistic, or derived statistics such as the contingency coefficient $C = \sqrt{\chi^2/N + \chi^2}$. See, for example, Kendall (1952).

Some Distributional Groupings

We examine the correspondence between syntactic and segment boundaries using several samples of morphemic data.

In many cases a zero-degree pair occurs in a manner which is only barely statistically significant. Let us compare.

	look	Sally
R/L	13/18	39/33
Sign (R-L)	-	+
	and	
	come	and
R/L	19/32	21/19
Sign (R-L)	-	+

For the sequence lookSally, the differences (R-L) appear to be statistically significant, but in comeand, we may wonder whether the slight positive value of R(and) - L(and) is due to sampling variations.

In a statistical version of our procedure, we test the hypothesis that R(and)$>$L(and). Since there is no exact sampling theory for this test, we construct an approximate test. The 4646 morpheme text is divided into approximately equal blocks, and R-L computed for each block separately. The values of R-L may be viewed as independent samples, provided the individual block size is large enough. We infer from the signs of R-L in each block that R(come)$<$L(come). But we may not infer that R(and)$>$L(and), since the positive difference occurs in only one trial in five. On the other hand, for the pair look Sally, R(look)$<$ L(look) and R(Sally)$>$L(Sally) in all five blocks.

Considering the 4646 morpheme text as a statistical sample, the inferred zero-degree segments are sai d, look Ted, look Sally, and run Ted. If they occurred, run Sally, look Ted, say Ted and say Sally would also have zero degree, while come run and come look would be of second degree.

The next sample is from a lower school reader. The corpus is the first 2100 morphemes from a simplified version of Robinson Crusoe. Even though this text is simplified, it is fairly representative of ordinary language and the frequency distribution follows Zipf's law. The words are morphemically simple, but many morphemes occur only once. For the first sentence, the morphemic representation and the groupings relative to samples of the first 300, 600, ..., 2100 morphemes follow.

The ship be ing fit ed out I go ed on board the one st of September 1659.

Theship be ingfittedout Iwent onboard thefirstofSpetemberl659
Theship beingfittedout I went onboard thefir st of Septemberl659
Theship beingfittedout I went onboard the first ofSeptemberl659
The ship beingfittedout I went onboard the first ofSeptemberl659

. . .

The ship beingfittedout I went onboard the first ofSeptemberl659

As soon as the sample reaches 1200 morphemes, the segmentation becomes stable. In this first sentence ing fitt ed and of September 1659 remain unsegmented since fit, September, and 1659 occur only once each

and we lack distributional information. The pairs being, fitt ed, wen t, and fir st are coextensive with printers' words. On board shows strong association and is operationally a word. The morpheme the shows strong disassociation in the context the////first, but neutral association in the context the//ship.

Several high-frequency morphemes tend to occur early in the text so that we have fairly extensive distributional information for the first sentence, but less information for morphemes occurring later in the text. In this sample there are 424 different morphemes. Of these 215 occur only once and 82 only twice, so we have little information for segmentation. On the other hand, the high-frequency morphemes the, ship, be, ing, out, ... all occur in the first sentence. A consequence is the poor performance of the procedure when applied to more than the first two sentences. See table .

A final example is Quine's Word and Object. We show the segmentation of this sentence relative to a sample of 900 morphemes. Even though the words tend to be polymorphic, the morphemic diversity is smaller than that found for the first 900 morphemes of Robinson Crusoe. The values are 4.0 and 5.1, respectively. It follows that morphemic combination in Word and Object is more restrained and the occurrence of longer words does not imply more freedom of morphemic combination.

The segmentation follows.

For // the / case / of //// sent // ence s //// gener al ly //// how ever // or //// even // the / case / of // e tern al // sent // ence s //// gener al ly sure ly //// there // i s ///

no / thing /// ap proach ing a //// fix ed //// stand ard //
of //// how / far /// in // di rect //// quot ation /// may ///
de viate /// from // the // di rect.

The morpheme groupings are:

For thecaseof sent ences generally however or even
thecaseof eternal sent ences generallysurely there is
no thing approachinga fixed standard of howfar in
direct quotation may deviate from the direct.

Some numerical results are summarized in Table 1. The measures of correspondence are between word boundaries and segment boundaries of degrees two, three, or four. In Table 1, Length refers to the text length in morphemes, and N is the number of boundaries for which the correspondence measures were computed.

Text	Length	Rule	N	X^2	C	Diversity
Ted and Sally	4646	Second Order	197	104.4	.59	2.8
Robinson Crusoe	2100	First Order	95	5.0	.07	7.0
Word and Object	900	First Order	95	35.8	.85	4.1

Table 1. Summary of word and segment correspondences.

The general conclusion is that words do correspond to segments of at least second degree in a statistically significant manner. The correspondence, however, is dependent on text length and style.

Left-Right Linguistic Asymmetry

In applying Harris's procedure to our test data, we observe that the segments obtained from the R's alone were different from the segments obtained from the L's alone.

Using entropy as a measure of freedom of co-occurrence, and segmenting after each macimum in E_R, we obtain the first-order segments:

come Boots / sai d Ted /
come and ride /
come / and ride / in my wagon /
jump in Boots / sai d Ted /

Placing a boundary before every maximum in E_L, we obtain the segments:

come / Boots sai d Ted /
come / and ride /
come and ride in / my wagon
jump in / Boots sai d / Ted

Combining E_R and E_L, we obtain the segments:

come // Boots // sai d // Ted ////
come and // ride ////
come and // ride //// in my wagon ////
jump in // Boots sai d Ted ////

Notice that the segments following from the E_R's alone are in better agreement with conventional syntactic units than those following from the

E_L's alone. Using just the E_L's we obtain: Boots sai d Ted, come and ride in, Boots said as segments which are not easily identifiable as phrases.

Notice also that fourth-degree boundaries coincide more often with those following from the E_R's than those following from the E_L's. This suggests that there is more information for segmentation in following units as compared to preceding units.

If we examine the phonemic examples in Harris's paper, e.g.

	ʒ	ə	s	a	y	l	o	w	w	o	h	l	z	w	ə	r	ə	p
R	5	29	15	15	28	7	5	29	7	1	8	29	29	7	2	29	9	29
L	24	3	23	10	2	27	5	3	23	16	1	8	18	23	24	5	23	11

The silo walls were up

or

	i	t	k	ə	n	t	e	y	n	z	ə	l	u	w	m	i	n	ə	m
R	10	28	11	11	27	7	6	6	3	28	21	9	2	9	28	4	10	2	28
L	22	19	21	1	1	7	7	3	7	16	22	1	1	1	2	1	5	13	9

It contains aluminum.

we find that the range of following phonemes is larger than that of the preceding. In It contains aluminum, for example, the range of successors is 28-2 = 26 and that of predecessors is 22-1 = 21. Moreover, the R's and L's give different segments. From the R's we obtain

it/ kən/teynz/əluwm/in/əm

From the L's alone we obtain

it/kən/teynz/əluwmin/əm

Another example of different segmentation resulting from following and preceding units is found in Gammon (1963). In this study the linguistic units were Fries' classes, and the sample a text of 5000 words. The second-order segments from the following classes are

If one believes/ that all questions raised/by science/...

The reverse segmentation gives:

If/one believes that all/questions raised by/science .

In this text, the variance of E_R is larger than that of E_L.

A related result is Johnson's (1965) experiment which relates constituent structure to memory blocks. Carried out in reverse order, where Ss are expected to remember preceding words, constituents are not so well isolated.

In our primer data, following morphemes are more variable than preceding morphemes. Using entropy as a measure of diversity,

$$E(E_R) = E(E_L) = 3.18,$$

where E indicates expected value. It may be shown that the expected value of right and left entropies must be equal. But for the variances we find

$$Var(E_R) = 2.33 \quad \text{and} \quad Var(E_L) = 1.98.$$

The difference $Var(E_R) - Var(E_L)$ is significant for this sample.

For the application of our segmentation rules it is of interest that $E_R - E_L$ is more closely correlated with E_R than it is with E_L. And, in

fact, in all the English samples that we have considered, $Var(E_R) > Var(E_L)$. Moreover, in these samples $Cor(|E_R - E_L|, E_R) > Cor(|E_R - E_L|, E_L)$. The variances and correlations are shown in Table 2.

	Length	$Var(E_R)$	$Var(E_L)$	$Cor(\|E_R-E_L\|, E_R)$	$Cor(\|E_R-E_L\|, E_L)$
Ted and Sally	4646	2.33	1.98	.61	.51
Robinson Crusoe	2100	3.63	3.46	.32	.24
Word and Object	900	2.19	1.99	.37	.19

Table 2. Variances and Correlations.

These measures of directional diversity apparently reflect that the language is a unidirectional process. This is to be expected in a suffixing language such as English. We wonder if some directional asymmetry is a property of all natural languages.

One purpose of this paper was to clarify the distributional nature of the word. The assumption has been that a word is a cluster of morphemes. A quantification of what one might mean by "cluster of morphemes." leads to the segmentation rules, and we have presented the results of their application in numerical detail.

The hypothesis that words are clusters of morphemes according to our interpretation is partially verified by the data that have been presented, but the results remain suggestive rather than definitive. Printers' words and distributional groupings are coextensive with a much greater-than-chance frequency. Moreover, in one case at least, there is a close correspondence between the degree of distributional separation of morphemes and the corresponding syntactic boundaries.

An ofttimes unstated assumption in statistical studies of language is that the results would become better if the sample size were larger. This assumption is confirmed, but only in a restricted sense. In the specialized language of the primer Ted and Sally, we used a large sample procedure to eliminate zero-degree segments and obtain a closer correspondence with printers' words. This procedure is applicable to the closed vocabulary of this primer, in which every morpheme is used many times. It would not be applicable to texts where Zipf's law holds, and most morphemes are used only once.

A study of the relationship between segmentation and sample size shows that segments are quite stable and do not change with respect to longer and longer portions of a text. In some cases, of course, larger

samples break up segments which occurred initially for lack of distributional information. The general conclusion is that the distributional freedom with respect to limited contexts may be established from relatively small samples.

With regard to establishing the distributional reality of printers' words, morpheme segments of fixed order do not necessarily approach words as the sample size increases. The distributional clusters which do not correspond to printers' words furnish style indicators. Thus, we have the segments: lookTed, saySally in Ted and Sally; onboard and onshore in Robinson Crusoe; and however and thecaseof in Word and Object. These stylistic groupings show the same strong association that is found between the morphemes occurring within words. These groupings are not necessarily the most frequent in a sample.

The groups onboard, thecaseof, etc. function as compounds in their respective texts. We may speculate about the role of morpheme frequency in the formation of compounds. To use our theory in a predictive sense, we would assert morphemes showing strong association, in the sense we have defined it, operate as compounds.

Our rules enable us to make statements about the relative ease of combination of linguistic units. We have already pointed out that in the Robinson Crusoe sample the, in the context the // ship , shows neutral association, while in the context the //// first, the disassociation is strong. A parallel example, also in Robinson Crusoe, is on where we find onboard, onshore. On the other hand, in the context of the prepositional phrases on us and on them, we find the neutral associations on // we and on // they.

These examples suggest that there are degrees of distributional freedom and that instead of hoping to give an absolute distributional characterization of the word, we should speak of degrees of distributional word-hood. The degree of boundedness of the morphemes of a word is not an absolute property but depends on the corpus containing them, and in addition the context of surrounding morphemes.

Graphemic Grouping

The segmentation rules are numerical procedures for grouping linguistic units. Here we apply these rules to graphemic data. For a graphemic application we compare Ted and Sally and Word and Object. Using letters, we can process much larger samples than we could using morphemes. Relative to the first 16,640 letters of Ted and Sally, we obtain the segments

Come Boots said Ted

In this simple text almost all words can be isolated from letter samples. In contrast, consider the sentence fragment from Word and Object:

What counts as a word as against a string ...

Relative to a sample of 15,889 letters, the second-order segments from maxima in R:

What cou nts asa word asa gain stas tring ...

From maxima in L:

Wh at counts asaw ord asaga ins tast ring ...

Combining the information from the R s and L s we obtain the segments.

Whatcounts asa word asaga insta string ...

The complexity of the text makes a marked difference in the operation of our segmentation rule. We obtain many word boundaries but also asaga, insta. In a text of one syllable words such as Ted and Sally, such combinations do not occur.

A text intermediate to the last two is the lower school reader All Around Me. The segmentation relative to 15760 letters shows an isolation of meaningful letter sequences, which are not necessarily words. The text begins:

Now Whitey was eleven years old, or thereabouts. He had ...

Our segmentation rule gives:

Now Whiteywas eleven y ear s oldor there about she had ...

This text illustrates that the segmentation follows distribution, giving y, s, she, ... , as segments. No punctuation was involved in these letter samples. An introduction of punctuation would give s. he rather than she, but not change the other groupings substantially for a sample of this size.

REFERENCES

Bloomfield, Leonard, "A Set of Postulates for the Science of Language," Lang., 2, 1926, pp. 153-164.

Bloomfield, Leonard, Language, New York, 1933, p. 178

Chomsky, Noam, Word, 12, 1958, p. 217.

Gammon, E., Proc. IX Inter. Cong. of Ling., 1963, pp. 507-13.

Greenberg, Joseph, Essays in Linguistics, Chicago, 1957, p. 27.

Harris, James, Hermes or a Philosophical Inquiry Concerning Universal Grammar, London, 1771, pp. 20-21.

Harris, Zellig, "From Phoneme to Morpheme," Lang., 31, 1955, pp. 190-234.

Hjelmslev, L., Omkring Sprogteories Grundlaeggelse, 1943, p. 66.

Jakobson, R., Actes du IV^{me} Congres de Linguistes, 1938, pp. 133-34.

Johnson, N.F., "The Psychological Reality of Phase Structure Rules", J. of Verbal Learning and Verbal Behavior 4, 1965, pp. 469-475.

Juilland, A. The Word, unpublished MS

Kendall, M.G., The Advanced Theory of Statistics, Vol. 1, New York, 1952, p. 290 et seq.

Khinchin, A. I., Mathemadical Foundation of Information Theory, New York, 1957, pp. 2-4.

Togeby, Knud, "Qu'est-ce qu'un mot?" Travaux du Cercle Linguistique de Copenhague, V, 1949, pp. 97-111.

TEXT SAMPLES

Defoe, Daniel, "Robinson Crusoe, " Beacon Third Reader, Ginn and Co., Boston, 1914.

Francis, N., The Structure of American English, New York, 1959.

Gates, A. I., and Bartlett, M. M., All Around Me, New York, 1957.

Gates, A. I., Haber, M. B., and Salisbury, F.S., Ted and Sally, New York, 1957.

Quine, W., Word and Object, M.I.T., 1960, pp. 13-14.