PAPERS IN STRUCTURAL AND TRANSFORMATIONAL LINGUISTICS
FROM PHONEME TO MORPHEME

0.1. Introduction
The following investigation\(^1\) presents a constructional procedure segmenting an utterance in a way which correlates well with word and morpheme boundaries. The procedure requires a large set of utterances, elicited in a certain manner from an informant (or found in a very large corpus); and it requires that all the utterances be written in the same phonemic representation, determined without reference to morphemes. It then investigates a particular distributional relation among the phonemes in the utterances thus collected; and on the basis of this relation among the phonemes, it indicates particular points of segmentation within one utterance at a time. For example, in the utterance /hiyzkwikər/ He's quicker it will indicate segmentation at the points marked by dots: /hiy.z.kwik.ər/; and it will do so purely by comparing this phonemic sequence with the phonemic sequences of other utterances.

The interest of this procedure is as follows. At present we have distributional procedures for finding the phonemes of a language, and we have procedures for testing morphologically whether any arbitrary segment is or is not a morph, by describing its distributional relations to other segments. We lack, however, any direct procedure for finding the segments which morphological analysis would show to be morphs. We have had to depend upon various indirect methods, or upon distributional or semantic guesses, to tell us what segments seemed worth testing for morphemic status.\(^2\) The procedure described below may serve to fill this gap. It segments each utterance to which it is applied. When we test the resulting segments by the usual morphological methods, we find that most of our procedurally-obtained segmentations occur at the word and morpheme boundaries for that utterance.\(^3\) This procedure can therefore provide the segments for morphological analysis, even though it does not tell us directly that any particular segment of an utterance (e.g. the /ər/ above) is a morph, or a word; and obviously it cannot tell us anything about its meaning. The decision as to morphemic status is made, as before, with the usual morphological methods. And when these distributional methods are applied to our procedurally obtained segments, they

\(^{1}\) See Language, 31 (1955), 190-222.

\(^{2}\) For example, the segment /ər/ in He's quicker is clearly a morph in the sense that it cannot be divided up into words.

\(^{3}\) This is true only if the utterance is long enough to have both word and morpheme boundaries.
not only test the morphemic aptness of the segmentation but also adjust it (whenever it does not fall on morpheme boundaries) to accord with morphological relations. For methodological purposes and for special problems—though certainly not for practical work—this procedure can therefore replace the less orderly search for morphemic segments. In particular, all or almost all word boundaries come out from the procedure.

This completes a chain of procedures covering phonology and morphology: first, the procedure that starts with the sounds of utterances and yields a phonemic spelling of them, then the present procedure which starts with the phonemic spelling of utterances and yields a segmentation of them into phonemic sequences; and lastly, the morphological procedures which describe the structure of the utterance as a distribution of these segments (and in so doing correct the segmentation to obtain better structural elements). The present procedure requires only activities which are already available at its position in the chain: a phonemic representation of utterances, and the known activities of eliciting (asking an informant for utterances) and counting. In particular, it involves no reference to the meaning of morphemes—that is, no knowledge or judgment of meanings or meaning-differences, and no reliance on the speaker’s ability to respond in terms of the meaning or meaning-differences of morphemes.

0.2. Plan of the Paper

In this paper, the method will be presented empirically. One can raise deeper questions about the method: what relation between phonemes and morphemes makes it possible to find morpheme boundaries from a particular distributional relation among phonemes? Why should we have expected the particular relation described here to yield these boundaries? What implications may be drawn from this result, what further investigations suggested? These questions will be touched upon in § 6, but only briefly, because they require an independent discussion, too ramified to serve as an introduction or conclusion to the present procedure (which requires enough space as it is).

The procedure will therefore be described without theoretical introduction. We will simply consider a particular way of counting phoneme distributions; and we will see how the segmentations indicated by this counting procedure coincide with the morpheme boundaries as we know them from morphological analysis. There is first a basic procedure, which gives the bulk of the desired correlations. To this are added several related procedures which increase the correlation between the counting method and morphological boundaries.

The basic procedure (§ 1) is to ask how many different phonemes occur (in various utterances) after the first $n$ phonemes of some test utterance.
will be found that the number of these possible successors to the first \(n\) phonemes varies with \(n\): in the test utterance of § 1, 14 different phonemes occur (in one sentence or another) after the first two phonemes of the test utterance, while 29 occur after the first three. We segment the test utterance at the points where the number of successors reaches a peak.\(^5\) To carry out this procedure, it is necessary that each phoneme be pronounceable. This raises certain difficulties, as in the case of junctures, which are discussed in § 2.

By the side of the basic procedure a number of modifications may be considered. The most fruitful of these are: to carry out the procedure backward from the end of the test utterance (§ 3.1); to consider the insertion of phonemic sequences at each point of the test utterance (§ 3.2); and to include an additional count of how many \((n+2)\)th successors can be found after each \((n+1)\)th successor (§ 3.3). Major additional results are obtained when we consider not only the number of phonemes at each point but also which particular phonemes are the ones that occur there. In all the utterances that begin with /hiyzk/ there are 11 phonemes which ever occupy the sixth place: /l, r, w, y, i, e, æ, a, o, u/. We call this the variety of phonemes after utterance-initial /hiyzk/.\(^6\) In § 4 we correlate the variety of successors at each point, and their number, with the phoneme whose successors they are.\(^7\) With the aid of this we find a periodicity in each sentence. If we segment the utterance on the basis of this periodicity we obtain an even higher agreement with morphemic segments as we know them through morphological analysis. Indeed, this periodicity would suggest to us that sentences can be segmented into morphemic elements, even if we did not know beforehand that such elements exist in language.

How fully all these methods correlate with the boundaries derived from morphological analysis will be briefly considered in § 5. The relation of such methods to linguistic structure is touched upon in § 6. A summary is given in § 7.

1. THE BASIC PROCEDURE: SUCCESSOR COUNTS

We take any utterance \(U\) written as a sequence of phonemes. (For the relevance of phonemic contours and junctures, see § 2.) We collect many utterances which begin with the first phoneme of our test utterance \(U\), and count how many different phonemes occupy the second position in these utterances: these are the successors of the first phoneme of \(U\). Then we collect many utterances which begin with the first two phonemes of \(U\), and count how many different phonemes occupy the third position in these utterances: these are the successors of the first two phonemes of \(U\). And so on down to the end of \(U\). That is to say, for each utterance-initial sequence
of U up to its nth phoneme \((n=1, 2, \ldots \text{ up to the last phoneme of the utterance})\), we count all the different \((n+1)\)th phonemes in the various (associated) utterances that begin with this same sequence of \(n\) phonemes. As we proceed along the phonemes of U, we find that for each \(n\) of U the number of successors, i.e. \((n+1)\)th phonemes in the associated utterances, falls, then rises to a peak, then falls again, rises to a peak again, and so on. At the points at which this number reaches a peak, we place our tentative segmentation of U.

For example, consider the short utterance \(\text{He's clever} /\text{hiyzkle\v{o}r}/\). We collect utterances beginning with /h/: \(\text{When did you come?} /\text{Humans act like simians.} /\text{His ship's in.} /\text{He's out.} /\text{Hell, what's the use?} /\text{Had to, sorry.} /\text{Have you got it?} /\text{Harping on it won't help.} /\text{Hot coffee.} /\text{Her timing's off.} /\text{Hunting's a dumb thing to do.} /\text{Hope for the best.} /\text{Who is it?} /\text{Hook and ladder company.}\)

The first phoneme in all of these is /h/, as in our test utterance. The second is one of the following: \{w, y, i, e, æ, æ, ə, o, u\}. This is the variety of phonemes in second position, in those utterances whose first-position phoneme is the same as in \(\text{He's clever}\); the number of successors for /h/ is 9.

We next collect utterances beginning with /hi/: \(\text{Hip-high in water.} /\text{Hit it back!} /\text{Hickory nuts are still available.} /\text{Hidden meanings were discovered.}\)

And so on. When we continue this collection of utterances beginning with /hi/ we find one or another of the following phonemes in the third position after /hi/: \{p, t, k, d, g, ð, s, ɔ, z, l, m, n, h, y\}. This is the variety of phonemes in third position, in those utterances whose first two positions have the same phonemes as \(\text{He's clever}\); the successor count for /hi/ is 14.

In Table I, two test utterances are compared. In various utterances beginning with /h/ there are 9 different phonemes following the /h/; in those that begin with /hi/ there are 14 phonemes after the /hi/; in those that begin with /hiy/ there are 29 phonemes after /hiy/: \{p/ \text{in Heaps of them,} /d/ \text{in He didn't,} \} and so on. In utterances beginning with /hiyz/ there are again 29 phonemes after /hiyz/, as in \(\text{He's pretending, He's trying to, etc.; in those that begin with /hiyzk/ there are 11 phonemes after /hiyzk/}, \) as in \(\text{He's cranky, He's quiet, etc.}; \) in those that begin with /hiyzkl/ there are 7 phonemes following, as in \(\text{He's clinching it, He's close}; \) and so on.

The second test utterance is \(\text{He's quicker.}\) Up to the first /k/ the string of phonemes is the same as before, so that the associated utterances and the count of successors at each position are the same. When we get to /hiyzkw/ we find, in new utterances beginning with this sequence, that 6 different phonemes follow the /hiyzkw/; 10 phonemes follow the /hiyzkw/ in associated utterances that begin with that; and so on.

In \(\text{He's clever}\) we find peak numbers for the successors to /y/, /z/, /ɪ/, dividing the utterance into /hiy/, /z/, and /kle\v{o}r/. In \(\text{He's quicker}\) we have
<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>9</th>
<th>14</th>
<th>29</th>
<th>9</th>
<th>14</th>
<th>29</th>
<th>11</th>
<th>6</th>
<th>10</th>
<th>28</th>
<th>14</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>J</td>
<td>K</td>
<td>J</td>
<td>K</td>
<td>K</td>
<td>L</td>
<td>N</td>
<td>L</td>
<td>J</td>
<td>K</td>
<td>K</td>
<td>L</td>
<td>J</td>
</tr>
<tr>
<td>L</td>
<td>J</td>
<td>J</td>
<td>K</td>
<td>K</td>
<td>L</td>
<td>N</td>
<td>L</td>
<td>J</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| h | i | y | z | k | l | e | v | o | r | h | i | y | z | k | w | i | k | o | r |
|---|---|---|----|----|---|----|----|----|---|----|----|----|----|---|----|---|---|---|----|----|---|----|---|----|---|---|
| k | D | B | B | B | B | D | B | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B |
| b | B | B | B | B | B | B | B | B | B | B | B | D | C | B | D | C | B | D | C | B | D | C | B | D | C | B | D | C |
| ō | D | B | B | B | B | D | B | B | D | B | B | B | C | B | B | C | B | B | C | B | B | C | B | B | C | B | B | C |
| s | A | B | B | B | B | A | B | B | A | B | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C |
| š | A | B | B | B | B | A | B | B | A | B | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C |
| ć | A | A | B | B | B | A | A | B | A | B | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C |
| t | B | B | B | B | B | A | A | B | A | B | B | C | B | A | B | C | B | A | B | C | B | A | B | C | B | A | B | C |</p>
<table>
<thead>
<tr>
<th>y</th>
<th>D</th>
<th>A</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>D</th>
<th>A</th>
<th>B</th>
<th>D</th>
<th>A</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>ə</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>œ</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>D</td>
<td>B</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.2</td>
</tr>
<tr>
<td>11.2</td>
</tr>
</tbody>
</table>

Notes:

If a point in a given column and row is filled (by A, B, C, or D), the phoneme at the head of the row occurs as the \((n+1)\)th successors to the sequence (above) ending in the \(n\)th phoneme at the top of the column. For example, the sequence /hi/ may be followed by /k/ (as in hickory) but not by /tl/; here \(n=2\) and /k/ is one of the \((n+1)\)th phonemes. The sequence /hiykle/ may be followed by /p/ but not by /tl/; here \(n=7\) and /p/ is one of the \((n+1)\)th phonemes.

N: the number of \((n+1)\)th successors to the sequence that ends in the \(n\)th phoneme beneath the number.

T: the type of successor variety (§ 4) after the same phoneme sequence.

A, B, C, D: the number of \((n+2)\)th successors to the sequence indicated by the column and row in question (§ 3.3); for example, the number of such successors to /hip/ is A, the number of such successors to /hiykl/ is B. The totals below indicate how many \((n+1)\)th successors to the sequence ending in the \(n\)th phoneme at the top of the column have A \((n+2)\)th successors, how many have B \((n+2)\)th successors, and so on. The average is of the B, C, D \((n+2)\)th successors only (notes 17 and 18), per \((n+1)\)th phoneme.
peaks for the successors to /y/, /z/, the second /k/, and /r/, dividing the utterance into /hiy/, /z/, /kwik/, /şr/. Very small rises in number, such as the 8 for the successor to /e/, are not in general a basis for segmentation (§ 5). In this simple example, then, the peaks accord with the boundaries of words and morphemes. We will see below that in some situations the results are not so simple, and auxiliary operations will be introduced to obtain such cuts in the utterance as will correlate with the usual morphemic segmentation.

Sources of Data. The procedure requires a large number of associated utterances sectionally identical with U, some in their first phoneme, others in their first two phonemes, and so on. We could draw these utterances from some written corpus; but the corpus would have to be prohibitively large if we are to be able to find in it, for any U we choose, enough associated utterances for each n of U. The only practicable way of finding the required utterances is to elicit them from an informant, i.e. to ask him for any utterances beginning with /h/, then for any utterances beginning with /hi/, and so on. Such eliciting in no way prejudices our result, since it merely selects, from among the utterances which the informant can make, those whose first n phonemes are the same as in U. It cannot bring up successors which would not occur normally, for example, eliciting cannot make an informant produce an utterance with /v/ as successor to /hiyzk/. And such eliciting does not involve knowledge of morpheme boundaries on the part of the informant or the investigator.

2. PROBLEMS OF PHONEMIC REPRESENTATION

The procedure above measures the change of successor variety as we proceed through the phonemes of an utterance. In order to obtain a clear picture of how the growing string of its phonemes affects the successor variety, it would be natural to consider at each phonemic position all the phonemic distinctions that are made at that position. Such a representation of the successive phonemic distinctions in an utterance is precisely its phonemic spelling. However, simplifications of phonemic systems have led to various departures from the pure successive representation of simple segmental phonemes. Intonations are sometimes marked by contour indicators like /!/ or by tone-numbers at the change points, although the tones are pronounced not by themselves but simultaneously with the vowels and with some of the consonants. Stress is marked before the syllable or over the vowel symbol, but it is pronounced throughout the syllable. Juncture is written as a phonemic mark among the phonemes; it is pronounced as part of the neighboring consonants and vowels. Long vowels (and consonants) are sometimes written
as doubles. And restricted phonemes can be expressed by long components.

For graphic representation it does not matter how these nonsegmental phonemic distinctions are marked, so long as their placing is defined. But in a procedure which measures the effect of successive marks, each phonemic contribution should be marked in the place where it physically occurs. In particular, eliciting from informants (who are not acquainted with the phonemic analysis) requires that each successive phonemic location be pronounced with all the phonemic distinctions it introduces, and that there be no phonemic element which is unpronounceable (such as a zero phoneme), since the effect of such a phoneme upon the successor variety could not be communicated by the informant.

The required phonemic writing is exemplified in Table II. Instead of a segmental juncture phoneme, we write a suprasegmental juncture symbol over the first phoneme which exhibits junctural phenomena: thus /ɔ/ indi-

<table>
<thead>
<tr>
<th>TABLE II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>5 29 8 12 7 29 7 14 14 29 10 7 7 2 1 28 7 9 29</td>
</tr>
<tr>
<td>δ ə n a y t r ē y t s p r i t i y l o w</td>
</tr>
<tr>
<td>29 7 14 9 28 28</td>
</tr>
<tr>
<td>+ + +</td>
</tr>
<tr>
<td>δ ə n a y t r ē y t s p r i t i y l o w</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5 5 1 1 29 29</td>
</tr>
<tr>
<td>+ + +</td>
</tr>
<tr>
<td>δ ə n a y t r ē y t s p r i t i y l o w</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>31 1</td>
</tr>
</tbody>
</table>

The night-rate (nitrates)'s pretty low

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10 13 29 27 16 11 28 10 29</td>
</tr>
<tr>
<td>b a y t æ k s i y</td>
</tr>
<tr>
<td>10 13 29 10 14 5 2 2 29</td>
</tr>
<tr>
<td>+ + +</td>
</tr>
<tr>
<td>b a y t æ k s i y</td>
</tr>
<tr>
<td>2 3 1</td>
</tr>
<tr>
<td>2 3 3 +</td>
</tr>
</tbody>
</table>

By taxi

Notes:

Numbers below the letters indicate tones. (The last line applies to the statement intonation 231 of By taxi/. and the question intonation 233 + of By taxi?/.) Eliciting sequences from an informant would be difficult if we were to treat junctures as segmental phonemes — for instance, if after obtaining the successors to /ɔ/, and the successors to /ð/ we were to ask next for the successors to /ð+/. Accordingly, instead of segmental junctures, we use junctural allophones of neighboring phonemes. A plus over a letter indicates the prejunctural allophone; the following phoneme then appears in its postjunctural allophone.
icates the prejunctural allophone of /p/. Instead of intonation (and stress) contours, we write a suprasegmental tone (and stress) symbol beneath (and above) the first phoneme with which the contour is audible. The contour and juncture phenomena on any following phonemes are unmarked since they depend automatically on the first mark; for example, the continuation of a tone on successive vowels, or the appearance of a post-junctural /ɾ/ after a pre-junctural /t/ in /nátreyt/. Analogously, long vowels are written as single phonemes.

In Table II we obtain, when junctures and contours are not specified, the following peak-segmentation: /dɔ.nayt.reyt.s.priti.y.low./. The phonemic representation here does not specify as between night-rate and nitrate, and the count yields the fuller division, into night-rate. What we get at each point are the successors both for night-rate and for nitrate; after the first /t/ the successors for night, being a peak, include and thus mask the successors for nit-. In the second line, where the junctural allophones of night-rate are specified, we get the same segmentation, correlating with the morphology of night-rate, but with lower successor-counts in rate since the successors due to nit- (and due to night-ray) are absent. And in the third line, where the junctural allophones of nitrate are specified, we get /dɔ.naytreyt.s.priti.y.low./ which correlates with morphology, with no segmentation after /nayt/.

In the second set we obtain, when junctures and contours are not specified, a segmentation /bæt.eks.ɪ./ By taxi. The peak after /t/ is due to bite; this morphemic division would not satisfy the remainder of the utterance, but that is not known when we are eliciting successors to /bayt/, and in any case that would be a morphological rather than a phoneme-count consideration. The peak after /s/ is due to by tax. In the next line we take the counts both for By taxi and for By taxi?, specifying juncture and contours in each case. The counts for both cases are the same, and correlate with morphology: /bæt.eksɪ.ɪ./.

Operating without Juncture. These results suggest that when we apply the successor count to utterances in a complete phonemic writing, we get a segmentation which agrees well with morphological analysis. Nevertheless, there are reasons for using a simplified phonemic representation which does not specify junctures and intonation or stress contours. One might ask, How do we know that any segmentation obtained without specifying junctures or contours would also be obtained in the real utterance, which specifies junctures and contours? In particular, could it perhaps introduce unjustified correlations with morpheme boundaries, i.e. segmentations which agree with morphological analysis but which would not be obtained if we had specified the junctures and contours? To answer this, we first note from Table II that
segmentations which correlate with morpheme boundaries can be obtained both with and without specifying junctures and contours. Indeed, when junctural phenomena are specified, the segmentation accords all the better with morphological analysis (cf. also Summers, we’re out in Table IV). Hence any correlation with morphology that is obtained without specifying junctures or contours appears to be valid as far as it goes, since it would in general also be obtained when these are included.⁹

One might then ask further, Granted that the segmentation obtained without specifying the junctures is generally similar to (but not quite as good as) that for the complete phonemic writing, why nevertheless should we wish to omit the junctures and contours? The answer is as follows. First, many utterances have alternant forms with and without juncture (e.g. after a). Second, informants who are not linguistically trained usually cannot maintain particular junctural and contour allophones as they repeat the sectional strings of phonemes, e.g. they cannot repeat /səmərˈzwiːr/ of Summers, we’re out as distinct from the /səmərˈzwiːr/ of Summers, we rent a cabin, even though the juncture is different. Third, not all phonemes are equally audible, and their presence in an utterance is not equally definite. In particular, junctures are often less audible; and if we are given a sentence pair which is distinguished by a juncture, we often mishear or are uncertain. Hence a test which counts the effect of juncture equally with that of other phonemes is hard to carry out.¹⁰

In view of this, it is permissible and preferable to use just the segmental phonemes for most test utterances in making successor counts. As to alternative analyses of the segmental phonemes, it matters little which one is used, so long as the same analysis is used for the test utterance and all associated ones. For the segmentation is decided not on the basis of the actual successor count, but on how the count rises and falls; and this is more or less the same no matter what the phonemic analysis.¹¹

3. MODIFICATIONS OF THE PROCEDURES

Various modifications of the basic procedure can be devised. The three modifications presented below yield segmentations that agree more closely with morphological analysis.

3.1. Predecessor Count (Backward)

The first modification is to carry out a predecessor count, going backward from the end of the utterance, similar to the successor count forward from the beginning. This means taking the last phoneme of a test utterance, then its last two, then its last three, and so on, and asking in each case for various
utterances that could end with this sequence. We count the predecessors in the associated utterances, i.e. what different phonemes occur in \((n+1)\)th place counting backward before these utterance-final \(n\) phonemes. We segment the test utterance before each peak; the agreement with morphology can be seen in Tables IV and VII.\(^{12}\)

In English and many other languages, there are three main situations in which a backward test corrects the results of a forward test. The first is where we fail to get a peak before a morpheme because it is in grammatical agreement with some preceding morpheme. For example, in \textit{It disturbs me} we get only 3 successors to /\textit{itdistorb}/, because -\(s\) is always present here after \textit{it} unless -\textit{ed} or -\textit{ing(ly)} occurs; in \textit{They disturb me} that is not the case, and we get 29 successors (and a segmentation) after /\textit{deydistarb}/. In such cases, when we test the utterance backward, we find 18 predecessors (indicating a segmentation) before /\textit{zmiy}/.

The second situation is when we fail to get a peak after a morpheme which is rather limited in its distribution in respect to what follows it. For example, in \textit{Let me qualify this} we get only one successor to /\textit{letmiykwal}/, namely /\textit{j}/: \textit{qualify, qualitatively at least ..., etc}. But when we work backward, we find about 13 predecessors before /\textit{ifayb}/, suggesting a segmentation here. A special case arises when a morpheme has a morphophonemic form of limited distribution. In \textit{He's a dramatic speaker}, we find only one successor /\textit{t}/ after /\textit{hiyodramet}/, and only one successor /\textit{i}/ after /\textit{hiyodramet}/, since \textit{dramat} as alternate of \textit{drama} occurs only before -\textit{ic} (and -\textit{is personae}). But backward, we find about 15 predecessors (yielding a segmentation) before /\textit{ikspiyk}/: \textit{She's a terrific speaker, forensic}, etc.

The third and most important situation is that of sectional homonyms (§ 5), where the first part of a morpheme is identical with some whole morpheme: e.g. \textit{They left} has a peak after /\textit{deyl}/ because this is homonymous with \textit{they'll}; but when we go backward we get no peak before /\textit{eft}/, only before /\textit{left}/.

The backward operation is then no closer an approximation to morpheme boundaries than is the forward\(^{13}\); but it is a check on the forward operation. In many cases the two will yield peaks at the same points. These will usually be points of morpheme boundary for the test utterance; but not necessarily, for we may be dealing with a case where both the beginning and the end of the stretch happen to have homonyms.\(^{14}\) In some cases, one direction yields a peak at a given location where the other does not. Then either the peak is wrongly due to a sectional homonym, or the lack of a peak is wrongly due to a directionally limited morpheme or alternant. The decision among these possibilities can be made only by morphological tests or with the procedure of § 4.
3.2. Insertion

The procedure of §1 is sometimes inadequate for morphemic correlation chiefly because it measures the dependences of successor count or variety only in respect to what precedes. Carrying this procedure out backward yields, in addition, the dependences on what follows in the utterance. Even the sum of these two operations will not give perfect morphemic correlation, because it has to count the dependence due to each side separately. Morphemes, however, are set up in the grammar on the basis of their relation to a whole sentence. If we wish to find divisions that will correlate more fully with the morphemic boundaries, and to do this by a constructional procedure of counting phonemic variety instead of by the morphological procedure of comparing utterances and substitution, we have to replace the successor and predecessor operations by a single operation of insertion.

We take a test sentence, and insert between the $n$th and $(n+1)$th phonemes any phonemic sequences (containing whatever morphemes or morpheme parts) such that the total constitutes an utterance of the language. For example, if our test sentence is *This is new* /ðisisiznyw/, we insert between /ð/ and /i/ such phonemic sequences as /œœæl/ making /ðœœælisisiznyw/ *The chalice is new*, or /æθ/ to make *That hiss is new*, or /owzpiypælsedœœæl/ *Those people said the chalice is new*, and so on. Then we insert other sequences between /i/ and /s/: /yzmarksšowðætðobak/ making *These marks show that the box is new*. Then we insert fitting sequences between /s/ and /i/: /buk/ making *This book is new*, /[ilnɔs]/ *This illness is new*, /[meyk]/ *This make is new*, /[tuw]/ *This too is new*, /[oʊərwen]/ *This other one is new*, etc. And so for every other position in the test utterance. We now count the end-variety of the insertions – how many different first phonemes and how many different last phonemes there were in all the insertions at a given position. Where this count is at a peak we introduce a segmentation. In the example above, we have an end-variety peak for the insertions between /ðis/ and /iznyw/.

In cases of agreement, limited bound morphemes, and morpheme alternants, the insertion end count shows at a given location how the location has different degrees of morphemic independence in the two directions. In the case of homonyms, the interference is reduced by the fact that each insertion has to fit into the whole utterance. Even when we have homonyms both forward and backward, as in *The silo walls were up*, the undesired peak will often be kept out by the fact that the insertions would have to fit both homonyms at the same time.

The method of insertion is, however, considerably less convenient and direct for informant testing than the forward or even the backward procedures.
3.3. Counting the \((n+2)\)th

One important modification gives a closer approximation to the morpheme boundaries. This is to count at location \(n\) not only how many different \((n+1)\)th phonemes occur, but also how many different \((n+2)\)th phonemes occur for each \((n+1)\)th.\(^{15}\)

What can this additional information show about the \(n\)th place? To answer this, we consider whether there are any regularities in the sequence of successor counts. If there are any connections between neighboring counts, we might learn something about the \(n\)th place not only from its own successor count, but also from the next successor count, at the \((n+1)\)th place. In looking over all the tables of data, we see that the successor counts generally decrease as we proceed through the phonemes from one peak to the next, except for slight rises at some of the vowels: *He'll admit my family* has the sequence of counts \(9, 14, 29, 29, 18, 12, 4, 2, 29, 8, 12, 28, 10, 13, 2, 3, 1, 1, 1, 28; ... a dramatic* has \(22, 13, 7, 2, 1, 1, 1, 2, 1, 1, 1, 1, 28\). We can draw some empirical conclusions from the results. For example (in English), if the \((n+1)\)th phoneme is a consonant and has around 10 successors, there is likely to be a peak (i.e. a morpheme boundary) after the \(n\)th phoneme. If the \((n+1)\)th phoneme has only 1 or 2 successors, it is likely that there is no peak (and no morpheme boundary) after the \(n\)th place, i.e. the \(n\)th and \((n+1)\)th phonemes are together inside a morpheme.\(^{16}\)

The regularity is statistical: it may not appear in the successive counts of a particular test utterance. But we can bring its effect into play by recording for each \(n\)th phoneme of our utterance not only how many successors can occupy the \((n+1)\)th place (in the associated utterances), but also how many \((n+2)\)th successors each of these successors could have there, i.e. how many different phonemes can occupy the \((n+2)\)th place for each phoneme in the \((n+1)\)th place in the associated utterances. For this purpose it is not necessary to give the exact number of the \((n+2)\)th phonemes, as was done in note 15. Because of the numbers which are characteristic for each inter-peak place, as indicated for English in note 16, it suffices to record whether they are, say, around 28 (category A) or around 15 (B), or around 6 (C), or around 2 (D).\(^{17}\) Then we find that if two phonemes are inside the same morpheme, the second will have fewer of its \((n+2)\)th successors in the higher categories and more in the lower.

In Table III, we see a correlation between the distribution of A, B, C, D and morpheme boundaries (as these are known from morphology). When there is a word peak at \(n\), there are usually 29 successors (initial in the next word); and usually of these, 20 have 8 to 18 successors (as second phonemes of the word), 8 have 5 to 7 successors, 1 (namely /u/) has 1 successor. When
### TABLE III
Count of \((n+2)\)th phonemes for each \((n+1)\)th phoneme

<table>
<thead>
<tr>
<th></th>
<th>It disturbed ...</th>
<th>... family</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>6 2</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1 20 5 1 4</td>
<td>20 1 1</td>
</tr>
<tr>
<td>C</td>
<td>1 8 4 4 11 2</td>
<td>8 7 8</td>
</tr>
<tr>
<td>D</td>
<td>8 1 10 7 3 1 1 1</td>
<td>2 5 1 3 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>16 28 10 17 22 5 1 1 3 28</td>
<td>10 13 2 3 1 1 1 28</td>
</tr>
</tbody>
</table>

... left

<table>
<thead>
<tr>
<th></th>
<th>...</th>
<th>Some ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>7 3</td>
<td>8 2 20</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>6 8</td>
</tr>
<tr>
<td>D</td>
<td>13 1</td>
<td>1 9 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>7 16 1 29</td>
<td>15 14 29</td>
</tr>
</tbody>
</table>

He's a dramatic ...

<table>
<thead>
<tr>
<th></th>
<th>h i y z æ d r æ m æ t i k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>8 1</td>
</tr>
<tr>
<td>B</td>
<td>5 20 20 15 6 3</td>
</tr>
<tr>
<td>C</td>
<td>3 2 8 8 7 5 2</td>
</tr>
<tr>
<td>D</td>
<td>14 1 1 2 2 2 1 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>9 14 29 29 22 13 7 2 1 1 1 1 28</td>
</tr>
</tbody>
</table>

Notes:
- Q: the total number of \((n+1)\)th successors to the \(n\)th phoneme above.
- A: Of the total in Q, so many had about 28 \((n+2)\)th successors in turn.
- B: Of the total in Q, so many had about 15 \((n+2)\)th successors in turn.
- C: Of the total in Q, so many had about 6 \((n+2)\)th successors in turn.
- D: Of the total in Q, so many had about 2 \((n+2)\)th successors in turn.

Some of the successors have about 28 successors in turn (A), that means that there can be a word boundary after these successors. Aside from these cases of A, we find that as we proceed inside a morpheme the average number of \((n+2)\)th phonemes per \((n+1)\)th always decreases. In some the total number of \((n+1)\)th successors to \(n\) decreases only from 15 to 14; but the average of \((n+2)\)th successors per \((n+1)\)th decreases clearly, from 8 B + 6 C + 1 D to 2 B + 9 D.

Results

We have found that the \((n+2)\)th average decreases monotonically in most inter-peak stretches, and that the first, second, and third places after a peak (or utterance beginning) have characteristic \((n+2)\)th averages. Now we can...
use this information in checking on some peculiar counts (chiefly those due
to vowel-consonant differences, or to somewhat limited bound morphemes,
or to certain morphophonemics).

For example, in family there is a slight peak in the \((n+1)\)th count from
10 after /f/ to 13 after /fæ/; but the \((n+2)\)th average falls from 8 B+2 D
to 1 B+7 C+5 D. In left there is a bigger peak in the \((n+1)\)th count, from
7 after /l/ to 16 after /le/; but the \((n+2)\)th average falls clearly from 7 B to
3 B+13 D. In both cases, then, the peak for the \((n+1)\)th count does not
indicate a morpheme boundary. On the other hand, in disturbed we have a
peak of 22 after /dis/ which is not relatively larger than these others (though
it reaches a higher number); but the \((n+2)\)th average shows that there is a
segmentation here: it first drops (within dis-) from 5 B+4 C+1 D to 1 B+
4 C+10 D, and then rises to 4 B+11 C+7 D to fall again to 2 C+3 D and
on to just 1 D. This means that more successors to /dis/ than to /di/ had a
high number of successors in turn, which is due to the fact that many
successors to /dis/ were the initial phonemes of new morphemes and there-
fore had the relatively high successor variety which characterizes the early
part of a morpheme.

Finally, in He's a dramatic... we see another 22 after /hiyza/. This 22 is
not a peak, but the sequence 22, 13, 7, etc. begins with suspiciously high
numbers. When we consider the \((n+2)\)th average for these 22, we find that
it is little lower if at all than the \((n+2)\)th average of the two preceding
positions, which were word boundaries. For /hiyz/ we have 20 B+8 C+1 D,
while for /hiyza/ 15 B+7 C; thereafter there is a clear drop to 6 B+5 C+2 D
and on to 3 B+2 C+2 D and to 2 D. We judge then that in respect to the
\((n+2)\)th average /hiyza/ appears like a word boundary; and morphological
investigation shows that in this position /a/ and a morphophonemic alternant
/ən/ fill out between them the position of a morpheme. For contrast, note
that in we're out (Table IV) the \((n+1)\)th count is rather similar: 29 after
/wiyr/ and 20 after /wiyræ/. But when we check the \((n+2)\)th average there,
we find that it drops sharply from the 20 B+8 C+1 D for the 29 successors
of /wiyr/ to (4 A+) 2 B+7 C+7 D for the 20 successors of /wiyræ/. We
therefore do not suspect a morpheme boundary after /wiyræ/ as we do after
/hiyza/ above.

4. SEQUENCES OF SUCCESSOR VARIETY

The operation of § 1 counts the successor variety at each \(n\). The location of
cuts depends only on the peaks of the successor counts. The modifications
of § 3 deal only with the counts. We now consider a related operation, which
uses the procedure of § 1, but differs in one respect. Instead of merely count-
ing how many different successors there are at \( n \), we list the specific ones which are there. The importance of this new step is that we find certain recurring sequences of successor varieties.

Recurring sequences of exactly the same successor varieties are rare; but it is possible to group certain similar varieties into classes in such a way that repeating sequences of those classes are common. For example, in *He's clever*, /h\( iy/ \) and /hiyz/ both have the same number of successors, 29, and the variety which makes up this number is identical: all the phonemes except /\( ð/, \( η/ \) – i.e. all the phonemes that occur at utterance initial. Let us call this variety J. Again, in Table I, /hiyzkle\( v/\), /hiyzkwik/, and /hiyzkwik\( æ/\) all have 28 successors, again identical: all the phonemes except /\( ð/, \( η/, \( u/\). We will call this variety J 28; often it will be sufficient, as we will see, simply to call it J. After /h/ we find 9 successors – the 7 vowels and /\( y/, \( w/. After /hiyzk/ we find 11 successors – the above and /\( l/, \( r/. In each case these are all the phonemes that would occur after the \( n\)th phoneme, /h/ and /k/ respectively, when these consonants are first in an utterance. Let us call this variety K; when a sequence ending in a consonant \( x \) has the K variety of successors, that means it has the same successors that \( x \) has when \( x \) is in utterance initial. K after phoneme \( x \) therefore indicates the phonemes that occur in second place in all utterances beginning with \( x \); it almost always includes all the vowels, and usually also a few consonants, depending on the phoneme \( x \).

We further notice that /hiyzkl/ has 7 successors, just the vowels, which is what follows /l/ after utterance initial consonant, and /hiyzkw/ has only 6 of the vowels, as does /w/ after an utterance-initial consonant. We can call these varieties \( K' \): when a sequence ending in consonants \( xy \) has the variety of successors \( K' \), it has the same successors that \( xy \) have when they are an utterance-initial cluster. We may compare this with /hiyzkle\( v/\), which has only one successor, and is thus quite different from initial /\( v/\).

For each language we notice that not all varieties of successors occur, at one point or another; only certain classes or types of varieties (e.g. J or K) occur as successors, at least with any frequency; and some particular classes (e.g. J) occur very frequently. What is even more important, only certain sequences of these successor-variety types (J, K, etc.) occur. We have here, then, a situation which is well known in phonology and morphology; and we can carry on these successor varieties the same kind of linguistic operations that we carry out on allophones or morphemes. We find that it is possible to group certain similar varieties into classes, like the class K above, in such a way that regularities can be asserted about the occurrence of these classes: a certain one always occurs with a certain other; or one can be treated as a positional alternant of another; or most of the stretches from peak to peak exhibit a few fixed sequences of these classes. And if in
some utterance the peaks do not come at the boundaries of these regular sequences of variety classes, we may find that the stretch covered by the regular sequence of varieties correlates better with morphological boundaries than does the stretch from peak to peak.

4.1. For English, we set up empirically the following classes of successor varieties:

J: all phonemes except /ɛ, ɪ/.

K (after x): those phonemes which occur after utterance-initial consonant x – the vowels and from 0 to 8 consonants, depending on x.

K' (after [z]xy): those phonemes which occur after the utterance-initial consonant cluster (z)xy – the vowels and from 0 to 2 consonants.

L: any number of consonants only, as successor variety to vowels or post-vocalic /w, y, h/, or as successor variety to a consonant which precedes a consonant whose successor variety is J.

M: a small number (usually 10 or fewer) of consonants and vowels, not satisfying the conditions for K, K'; usually we say that x has successor variety M only if M contains some phonemes which are not included in K.

N: vowels only, usually 4 or fewer.

In terms of these, we find that in inter-peak stretches, i.e. after one successor peak (or from the beginning of an utterance) up to and including the next peak, the sequences can generally be expressed by the following formula:

\[
\text{Syllabic character of the } n\text{th phoneme: } \begin{cases} \text{consonant} & \text{vowel } (w, y, h) \\ \text{K (K' K')} & L \\ \text{MN L} & (w, y, h) \end{cases} \text{ (consonant)}
\]

All parenthetic sections can be independently omitted. The square brackets may yield one or several consonants (of which only the last or none has successor N) before the vowel, and the whole sequence in square brackets may be repeated several times. The only case where the first vowel with successor L is lacking is the case where the first vowel in the inter-peak stretch has successor variety J, i.e. is the last phoneme of the stretch; this occurs in a, the. The chief additional observation is that the number of phonemes in each variety falls sharply as we go through the inter-peak stretch; and the numbers in N, K, K' (which contain vowels) are smaller than the numbers in L, M for the same position between peaks.

That this summary satisfies most of the inter-peak stretches can be seen from the data in Table IV. What is of interest, however, is that when apparent irregularities are interpreted as special cases of this very sequence, we obtain a closer correlation with the usual morpheme boundaries than the numerical
TABLE IV

This table shows the results of §1, §3.1, and §4. The number above a phoneme indicates the number of \((n+1)\)th successors to the utterance from the beginning up to and including that phoneme (the \(n\)th); the capital letter above it indicates the type of successor variety after that phoneme. The number below a phoneme indicates the number of predecessors to the utterance end, from that phoneme to the close of the utterance; the capital letter below it indicates the type of predecessor variety before that phoneme. J yields a segmentation after the phoneme below it; Z yields a segmentation before the phoneme above it. Dots are placed after each J (in French after I, E, F) and before each Z to show the segmentation from each direction, so as to facilitate comparison with the dots placed between phonemes. The latter dots indicate (not phonemic junctures but) the division of this utterance as it is established by the usual morphological methods: a colon marks a division between words; a single dot marks a division which has a bound morpheme (or morpheme variant) on at least one side. Small differences in phonemics and in successor numbers are due to differences between informants.

Small raised letters refer to the notes after the table.

<table>
<thead>
<tr>
<th>K L J</th>
<th>K L N L J J K L J</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 2 29 10 3 1 1 28 28 7 9 28</td>
<td></td>
</tr>
<tr>
<td>y u w : b a ð e r d : m i y</td>
<td></td>
</tr>
</tbody>
</table>

*You bothered me*

<table>
<thead>
<tr>
<th>L J K L J K L L M</th>
<th>J K L J</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 29 10 17 22 5 1 1 3 29 7 9 28</td>
<td></td>
</tr>
<tr>
<td>i t : d i s t o r b d : m i y</td>
<td></td>
</tr>
</tbody>
</table>

*It disturbed me*

<table>
<thead>
<tr>
<th>K L J</th>
<th>L J K L J J L J J</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 14 29 17 29 29 7 11 29 29 22 28 29</td>
<td></td>
</tr>
<tr>
<td>s ð m o r z : w i y r : æ w t</td>
<td></td>
</tr>
</tbody>
</table>

*Summers we’re out*

Without juncture after the second /æ/. With juncture, the last three phonemes have L 20, L 4, J 29, and there is no homonym at //æw//

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9 14 29 29 18 12 4 2 29 8 12 28 10 13 2 3 1 1 28</td>
</tr>
<tr>
<td>h i y l : æ d m i t : m a y : f æ m i l i y</td>
</tr>
</tbody>
</table>

*He’ll admit my family*

The M 12 suggests a possible boundary after /æd/.

<table>
<thead>
<tr>
<th>K L L J J L L J L J K L J K K</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 3 6 29 29 11 1 28 12 28 9 14 20 8 1</td>
</tr>
<tr>
<td>d o h g z : a h r : i n d i s p y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L L J L M N L J K K’ L J L J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 28 1 2 1 1 28 11 6 9 28 2 28</td>
</tr>
<tr>
<td>u w t æ b l i y : k w i k o r</td>
</tr>
</tbody>
</table>

*Dogs are indisputably quicker*

<table>
<thead>
<tr>
<th>K L J J K K’ L N L N M L J K L M L</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 14 29 29 22 13 7 2 1 1 1 1 1 28 15 12 7 4 28</td>
</tr>
<tr>
<td>h i y z : æ d r æ m æ t i k : s i n æ r</td>
</tr>
</tbody>
</table>
He's a dramatic singer

K K' L J * K L' m J * K L J * K L L J * L J
9 5 1 29 10 19 28 8 12 28 5 4 1 29 11 28
h w o t : d i : d : h i y : θ i n k : ov
22 4 7 18 23 1 3 9 12 4 22 15 13 12 23 6
Z S S Zr Z S T S T Z S T T T Z T

What did he think of?

The sectional homonym ink of has not enough predecessors to make a peak backwards.
T 3 is low because few verbs occur before he think of. For a cut at S 9 before /h'iy/, see § 4.2

K K' L J * K L' m J * K L J * K L L
9 5 1 29 10 19 28 8 12 28 5 4 1
h w o t : d i : d : h i y : θ i n
20 1 1 8 15 1 3 10 6 4 17 3 4
Z S S T T S T T S T Z S T

J * L Jh * L J * J K L J * K L L J
29 11 28 1 28 7 2 4 24 7 3 1 28
k : ov i n : L z : w o r : f o hr
11 22 3 8 8 17 23 24 3 23 20 6 3
T Z T T S T Zr Z Z T Z T T T T

What did he think ovens were for?

S 10 before /h'iy/ can be suspected of being a low Z. The grammatical position and the selection permit few words before this he

L J * K L J * K L L L J * Jh * M L L Jh * L M L J
10 28 11 11 27 7 6 6 3 28 21 9 2 9 28 4 10 2 28
i t : k o n t e y n z : 5 1 u w m i n 5 m
22 19 21 1 1 7 7 3 7 16 22 1 1 1 2 1 5 13 9
Z Zrh Z S T R S T T Z Z T T S T R S R S T

It contains aluminum

The M 10 after /in/ can be taken as a J which is very low because of selection; it comes from the homonymous It contains a loom in ... The S 13 before /am/ may be taken a Z low because of selection; many of the 13 predecessors come from words ending in -um (which might be morphologically analyzed as a separate morpheme). The R 7 before /teynz/ raises the question of a morpheme boundary here (§ 4.2 end). See § 4.1

K L J * Jh * L L J * K L J * L L M L L L N L L J
6 4 29 29 13 1 28 14 15 28 18 10 2 4 1 1 2 1 28
δ e y : l e f t : s a m : 5 1 u w m i n 5 m
24 12 4 24 2 3 18 23 7 7 22 1 1 1 2 1 1 5 13 9
Z S T Z S T T Z Z T T S T R S R S T

They left some aluminum

K L J * K L L J * J K L L J
6 4 29 14 1 8 1 29 29 14 13 28
δ e y : s w a y p t : s a m
23 10 4 23 1 3 3 9 18 23 8 9
Z S T Z S T T Z Z T T S T
They swiped some
With voiceless allophone of /w/ after /s/, hence no homonymy with wipe, going backwards

\[
\begin{align*}
J \cdot K \cdot K' L & \quad J \cdot K L L N L J \cdot L J \cdot K L N^m L J
22 & \quad 10 7 10 28 15 8 8 2 1 28 9 1 8 1 1 28
& \quad \text{with vowel allophone of /w/ after /s/}
K & \quad 23 3 9 8 22 4 3 27 17 4 24 4 24 9 14 24 8
Z & \quad Z R S T \cdot Z S T \cdot Z^h S R \cdot Z \cdot Z^h \cdot Z S T \cdot Z T
\end{align*}
\]

A brick silo is better

\[
\begin{align*}
K & \quad J \cdot J \cdot K L J \cdot K L J \cdot K L J \cdot J K L J L J
5 & \quad 29 15 12 28 7 5 29 7 1 8 29 29 7 2 29 9 27
& \quad \text{with vowel allophone of /w/ after /s/}
\delta & \quad \sigma: s a y l o w : w o h l \cdot z : w \sigma r : o p
24 & \quad 3 23 10 2 27 15 3 23 16 1 8 18 23 24 5 23 11
Z & \quad Z S Z S T \cdot Z^i S T \cdot Z S T' T \cdot Z^i \cdot Z^h T \cdot Z T
\end{align*}
\]

The silo walls were up

\[
\begin{align*}
K & \quad J \cdot K L J^h \cdot K L J \cdot K L J \cdot K L M J \cdot L
5 & \quad 29 15 12 28 7 5 29 7 5 11 28 10 11 9 14 5
& \quad \text{with vowel allophone of /w/ after /s/}
\delta & \quad \sigma: s a y l o w : l a y k : b i l d \cdot i
23 & \quad 23 22 1 4 9 10 3 22 7 3 13 23 3 4 11 22
Z & \quad Z^h \cdot Z S T' R S T \cdot Z S T' T \cdot Z S T' T \cdot Z
\end{align*}
\]

The silo-like building is a water tower

For the final Z sequences, see § 4.2 end

Notes:

- A cut is lacking here because the following morpheme is in grammatical agreement (in the direction in which we are going). The cut would be obtained in the reverse direction.
- A cut is lacking because our morpheme occurs in this grammatical position with only a few successors. The cut would be obtained in the reverse direction.
- An inappropriate cut is obtained here because the first part of the morpheme (proceeding in the direction in which we are going) is homonymous with a whole morpheme. When our morpheme occurs in other grammatical positions this cut may be avoided; and it is avoided in the reverse direction.
- Same as the preceding, except that the remainder of our morpheme is homonymous with still another morpheme. For this reason the cut is obtained also in the reverse direction.
- A cut is lacking because our morpheme occurs here in a morphophonemic form that has only one or a few successors. The cut would be obtained in the reverse direction.
- A cut is lacking because the bound morpheme here has too few successor morphemes. The cut would be obtained in the reverse direction.
peaks had afforded. Our cuts will now be made not simply at peaks, but at the points where the above sequence ends, namely after J.

For example, one successor variety X which seems not to belong to any of these classes is the 22 consonants which we obtain after /ə/ in some (not all) of the instances where /ə/ comes at utterance beginning or after a peak (in /əbrək/ a brick but not in /səməlu.wəmənəm/ some aluminum). We notice that this X appears before KK’ (a brick, a dramatic); in fact all the sequences of variety classes which occur after X are sequences which occur after J. We notice further that X includes all the consonants except /ʒ, η/, which are precisely all the consonants of J. Finally the (n+2)th variety for each of the 22 members of X is the same as the (n+2)th variety for the corresponding 22 consonants in J. Hence we tentatively include the X successor variety of /ə/ as a variant of J, thus assuring a segmentation after the /ə/ which has X successors. This assignment is made independently of any complementary relation to an, but it will be supported when we find that there is a sequence an whose successors are precisely the vowel members of J. In contrast, the 22-consonant variety after /wirə:/ of we’re out is not classified in J, because it includes /ʒ, η/, and because its (n+2)th average (4 A + 2 B + 7 C + 7 D) is far smaller than the (n+2)th average of J (27 B + 1 C + 1 D).

In It contains aluminum we find the following successors after the successive phonemes of /əlu.wəmənəm/: J 21, M 9, L 2, L 9, J 28, L 4, M 10, L 2, J 28. Now the inter-peak sequence MLLJ does not otherwise occur, but the sequence LMLLJ does (see the formula above). If we wish to interpret this case in terms of our formula, we can regard it as the sum of two morphemic possibilities: One of these would take /ə/ as a separate morpheme, with successor J; then the next phoneme /l/ is the first of a word (with roughly the same successors as when it is the first of an utterance) and hence has successor K. The sequence would be JKLLJ. The other possibility would take /ə/ as the first phoneme of a longer word (since it occurs after a peak) with certain successors L (adrenalin, ascorbic acid, allusions, etc.); then the next phoneme /l/ is the second of the word and has a smaller number of successors M (ulterior, albinos, alert..., etc.). The sequence would be LMLLJ. Both possibilities would fit after It contains. But if a string of phonemes can yield both JKLLJ and also LMLLJ (if the first phoneme may constitute a whole morpheme or homonymously the beginning of a longer morpheme), then the sum of both JKLLJ and LMLLJ for that string of phonemes will be JMLLL. For J includes L, so that we cannot observe the presence of L in the total successors to /ə/; and M includes K, so that we cannot observe the presence of K in the total successors to /əl/. We therefore interpret the aberrant JM... as being merely a resultant of JK... and LM... This means that there are two (homonymous) morphemic analyses here, a l... and al...;
and this correlates with the morphological analysis. Whether we locate a segmentation here depends on which analysis fits the remainder of the utterance. In this way we can recognize certain homonymous alternatives, which are added together (and therefore hidden) by the successor counts. Some but not all of these cases could also be recognized directly if we specified the junctural allophones.

When a small successor peak occurs, the sequence of variety-classes established here may help in deciding whether a cut should be placed after the peak. For example in a water tower we find the following class and number of successors for /swɔːtər/: J 21, K 7, L 2, L 8, N 1, L 1, J 27. In considering the small peak at L 8, we ask whether it might not be taken as an unusually low J. If we did, then our next inter-peak stretch would be N 1, L 1, J 27, which would be an otherwise unknown type of inter-peak sequence. And if we try to correct this by saying that the N 1 is just an unusually low case of K, yielding the known sequence KLJ, then we face the unusual situation of having an inter-peak stretch begin with such a low successor count as 1. Hence we make no cut between L and N here. In contrast, an unusually high M, even if it does not reach a peak, will often be found to correlate with the morphologically establishable morpheme boundaries, as in admit /ədmit/ with L 18, M 12, N 4, L 2, J 29.

4.2. Similar sequences can be set up for the classes of predecessor varieties, going backward. Most inter-peak stretches turn out to have the same sequence, if we group the predecessor-varieties into the following classes:

Z: All the consonants and /ɔ/, except that /ð/ is often lacking, /ʃ/ extremely rare (depending on dialect), and /h/ almost only after /o/. /ɔ/ is frequently lacking, especially in what will turn out to be certain grammatical positions.

T (before x): all or most vowels and a few consonants, the consonants being some (not necessarily all) of those which appear before x when x is the utterance-final consonant.

T': only the vowels, as predecessors to consonants (including /w, y/).

S: consonants, predecessors to vowels.

R: some vowels and some consonants (including some not included in T).

The sequence of these before a peak (or from utterance end) up to and including the preceding peak usually follows this formula:

Syllabic character
of the nth phomeme:
 Its predecessor variety \[ Z \begin{pmatrix} \text{vowel} \\ S \end{pmatrix} \begin{pmatrix} \text{consonant vowel} \\ T' \end{pmatrix} \begin{pmatrix} \text{consonant} \\ R \end{pmatrix} \begin{pmatrix} \text{consonant} \\ S \end{pmatrix} \begin{pmatrix} \text{consonant} \\ T' \end{pmatrix} \begin{pmatrix} \text{consonant} \\ T \end{pmatrix} \]
Parenthetic sections are independently omissible. In the square brackets there may be none or several R, and one or no T'; and the parenthetic sequences may be repeated. In the few inter-peak stretches which end on /ɔ/ instead of a consonant, the predecessor of /ɔ/ is a restricted S instead of T. The numbers in each predecessor class generally go down as we go backward in the inter-peak stretch and in the utterance.

Here again we can analyze irregular phenomena by interpreting them in terms of this sequence. Various rarer predecessor varieties can be assigned to the above classes on the basis of the sequences in which they occur. For example, a variety Z', of all voiced phonemes except the sibilants, occurs often before /z/ when /z/ is before a peak or at utterance end; and a variety Z", of all voiced phonemes except /d/, occurs often before /d/ in similar circumstances. We notice that the sequence before it often starts (going backward) with T, rather than T'; thus, the predecessors for /svinz/ ovens are ZT'USTZ', where the sequence ZT'ST is just what we expect before Z. Because of this, as also because the (n+2)th predecessor – i.e. the predecessor variety of each member of Z' or Z" – is the same as for the corresponding member of Z – we regard the Z' predecessor variety of /z/ and the Z" predecessor variety of /d/ as two alternants of Z. We can find this situation even in what seem to be the ordinary predecessors of /s, t/ before peaks or at utterance ends; and we can regard those predecessor varieties as alternants of Z' and Z" respectively. For we sometimes find that before /s, t/ also the sequence starts (backward) with T rather than T', as in /swaypt/ swiped Z 23, T' 1, S 3, T' 3, T 9, T 18. Our formula admits ZT'ST'T but not ZT'ST'TT; hence we interpret the last T 18 as a low Z". Later we will find that in certain positions the unique predecessor varieties of /z, s, ɔz/ (before peak or utterance end) complement each other, as do those of /d, t, ɔd/, and that these can be analyzed as alternants of a complete Z variety.

In What did he think of we find for /didhi/ Z 23, S I, T 3, S 9, S 19, T 4. The sequence is irregular. The peak of 19 would suggest a cut before /iy/, interpreting S 19 as a low Z. But this would leave ZSTS, which does not otherwise occur. The fact that the numbers decrease in grammatically restricted positions permits us to consider the S 9 as a very low grammatically-reduced Z (the members of this S 9 being all included in Z), yielding two regular sequences ZSTZST which correlate with the morphemic division did he.

In The silo walls were up we have for /saylow/ Z 23, S 10, T 2, R 27, S 15, T 3, with an unusually high R before /low/. In such cases we may suspect that the peak is due not to a sharp increase of variety in the middle of a regular sequence, but rather to adding two alternative (homonymous) morphemic analyses (low and words like furlough, hollow, silo), one of which
contributes a predecessor Z at this point and the other R or T' (whence the vowel members of this R 27 variety). The resultant is Z+R, and the segmentation is optional, depending on which analysis fits the rest of the utterance. If the sentence had not *silo* but *hollow* /halow/ – where the predecessors would be Z 23, S 2, R 27, S 15, T 3 – it would be clear that only the R contribution and not the Z is relevant for the R 27 in this utterance; because if we took the Z contribution and placed a cut there, the preceding Z 23, S 2 would have to constitute an inter-peak stretch in conflict with the formula above. In this way we are able in some cases (for *hollow* but not for *silo*) to reject a homonymous analysis of one section because it makes a neighboring section irregular.

In general, a small peak can be disregarded if it occurs inside an undividable sequence, e.g. in another /saylow/ where the predecessors are Z 22, S 1, T' 4, R 9, S 10, T 3. But where the sequence is divisible, an R which is not low may be suspected of including a (morpheme-bounding) Z as one of the alternative analyses: e.g. *admit* /ædmit/ Z 23, T' 1, R 7, S 3, T 10; *contains* /kɔntɛnz/ Z 21, S 1, T' 1, R 7, S 7, T' 3, T 7, Z' 16. In this way we recognize the possibility of a segmentation before *mit* and *tain*.

In some cases we find a possible but rare situation, which calls our attention to the need for a morphological decision (§ 0.2) at that point. E.g. in *tower* /tɔwər/ Z 23, Z 23, Z 23, Z 24, T 6, the string of Z raises our suspicions; it is due to a number of overlapping homonyms.

4.3. An interesting application of the sequence method arises in the case of French. If we look at the numbers of the successors in Table V, it seems that there are many peaks which do not correlate with the morphological boundaries of the sentences. However, we notice dependences among some of the successor varieties. For example, both /il/ and /ils/ have almost all the consonants and vowels as successors. But whereas each successor of /il/ has in turn a fair number of successors, we find after /ils/ that each consonant successor has about the same number of successors in turn as after /il/, while each vowel successor has very few successors in turn. Thus /p/ has many successors both after /il/ and after /ils/ (il peut, il pleut, il pense, il prend; il se plaint, ils se parlent, il se propose, il se peint). But whereas /e/ has many successors after /il/ (il épouse, il écrit, il éclate, il écarte), it has only one or two after /ils/ (il se hérissé). Let us call the variety after /il/ I and that after /ils/ E. Then we notice that the varieties neighboring E are restricted: chiefly, either E appears after a /ə/, in which case there will always be the particular successor variety F preceding it, or else E appears after some other vowel, in which case F may follow if a consonant plus vowel follow the /ə/ but not if two consonants or a vowel follow the /ə/. With this and some other regularities in mind we group the successor varieties as follows:
TABLE V

<table>
<thead>
<tr>
<th>1³h</th>
<th>I</th>
<th>F</th>
<th>E</th>
<th>H</th>
<th>E’</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>G</th>
<th>H</th>
<th>H</th>
<th>G</th>
<th>G</th>
<th>F</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>32</td>
<td>20</td>
<td>27</td>
<td>30</td>
<td>32</td>
<td>19</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>27</td>
</tr>
</tbody>
</table>

Il serait inconvenable d’y apporter un paraphyle

The E 27 after /ilso/ comes largely from se, a homonym of the se here. The G 17 after /... dia/ and the high G 22 after /para/ are suspect of being low I’ in view of the high (n+2)th average after them

| G | I | I²h | Hᵐ | E | F | Eᵐ | Hᵐ | I | H | E | H² | Eⁿ | H | F | G | H | G | H | G² | H | H | H | I |
| 11 | 32 | 34 | 16 | 31 | 19 | 32 | 11 | 33 | 24 | 31 | 23 | 32 | 18 | 16 | 16 | 15 | 9  | 2  | 1  | 1  | 1  | 31 |

Elle avait été sensible autrefois

The high H 23 after /sās/ suggests a low I’. Analyzing the 16 after /ible/ as F is supported by the high (n+2)th average, and by the free alternant /ible/ here, which ends with F 16, E 27. A peak after /otra/ would have been obtained if we had had l’autre here

| I | E | H | I | I | E | H | E’ | F | E | F | G | H | H | G | G | H² | G | I |
| 36 | 32 | 16 | 2  | 34 | 34 | 32 | 19 | 32 | 13 | 26 | 18 | 21 | 13 | 1  | 1  | 1  | 1  | 1  | 31 |

Le choix des signes est arbitraire

Comparison of many utterances shows additional automaticities of /ə/, among them that certain cases of I (after I or utterance beginning) occur before consonant plus vowel, whereas before a following consonant cluster they always have an additional /ə/ (with E as its successor variety). Such I E sequences then yield a single cut, not two

Notes:

For superscript letters see the notes to Table IV.

* Same as r; but the successor morphemes in turn have few predecessor morphemes, so that a cut would not be obtained in the reverse direction either.

I: All consonants but /ŋ/ and most vowels; with many successors after each of these, in the (n+2)th place, except for a few rare vowels and for fewer consonants (e.g. /z/).

I’’: Over half (usually) of the I (with fewer vowel successors after vowels), with (n+2)th average lower than for I and higher than for G, H.

I’: I plus /ɲ/.

E: I, except that the (n+1)th vowels have very few (n+2)th successors, and some of the (n+1)th vowels may be missing.

E’: E plus /ɲ/.
F: A few consonants and almost all the vowels, the vowels having many successors in \((n+2)\)th place and the consonants relatively few.

H: Varies from most vowels and consonants, to most vowels and a few consonants, but always with a middling number of \((n+2)\)th successors, clearly lower than in I or the vowel members of F; and down (as we proceed to the next cut) to very few vowels and consonants, with very few \((n+2)\)th successors each.

G: Consonants only, varying from most with a good number of \((n+2)\)th successors each, down to very few with very few \((n+2)\)th successors each; rarely there are one or a few vowels in addition to many consonants, the vowels having very few \((n+2)\)th successors.

The sequence of these classes of successor variety are usually:

\[
\begin{array}{ccc}
  & I & \\
  [G H] & (F) & E \\
  (E) & F & \\
\end{array}
\]

There may be from one to three G, followed by one to three H, or the reverse; this sequence may be repeated several times, or omitted entirely. This is followed by the sequence EF or FE, or by E or F alone, or by I.

We locate our segmentations at the end of this sequence, i.e. after I (or I'), EF, FE, and E (or E') or F alone. Since I' and E' include /n/, which characterizes G as against I and E, we conclude that I' = I + G, E' = E + G— in other words, that I' and E' indicate the possibility of alternative analyses, with and without a segmentation at that point. Whether or not we segment after I' and E' therefore depends on the surrounding sequences. The numbers for each class decrease as we proceed through a stretch between segmentations, and sometimes through a whole utterance or successive parts of an utterance. I' will be found to correlate with the end of a bound morpheme.

In the three French sentences, all the word boundaries, and some of the morpheme boundaries within a word, are given by these sequence-end segmentations. The other morpheme boundaries are missed because of morphophonemically or morphologically limited distribution, but only one of these would remain undiscovered after repeating the procedure backward. In addition, there are several segmentations due to homonyms, at points where there are no morpheme boundaries in these utterances; but of these too only one would remain after going backward.

5. Correlation of the Cuts with Morphemic Boundaries

It is of interest to note briefly under what conditions our procedure undercuts (i.e. fails to get a segmentation at a morphemic boundary) or over-cuts
(i.e. yields a segmentation at a point where there is no morpheme boundary).

We may fail to get a segmentation in the case of nonsegmental morphemes.\textsuperscript{27} For example, if a tone contour and a stress contour extend over the same phonemic stretch, our procedure will fail to distinguish them. We may also fail to find the second parts of discontinuous morphemes. In both cases, indirect evidence of the morphemic situation may nevertheless be found by this procedure. Finally, we may fail to get a segmentation for morpheme alternants which are restricted to a few following environments (e.g. \textit{left} and \textit{dramatic} in Table IV), or for morphemes which have a small morphological selection in the following position (§ 6.3).

On the other hand, we may get a peak where there is no morphemic boundary, due to nonmorphemic restrictions of phoneme distribution. For example, in a language (such as English) with a certain short-range periodicity of consonants and vowels\textsuperscript{28}, there are positions in which only consonants or only vowels or both may occur. Thus the number of phonemes which can occur as successors in a particular position is affected not only by the relation to morpheme boundary but also by this kind of syllabic periodicity. The variation in number of possible successors becomes greater if there are many more consonants than vowels or, as in Thai, many more tone-bearing vowels than consonants. One way of reducing the disturbing effect of this syllabic periodicity is to use a phonemic analysis with an approximately equal number of consonants and vowels.\textsuperscript{29} Another way is to use not the actual successor counts, but the ratio of the count after the \textit{n}th phoneme to the number of successors that ever occur after the \textit{n}th phoneme, or after the same short stretch of phonemes. This is the ratio of the actual successor count to the number of successors permitted by the syllabic structure in that position. This separates in part the phonemic restrictions due to syllabic structure from the phonemic restriction due to the particular utterance, and thus shows how much restriction is due to the occurrence of the \textit{n}th phoneme in this particular string, i.e. in this particular morphemic complex (cf. Table VII).

We may also get a peak where there is no morpheme boundary in cases of sectional homonyms. If our test utterance begins with /\textipa{θəsələw}/ \textit{The silo}, we will get a peak at /\textipa{θəsə}/ because of all the phonemes (in the associated utterances) that can follow \textit{The sigh}. This will not happen if the sectional homonym (\textit{sigh in silo}) is excluded by the intervening environment, e.g. if the test utterance is \textit{The brick silo}. Many sectional homonyms are avoided if we specify junctural allophones: for example, /\textipa{tæks}/ is a sectional homonym of /\textipa{tæksi\textsuperscript{iy}/, but /\textipa{tæks\textsuperscript{s}/ is not a sectional homonym of /\textipa{tæksi\textsuperscript{iy}/ (Table II). Going backward corrects some of these cases; e.g. we will get no peak before /\textipa{li}/, since only the first part of \textit{taxi} has a homonym here. But in \textit{silv}, where
both parts have homonyms, we may get a peak before /low/ (in *The silo walls were up*, not in *The silo-like building*). The variety-sequence method of § 4 also corrects some of these cases (e.g. *aluminum* § 4.1, *silo* § 4.2), the more so since it requires that the residue (after the sectional homonym, e.g. the /iy/ above) should have the same successor-characteristics as the other segments.

Furthermore, since there are segmentations at virtually all word boundaries and most morpheme boundaries, under-cutting or over-cutting usually occurs between correct cuts. The problems of morphological testing are thus limited to short stretches (note 3).

6. SUCCESSOR COUNTS AND LANGUAGE STRUCTURE

The method presented here involves a special case of a more general characteristic of language structure. Since the physical events which are observed in linguistics are in general occurrences of sounds relative to each other, it is reasonable to suppose that the structural features of a language can be expressed as particular types of relative occurrence of sounds. The question is whether we can find procedures for investigating these occurrences such as will yield, in some orderly way, the structural features (already or not yet known) which are of interest to us. One can look at such a set of procedures as successively investigating how the actual occurrence of sounds departs from a random occurrence, each investigation dealing with a departure from equiprobability that has not been treated by the preceding investigations, i.e. showing the extra contribution to nonrandomness at that level. The method of successor varieties belongs to such a set of investigations. These investigations will be briefly noted here with special reference to the morpheme-boundary correlation.

6.1. *Over Stretches Shorter than a Morpheme*

The frequency of each successor to each sound or phoneme can be studied in terms of the (transitional) probability of each phoneme in respect to its immediate neighbor alone. In general, frequency of occurrence correlates with what may be considered language use (or communication) as against language STRUCTURE; beyond this point the investigations will ask whether a sound (a subsequently defined element) EVER occurs in a given environment (or never, within a very large corpus) rather than how frequently it occurs there. Count and variety therefore do not include frequency.

Over a stretch only a few sounds in length, the successor and predecessor varieties of each sound determine the grouping of sounds into phonemes. Over slightly longer stretches, the successor and predecessor varieties of each
phoneme describe the phoneme distribution (phoneme classes, specially limited phonemes), and such phonologic periodicities as syllabic structure.

6.2. Existence of Morphemic Segments

When the successor count is applied not over arbitrary short stretches, but always to the whole stretch from the beginning of the utterance, we find a new periodicity over and above all the preceding ones. It is not only that we find a rise and fall of the counts: when we consider the sequences of \((n+2)\)th averages, and also the sequences of successor varieties, we find fairly regular periodicities. We find further that the beginnings and ends of utterances are almost always marked by certain of these sequences. We then segment at the points where sequences that characterize utterance boundaries appear within an utterance; these correlate in general with word boundaries. Other periodicities between word boundaries lead to subsidiary segmentations which correlate with morpheme boundaries.

When we know that a language must contain morphemes, these procedures yield segmentations which we can test to see if they satisfy morphemic relations. But if we had not known that such things as morphemes and words exist at all, then these procedures would reveal to us that every utterance is a succession of periodicities, and that these periodicities are occurrence relations of phonemes which depend on the whole string of phonemes (i.e. on everything that has been said) thus far in the utterance. The existence of something of the nature of words and morphemes could thus be discovered from this procedure.

6.3. Degrees of Independence

Morphological analysis in linguistics sets up as the morphemes of a language those phonemic stretches which are independent of (do not co-occur with) every other morpheme in at least one utterance. All morphemes have thus a certain minimum independence. Morphological methods do not – and in their present form cannot – distinguish among various degrees of independence; yet there are different degrees. In the successor count procedure different degrees of independence yield different results. Low degrees do not yield peaks, or yield them only in particular circumstances. High degrees yield peaks regularly. Methods like the successor count could be used for obtaining an indication of morphemic independence.

The major types of reduced independence are as follows. (1) Agreement, in which a morph is required in one position if a particular other morph occurs in another; e.g. \(-s\) after It contain. The successor count yields a peak if the position is not fixed in respect to what precedes; if the position is fixed there is no peak. The count responds therefore to the conditions rather than
the fact of dependence. (2) Morpheme alternants, as dramat for drama. The successor count yields no peak for alternants which occur in very restricted environments. (3) Bound morphemes (e.g. English *-ing*, or the *con-* and *-tain* of *contain*) have in general fewer neighbors than free morphemes. Hence the successor count at their end is in general lower than at the end of words. (4) The selection of a morpheme is the variety of other morphemes with which it occurs; thus, the verb selection of *people* is the list of verbs with which that noun occurs. Morphemes, and even words, end on reduced peaks if their selection in the next position is so small as to reduce the number of next phonemes.

### 6.4. *Over Stretches Longer than a Word*

In many cases, the successor counts, both peaks and troughs, become somewhat lower as we go from the beginning of an utterance to its end, or up to some medial point. This is in part due to the mounting restrictions of the grammatical structure. Some information about the grammatical structure, and the location of such divisions as phrase boundaries (or similar domains within which selection operates) can be found by means of secondary differences in the counts.

### 7. SUMMARY

This paper presents a method for counting, at each phonemic position $n$ of a test utterance, all the phonemes that occur in the $(n+1)$th place (in any utterances) after the particular string of phonemes from the beginning of the test utterance up to $n$. When this count is made for each $n$ of the utterance, it is found to rise and fall a number of times. If we segment the test utterance after each peak, we will find that the cuts accord very well with the word boundaries and quite well with the morpheme boundaries of that utterance. While the method works for a complete phonemic writing of the utterances, it is not disrupted but only somewhat reduced in effectiveness if we fail to specify junctures or contours. The disturbing effect of syllabic structure and the like can be reduced, e.g. by taking the ratio of the successor count at $n$ to the successors that ever occur after the $n$th phoneme even in other (syllabically similar) positions.

The method can also be applied backward through the utterance, or by insertions at each point within the utterance. This yields independent corrections on the forward results. In addition, we can consider the average of $(n+2)$th successors for each $(n+1)$th successor at $n$. And we can note not only the count but the actual variety (list) of successors at each $n$, and group these successor varieties into certain frequently occurring types. In these last two additions we obtain a new result: we find recurring sequences, in respect
to the class of the \( n \)th phoneme, of the \( (n+2) \)th average and of variety types; and we find that the utterance is largely a repetition of these regular sequences. If we segment the utterance at those points where the sequence looks as it does at utterance beginning or end, we get a segmentation which agrees very well with the word and morpheme boundaries for that utterance.

The regularity of these sequences means that this method could have led to the discovery of morpheme-like segments, even if we had not known otherwise that morphemes exist. The method as a whole can be viewed as part of an orderly set of kindred methods capable of yielding a large part of language structure in terms of the relative occurrences of sounds, these occurrences being the physical events of language. In particular, the present method can serve for the gap in procedures between phonology and morphology: using nothing more than phoneme distribution, it provides utterance segments which can be tested with fair success by morphological method, and can thereby be automatically corrected, where necessary, to yield the elements of morphology.

**TABLE VI**

<table>
<thead>
<tr>
<th></th>
<th>16</th>
<th>21</th>
<th>25</th>
<th>19</th>
<th>13</th>
<th>30</th>
<th>12</th>
<th>20</th>
<th>14</th>
<th>10</th>
<th>30</th>
<th>11</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d a s</td>
<td>k a n</td>
<td>j a</td>
<td>? i m e r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>30</td>
<td>14</td>
<td>9</td>
<td>30</td>
<td>21</td>
<td>15</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>25</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>n o x</td>
<td>m å l</td>
<td>š t u d i r e n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

_Das kann ja immer noch mal studieren_ 'That fellow might yet study'

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>13</th>
<th>24</th>
<th>24</th>
<th>24</th>
<th>24</th>
<th>24</th>
<th>19</th>
<th>2</th>
<th>24</th>
<th>15</th>
<th>24</th>
<th>5</th>
<th>24</th>
<th>18</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>h e m l o</td>
<td>a k x u e t h a s f a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>5</td>
<td>21</td>
<td>13</td>
<td>21</td>
<td>7</td>
<td>12</td>
<td>18</td>
<td>20</td>
<td>24</td>
<td>5</td>
<td>21</td>
<td>20</td>
<td>21</td>
<td>8</td>
<td>26</td>
<td>26</td>
</tr>
</tbody>
</table>

|    | 7 | 1 | 24 | 5 | 24 | 11 | 8 | 4 | 2 | 1 | 1 | 24 | 10 | 23 | 24 | 24 |
|----|---|---|----|----|----|----|---|---|---|---|---|----|----|----|----|---|---|
| r i m h a x a d a s i m s e l o |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 21 | 25 | 5 |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

_Hebrew:_ They did not take his new books; backwards: They did not take the books

The 20 predecessors to /ja/ have a very low \((n+2)\)th average, and so differ from the set of 20 or 21 predecessors which occur at the points where cuts are made.

<table>
<thead>
<tr>
<th></th>
<th>19</th>
<th>24</th>
<th>5</th>
<th>24</th>
<th>24</th>
<th>24</th>
<th>20</th>
<th>3</th>
<th>17</th>
<th>9</th>
<th>3</th>
<th>2</th>
<th>23</th>
<th>23</th>
<th>11</th>
<th>23</th>
<th>5</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>k o š e h i t k a r a v t i e l h a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>23</td>
<td>10</td>
<td>1</td>
<td>24</td>
<td>19</td>
<td>16</td>
<td>16</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>24</td>
<td>24</td>
</tr>
</tbody>
</table>

_Hebrew:_ When I approached the shoemaker's I called him
Dravidian languages furnish a severe test of the present method, because of their morphophonemic complexity and their many one-phoneme morphemes. In the sentences tested (the last obtained by Lisker, the others given to me by Krishnamurti), peaks are found at all word boundaries. Of the morpheme boundaries within words (as we know them from morphological analysis), about half are marked by peaks when the count is made in one direction, more when the count is made in both directions. We never get peaks at points where there are no morphological boundaries; but in some instances where a morpheme boundary may be variously placed (by alternative morphological analyses) or where it occurs in the middle of a long consonant, the peak appears not at the morpheme boundary but at the phoneme next to it.

The results of the count accord more closely with morphological analysis when we apply the correction of § 5 – i.e. when we take not the actual number of phonemes occurring after the nth phoneme in a particular sentence, but the ratio of that number to the number of phonemes that ever occur after this (nth) phoneme. Thus, about 40 phonemes occur after vowels in various contexts. (The sequence VV occurs only across word boundaries; but since we do not use morphological information in making these counts, all we know about each vowel occurrence is that it can be followed by roughly any consonant or any vowel.) Hence if a given vowel in the nth place of a test utterance has 13 successors, we say that occurrence in this nth place has restricted the number of possible successors from 40 to 13 – a restriction of about 68%. In contrast with vowels, most consonants have a total of only about 15 different successors in various contexts: every consonant is followed by the vowels, and most consonants have particular other consonants as occasional successors; CC is followed only by a vowel. If, therefore, a typical consonant in the nth place of a test utterance has 5 successors, we say that occurrence in this nth place has restricted the number of possible successors from 15 to 5 – about 67%, or as much as the restriction from 40 to 13 after a vowel.

We get still better accord with morphological boundaries when we find the recurring sequences of successor varieties, as in § 4. Some successor varieties consist of consonants only (after a vowel inside a word), some consist of most vowels plus a few consonants (after a consonant inside a word), and so on.

When I obtained these sentences and made the counts on them I had no knowledge of Telugu. I acquired the relevant morphological information only after obtaining the numbers and working out the peaks (including the C-V restrictions) by direct work with an informant.

→ 12 13 12a 40 10 17b 12 38 7 7 12c 40 11 40
ceppėmätatavinıipö
11 1 1 17 10 1 1 2 10 2 6 14 11a 13 ←

Having heard what is said, go

→ 9 14b 2 25 7 3 3 1 2d 22 11 8 1d 22 2 40
geđada:virigı:pađıdda:di

The wall, having broken, fell down

cëtulukadukkanınun
11 1 2 2d 1 1 7 1 2 3 4 12 1 5 1 7 4a 25 ←
Table VII continued

$I$ wash $my$ hands

$\rightarrow 7 \ 4 \ 17 \ 4 \ 3^e \ 1 \ 2 \ 2 \ 3^a \ 1^e \ 1^e \ 1 \ 34 \cdot 8 \ 4 \ 3$

prātyēkamaēyna:kro

$\rightarrow 1 \ 1 \ 34 \cdot 6 \ 8 \ 1 \ 3 \ 2 \ 1^d \ 6 \ 2 \ 3^a \ 11^3 \ 2^d \ 3 \ 34$

ttā:visayā-liēmi:liēvu

Notes:

a  This is a peak when the number is taken in proportion to the total number of possible successors after V, C, CC, etc. – that is, the restriction (mentioned in the headnote to this table) is at a minimum here. Going backwards, all counts of about 10 predecessors before a consonant are minima of restriction.

b  When the adjustment referred to in fn. a is made here, this is not a peak.

c  Since $/n/$ has many consonant successors, this is not a peak.

d  A peak is missing because of morphophonemic alternation. Double consonants were tested as both single and repeated phonemes, with minor differences as shown.

*  A peak is missing because of selectional limitations of distribution. The last sentence is presented here as given, even though the first word is a rare Sanskrit form, and krotta is a spelling-pronunciation of kotta.

NOTES

1  I have had the advantage of discussing the subject of this paper with Noam Chomsky. Bernard Bloch and Charles F. Hockett have devoted a great amount of their time to a careful reading of the paper, which now appears considerably modified as a result of their valuable criticism. For data and comments on particular languages I am indebted to Henry M. Hoenigswald (German), Carol Schatz (French), Fred Lukoff (Korean), and Leigh Lisker and Bh. Krishnamurti of Andhra University (Telugu). The English data were obtained with the aid of the Committee on the Advancement of Research of the University of Pennsylvania.

2  We are concerned here only with the segmentation at morpheme boundaries. The fact that some of the morphs are alternants of each other (allomorphs), and together comprise a single morpheme, is not relevant here. In /hīy iz leyt/ $He$ is late, there are three morphemic segments, even though the middle one is an alternant of a morpheme unit. We are here seeking a method that will locate cuts after the third and fifth and last phonemes (not counting junctures) in this sequence. Such a method will give us the morphemic segments of an utterance, whether or not some of these are alternants of other segments.

3  In some cases a segment, when morphologically tested, turns out not to constitute a morph. In almost all such cases the lack of correlation between this segmentation and the desired morphemic boundaries affects only a small portion of the utterance, and is automatically corrected by the ancillary procedures discussed in § 3 and § 4, or else by morphological analysis. For example, in /litdistorbārimi:/ $It$ disturbed me, the segmentation comes out /lit.dis.torb.dymi:/; we lack a cut at the morpheme boundary before /d/; but this affects only the segment /torbd/, which contains two morphs instead of one. When we test the morphological relations of this stretch, we find that it is not a morphemic segment, but that it can be divided into two morphemic segments. Analogously, in
/ðətɛksɪ/ The taxi, we get /ðə. tɛk.s.i/. with two cuts at points that are not morpheme boundaries. But when we test morphologically, we find that /s/ and /i/ and even their sum /si/ cannot be morphemic segments in this position, whereas the somewhat larger sum /tɛks/.i/ can be. Almost all cases where our segmentations do not coincide with morpheme boundaries fall within short stretches of this type; cf. § 5.

4 Note that we are asking not the frequency of the various phonemes, but only which ones ever occur in that position. In the example of § 1, the test utterance is /hɪz. kləʊr/ He's clever. After the first 5 phonemes of that utterance we find 11 different successors: that is, in all the sentences that begin with /hɪz/.k/ we can find 11 different phonemes after the /k/. Some of these are more frequent than others: the successor /s/ is frequent, as in /hɪz. kəʊr/ He's covered, /hɪz. kəmɪŋ/ He's coming; the successor /r/ is less so, as in /hɪz. kreɪzi/ He's crazy; and the successor /eɪ/ is rare, as in /hɪz.kjʊ. wɜːrd/ He's cured. We ask only how many different successors there are to the first 5 phonemes. We next consider the first 6 phonemes of the test sentence, and find that in all the utterances which begin with /hɪz.k/ there are only 7 different phonemes that ever occur after the /l/, again without regard to how frequent they are.

5 This is a special case, though the most common one. More generally: we segment the utterance at those points where the number and variety of successors (see below) is similar to that at utterance end. This formulation is needed, for example, in cases where strong syllabic and other phonemic restrictions are not corrected for (§ 5). It is also needed if, contrary to Table II, we wish to apply this procedure to a phonemic writing in which the juncture /+/ is kept as a separate segmental phoneme, e.g. /hɪz. kəʊr+ kwik.ər+/ for He's quicker.

6 The list of phonemes which occurs in any utterance after a particular utterance-initial sequence may be called the SUCCESSOR VARIETY for that sequence; while the number of phonemes in that list is the SUCCESSOR COUNT for that sequence.

7 Though we are here correlating the variety with the phoneme which it follows, we must remember that the variety and the count depend upon the whole utterance-initial sequence. In the example above, the 11-phoneme variety occurs after the phoneme /k/, but it is the successor variety of utterance-initial /hɪz.k/, not of /k/ in general. After /k/ in general we can find other phonemes in addition, for example /s/ after /k/ in pixie. The results of § 4 are obtained by correlating the successors of an utterance-initial sequence with the last phoneme of that sequence.

8 Here and in similar cases, it is understood that we refer to utterance-initial sequences.

9 The examples above suggest (without assuming morphological knowledge) that junctures and intonation or stress contours have a special relation to morpheme boundaries. Junctures and some contours correlate with morpheme boundaries; other contours correlate with phrase or sentence boundaries without regard to morphemes and words. In contrast, if we dropped some segmental phonemes (e.g. the vowels), we would not obtain a segmentation similar to that obtained with these phonemes included. This applies also to tones and stresses which are not part of long contours, such as the tones in 'tone languages'. Such tones have distributions like those of other phonemes of the language. We can therefore tell whether, in a given language, tones are parts of a contour (and can be omitted in these tests) by seeing what kind of phonemic distribution they have.

10 More exactly: if we want to segment an utterance which has a junctural pair, like the two sets in Table II, the junctural allophones must be specified. But if we are segmenting some other utterance, we can usually get a successor count for the segmental phonemes alone that is almost the same as (not better than) the count we get if we specify junctural allophones and contours. And this with less work and confusion on the informant's part.

11 In general, different phonemic representations will give somewhat different successor counts; necessarily, since the different analyses mean that the same allophonic facts are represented by different phoneme sequences. In most cases these differences will not suffice to yield different peaks, i.e. different locations for our tentative segmentations. But sometimes this will happen. In particular, phonemes with great restrictions of occurrence usually
yield very low successor counts (e.g. the successor of /o/ is usually only /h, w, y/; and this may raise a neighboring moderate count to a relative peak. A particular phonemic analysis may eliminate certain of these undesired low counts. But some difficulties are unavoidable, for frequently, when we are dealing with very restricted allophones, our phonemic representation will have to have one or another serious restriction, especially since solutions by means of long components cannot be used here (since they involve unpronounceable and non-successive elements).

15 When we counted the successors of \( n \) we approximated the morpheme dependence at position \((n + 1)\) upon the preceding phonemic sequence; but we made no use of any morphemic dependence of position \( n \) upon the following phonemic sequence. Sometimes (or always, depending on the structure of the language) the dependence upon the preceding sequence suffices to show whether there is a morpheme boundary before position \((n + 1)\). When it does not suffice, we may be able to find out whether there is a morpheme boundary before position \((n + 1)\) by counting the predecessors of position \( m \) from the end (where the utterance is of length \( n + m \)), thus finding the dependence of position \((m + 1)\) from the end (= position \( n \) from the end) upon the phoneme sequence which follows it.

16 Of course, all the inadequacies of the forward operations can also occur in the backward operation if the positions are reversed: if a morpheme in a particular position is in grammatical agreement with something later in the sentence; if a morpheme or alternant has limited distribution in respect to what precedes it; if the last few phonemes of a morpheme are identical with the total phonemes of some other morpheme. As an example of the limited morpheme: in *It disturbs me* we find, on going backward, only 2 predecessors before /turbzmly/; /s/ and /Rs/ (in *It perturbs me*); but on going forward we find a peak of 15 successors after /tids/.

17 E.g. *The silo walls were up* has a successor peak after /dssay/ (The sigh ...) and also a predecessor peak before /lownwolzwarwp/ (... low walls were up). In this case we would get a segmentation in the middle of silo.

18 For example, in *It disturbed me* we find 16 successors after the first /i/. Of these successors, which are in the \( (n + 1) \)th place, 6 had 29 successors after them in turn, in the \( (n + 2) \)th position: *it, if, itch, is, ill, in*; after these 6 successors a new morpheme could begin in the \( (n + 2) \)th place. Of the other 10 successors, 1 had 18 successors (\( /\gamma* \): eat, eager, easy, each, either, aeons, etc.), 1 had 10 successors (\( /\alpha \): imp, imbibe, immune, immediate, etc.), and 8 had from 1 to 4 successors (\( /\gamma* \): ink, English; \( /d* \): idiot, etc.).

To put it differently, the roughly decreasing numbers as we go from peak to peak when we interpret peaks as word or morpheme boundaries) mean that there are in English about 29 ways of choosing the initial phoneme of a word; then depending upon the choice of the initial there are about 6 to 18 ways of choosing the second phoneme; and depending on the choice of the second phoneme (and somewhat on the first too) there are about 2 to 15 ways of choosing the third phoneme; about 1 to 10 ways of choosing the fourth; and 1 to 3 ways of choosing each following phoneme up to the end of the morpheme.

19 These categories depend on the decreasing numbers between peaks. If we say that the \( (n + 2) \)th phonemes for a given \( (n + 1) \)th are in category B, we mean that there are about as many \( (n + 2) \)th phonemes here as we would expect to find if the \( (n + 1) \)th were the first phoneme of an utterance. On the basis of the successor varieties of \( 4 \), we can go back and modify these categories, so as to obtain categories which closely characterize the first few (and, backward, the last few) places of an utterance. Thus modified, the calculations of the present section yield segmentations that agree even more closely with morphological boundaries. The adjusted categories are: A for the class J and high-count M of \( 4 \); B for the class K and high-count L of \( 4 \); C for middle-count L and N, and low-count M; D for low-count L and N. Part of this adjustment can be obtained simply by doubling the value of every successor vowel, thus correcting for some of the difference between the possible number of vowel and consonant successors. The adjusted categories are used in Table I above; for purposes of the arithmetic averaging there we set \( B = 15 \), \( C = 5 \), \( D = 1 \). Then if after \( n \), 6 of the \( (n + 1) \)th phonemes have successors in category B (i.e. about 15 successors each) and 2 have successors in category C, and 1 has its successor
in category D, the total in \((n + 2)\)th place is \(6B + 2C + 1D\), and the \((n + 2)\)th average per \((n + 1)\)th phoneme is 11.2.

18 If the adjusted categories of fn. 17 are used, there is no drop at all between /hiyz/ and /hiyz/. The 29 successors to /hiy/ then have the following total of \((n + 2)\)th successors in turn: \(6A + 21B + 1C + 1D\) = an average of 14.2 per \((n + 1)\)th phoneme in the numerical values of fn. 17 and Table 1 above. The 29 successors to /hiyz/ have the following total for their successors: \(27B + 1C + 1D\) = an average of 14.3. The 22 successors to /hiyz/ have for their successors: \(22B\) = an average of 15. The average is virtually the same, A not being counted since its distribution differs from that of B, C, D. Almost all the \((n + 2)\)th places (including the C and D) have the distribution we would expect if the \((n + 2)\)th phoneme were the second of an utterance, that is, if there were a word boundary after the \(n\)th phoneme.

19 This does not mean that there is necessarily a morpheme boundary before \(xy\) in this sequence.

20 Thus the J 28 above is merely J at a later point of the sentence, or at a grammatically more limited point. However, after consonants of rare occurrence, a vowel may have low L successor even near the beginning of an utterance, as in they /dɪˈeɪ/, where the successors are K, 6, L 4, J 28.

21 K and L can occur in first position; K′ and L in second or third position; M, N, L can occur in the same medial positions.

22 The rise in L 8 can also be eliminated by correcting for the vowel-consonant distribution, as in § 5.

23 Before some consonants, T consists only of the vowels (i.e. they are always the first of a postvocalic cluster). In these cases we may write T before S, and T before Z; but one could also adopt some different convention. Similarly K′ after a few consonants contains only vowels, and is hence identical with N.

24 If a sequence ends in R instead of T, we understand (since R includes T) that it sometimes is a morpheme separate from the following stretch, and sometimes constitutes a single morpheme with the following stretch.

25 Recognizing this /t/ as a separate morpheme is less obvious than recognizing the /d/ (and so for /s/ as compared with /z/) because /t/ has the same predecessors when it ends a morpheme as when it is a suffix. However, the predecessor variety tells us that in a certain utterance position the sequences preceding /t/ are the same as those which can themselves end an utterance (and have the same \((n + 2)\)th predecessors), whereas in other positions we find fewer and more restricted sequences preceding the /t/. The phonetic possibilities may be the same in both positions, but the variety that we find is different. This is precisely the difference between the present method and a study of phonetic structure. In those cases where the predecessors to /t/ are the same as those we would find before a peak or utterance end, i.e. where the sequence ends in T rather than T′, we place a tentative cut.

26 I.e. 0 or 1 G followed by 0 or 1 H, repeated up to 10 or 15 times.

27 Note also that segmental morphemes which consist of one phoneme are not easily separated out, since their boundary may be overshadowed by the neighboring boundary. In any case, a plateau of two high numbers (as in 9, 14, 29, 29 for /hiyz/ He's) indicates two segmentations, even though there are not two separate peaks.

28 This can be established by purely distributional investigations of the successors and predecessors of phonemes, as in J. D. O'Connor and J. L. M. Trim, 'Vowel, Consonant, and Syllable – a Phonological Definition', Word 9 (1953), 103–22.


30 For the inclusion of such considerations in the bases of phonology, see Charles F. Hockett, A Manual of Phonology (Indiana University Publications in Anthropology and Linguistics), 1955.