MORPHEME BOUNDARIES WITHIN WORDS: REPORT ON A COMPUTER TEST

For the science of linguistics we seek objective and formally describable operations with which to analyze language. The phonemes of a language can be determined by means of an explicit behavioral test (the pair test, involving two speakers of the language) and distributional simplifications, i.e. the defining of symbols which express the way in which the outcomes of that test occur in respect to each other in sentences of the language. The syntax, and most of the morphology, of a language is discovered by seeing how the morphemes occur in respect to each other in sentences. As a bridge between these two sets of methods we need a test for determining what are the morphemes of a language, or at least a test that would tentatively segment a phonemic sequence (a sentence) into morphemes, leaving it for a distributional criterion to decide which of these tentative segments are to be accepted as morphemes.

The locating of morpheme boundaries within a word, by a recurrent process in the manner of the paper 'From Phoneme to Morpheme' (Language 31 (1955), 190–222; Paper II of this volume) has now been carried out on a computer. Briefly, the method is as follows: given the first $m$ phonemes of a given $n$-phoneme sentence, for every $m$, $1 \leq m \leq n$, we count how many different phonemes follow these first $m$ phonemes in all sentences which begin with these $m$ phonemes. The same procedure can be used to count the predecessors of the last $m$ phonemes of the sentence, for each $m$. The points in the given sentence at which the number of successors (or predecessors) forms a peak are, to a first approximation, the boundaries between the morphemic segments of the given sentence.

Since the corpus used in the computer test consisted of words, not sentences, the test could yield not word boundaries within a sentence but only morpheme boundaries within a word. The latter are much harder to determine than word boundaries within a sentence. In the present test, morpheme boundaries were sought from the beginning of each word up to the first hyphen or (if there was no hyphen) up to the end of the word. Since the dictionary was in conventional spelling, the test dealt not with sequences of
MORPHEME BOUNDARIES WITHIN WORDS

Testing from a virtually complete dictionary of the language gives results that have certain interest over and above an informant-test. However, an informant-test could also include words constructed by the informant on the basis of productive morpheme-combining rules; these would accentuate the morpheme boundaries.

Any tests using phonemes, informants, and whole sentences, should therefore give even stronger correlation than that obtained here between the next-neighbor count and morphologically justified morpheme boundaries. Furthermore, most of the words used in the present test were selected for being difficult words based on Latin morphemes only a fraction of whose combinations appear in English.

Forty-eight words were tested in the first computer test reported here. All are given below, with the count of predecessors, successors, and (for many words) the average of successors for each successor, as they were given by the computer. For each word, the program carried out the method of the second paragraph above (substituting 'word' for 'sentence'), checking all words in the dictionary for next-neighbor letters to the first (and last) m letters of the test word, for each m.

In many cases, the actual number of next-neighbors rose to a peak at the points of morphological boundary:

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>24</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>10.4</td>
<td>9.2</td>
<td>6.4</td>
<td>3.9</td>
<td>1.5</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The numbers above the word count the different predecessor letters (in all words of the dictionary) to the sequence of letters up to the end of the given word. The first line below counts the different successors to the initial sequence. And the second line below gives the average of different successors for each successor to the initial sequence. Thus there are 15 different second letters in the set of words beginning with d; and these 15 have an average of 10.4 different letters in third place, following them.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9</td>
<td>24</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>14</td>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>7.7</td>
<td>4.6</td>
<td>4.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In other cases, the peak consisted not of the number of next-neighbors, but of the ratio of the number of next-neighbors of the initial or final sequence in the given word to the number of next-neighbors to the preceding (or following) triple of letters whenever that triple might occur. This ratio measures the reduction in next-neighbors due to the whole initial or final sequence in the given word as against the purely phonological requirements of the local environment (vowels, consonant clusters, and word-juncture). Thus, in

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>26</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

fewer letters can follow arm in all words containing arm than can follow for in all words containing for; hence the 4 after m represents a higher ratio than the 4 after r. Similar cases are seen in: afterward, alligator, antithetical, deflationary, deformity, development, distastefully, perfection, etc.

In judging the next-neighbor count, it has to be understood that we cannot recognize a peak until the next-neighbor count has begun to drop. Hence, if the first letter of a word is a separate morpheme, we cannot recognize it in the forward count; nor can a last-letter morpheme be recognized in the backward count. More generally, single-letter morphemes cannot be recognized directly. In all these cases we have to use other indications: the count in the reverse direction, an overly high average of next-neighbors per next-neighbor, etc.

Every morpheme boundary appeared as a peak in at least one of the two directions. In these 48 words there were 83 interior points at which a morphemic segmentation would be made on purely morphological grounds. 49 of these points have a next-neighbor peak in both directions. The remaining 34 points appeared as a peak in only one direction. In about half of these cases, the other direction did not show a peak because the morphemes (almost entirely Latin) had very few neighbors in that direction (English having taken over only a few of the words composed of those morphemes in Latin); forward peaks are missed in alliterate (after liter), alabum (after alum), anomalous (after non), antepenultimate (after pen), devitalize (after vit and al), dormant, periodic (after od, even though this is obtainable as a word boundary), pestiferous (after fer); backward peaks are missing in inclined (before cli) before tell.

In the other half of the cases, the second direction did not show a peak because the morphemic segment is morphophonemically limited to the environment in this second direction: antipathetical (pathet) applicability (ap, plic, abl), application (ap, plic), apposition (ap), disembod (em), permissible (mis). Of the same nature are the combining forms in altitude, pestiferous (attributable to -tude, -fer), and in autograph, where the two
MORPHEME BOUNDARIES WITHIN WORDS

directions gave different preferences for the point of segmentation. An artificial morphophonemic form appears in permalloy (per- from permanent).

In addition, there were 17 peaks which did not occur at morpheme boundaries of the given word. These were due to morphemic homonyms of initial or final nonmorphemic segments of the given word. Thus the existence of a final suffix -er (frequent in the science dictionaries) produced a maximum of predecessors before -as in all words ending in -ous, etc. The suffixes -ally produced a morphologically wrong peak in dismally; -er produced it in answer, another; and the prefix dis- produced it in discus, dismally, disulfide; and de- produced it in deign, devilishly.

In no case was there a peak in the opposite direction. In many cases, the fact that the peak in question did not come at a morpheme boundary for the given word could be recognized by the fact that further next-neighbor counts (in the next one or two positions), and average of next-neighbors per next-neighbor, are low in comparison with phonologically comparable words which have morpheme-boundary peaks in both directions at the corresponding point. Compare, for example,

\[
\begin{array}{cccccccc}
4 & 7 & 1 & 1 & 5 & 24 & 12 & 26 \\
& & & & & & & \\
\text{disulfide} & \rightarrow & 15 & 24 & 24 & 5 & 2 & 4 & 2 & 1 \\
& & & & & & & 10.4 & 9.2 & 6.4 & 1.4 & 2.5 & 1.5 & 1 & 0
\end{array}
\]

with

\[
\begin{array}{cccccccc}
1 & 1 & 4 & 3 & 18 & 15 & 11 & 25 \\
& & & & & & & \\
\text{disembodied} & \rightarrow & 15 & 24 & 24 & 11 & 5 & 6 & 5 & 2 \\
& & & & & & & 10.4 & 9.2 & 6.4 & 2.5 & 2 & 2 & 1.2 & 0.5
\end{array}
\]

where the low count of next-neighbor and average of next-neighbors per next-neighbor after a in comparison with e suggests that a is not the first letter of a morpheme, whereas e is.

\[
\begin{array}{cccccccc}
2 & 1 & 3 & 5 & 4 & 4 & 20 & 23 & 6 & 21 \\
& & & & & & & & & & \\
\text{afterward} & \rightarrow & 26 & 14 & 5 & 1 & 20 & 5 & 3 & 1 & 1
\end{array}
\]

\[
\begin{array}{cccccccccc}
2 & 1 & 3 & 5 & 8 & 19 & 16 & 25 & 17 \\
& & & & & & & & & & \\
\text{aligator} & \rightarrow & 26 & 25 & 14 & 10 & 1 & 1 & 3 & 1 & 2 \\
& & & & & & & & & & 15 & 7.8 & 5.1 & 1.8 & 1 & 3 & 0.6 & 2 & 1
\end{array}
\]

72 STRUCTURAL AND TRANSFORMATIONAL LINGUISTICS
NOTES

1 The work was programmed by Dr. Philip Rabinowits, and the program was run by him on the CDC computer at the Weizmann Institute of Science in Rehovot. The English dictionary used was the one arranged on tape by Dr. A. F. Brown (now at Lehigh University) at the University of Pennsylvania, giving a forward and reverse alphabetization of the entries in Webster’s Unabridged Dictionary and in a number of major specialized science dictionaries.

2 Since the dictionary did not list plural, past, and -ing suffixes for each word, the computer results were adjusted to what they would have been if the given word with these suffixes were in the dictionary.

3 Certain other morphemic segmentations can not be directly recognized by this method, e.g. infixed and intercalated morphemes (or that alternant of the past-tense morpheme which appears in task). Secondary indications of this situation can be drawn from the number sequences, including the average of next-neighbors per next-neighbor.

4 Or somewhat more, depending on interpretation of the number sequences.

5 Although such a situation could occur.