4pSCb5. Deriving functional load of phonemes from a prosodically extended neighborhood analysis.

Mafuyu Kitahara*, Keiichi Tajima and Kiyoko Yoneyama

*Corresponding author’s address: School of Law, Waseda University, 1-6-1 Nishiwaseda, Tokyo, 169-8050, Tokyo, Japan, kitahara@waseda.jp

The functional load of phonemes is a long-standing, but not a main-stream notion in modern linguistics: that some pairs of phonemes distinguish more words than other pairs is intuitively plausible, but hard to quantify. Meanwhile, neighborhood effects in word recognition and production have been one of the central topics in psycholinguistics, leading to a wide variety of investigations. However, the Greenberg-Jenkins calculation, the most common definition of phonological neighborhood, deals only with deletion, addition, and substitution of phonemes, lacking any consideration of prosody. For example, homophones, which cannot be segmental neighbors and thus excluded in most neighborhood research, can be distinctive if lexical accent is specified. The role of onset/rhyme distinction in neighborhood calculation has been discussed, but morae, another basic unit of prosody, were not mentioned in the literature. We propose a novel method for calculating the functional load based on a prosodically extended neighborhood analysis. It is a frequency-weighted neighborhood density summed across neighbors for a particular phoneme. Accentual distinctions, morae or syllables, and context effects within a word are taken into account. The proposed method gives a better account for the difference in the acquisition order of segments across languages.
INTRODUCTION

The functional load of phonemes is a long-standing topic in linguistics[1]. The idea that some pairs of phonemes distinguish more words than other pairs is intuitively plausible, which leads to a hypothesis that the less the pair is used in distinction, the more the pair is susceptible to diminish in the process of language change. An entropy-based definition of the functional load was proposed [2], but a quantitative analysis of language change with an entropy-based method was not convincing[3].

Meanwhile, neighborhood effects have been one of the central topics in psycholinguistics, leading to a wide variety of investigations [4, 5, 6]. Neighborhood is commonly calculated on the basis of the Greenberg-Jenkins method: only one deletion, addition, or substitution of phonemes in the target word defines the set of neighbors[7]. The size of the set of neighboring words, called neighborhood density, is a property of the target word that influences speaker/hearer’s performance of speech recognition and production.

Neighborhood is closely related to the notion of minimal pair since the substitution of only one phoneme naturally leads to a set of minimal pairs. Deletion and addition can also form a minimal pair where the distinction is made between a phoneme and an empty position. Thus, neighborhood calculation of a reasonably large dictionary is a good starting point to redefine the notion of the functional load.

However, the Greenberg-Jenkins method lacks any consideration of prosody. Japanese, a pitch-accent language for example, has a triplet “ha*shi – hashi* – hashi (chopstick – bridge – edge)” only distinguished by accent. Accentually different but homophonal words are generally excluded in the literature of neighborhood research (but, see [8] for an analysis of the functional load of accent in Japanese). Accentual identity of neighborhood is another issue here. Segmental minimal pairs may or may not be identical in accents. If we interpret the identity strictly, the neighborhood of a word only consists of those with equal word length and the same accent. In the case of deletion/addition, however, word length may be different from the target word. For example, “x*yz” and “x*y” are neighbors by deletion/addition, with the accent on “x”. When we count the accent from the head of the word, they share the same accent. When we count the accent from the tail, it is not the case, though. Thus, the accent location calculation should be varied and compared across in order to assess the validity of incorporating prosody in the neighborhood calculation.

In the present paper, a novel method for calculating the functional load based on a prosodically extended neighborhood analysis is proposed. It is a frequency-weighted neighborhood density summed across neighbors for a particular phoneme. Accidental distinctions, morae or syllables, and context effects within a word are taken into account. The proposed method gives a better account for the difference in the acquisition order of segments across languages.

DATABASE

The database used in this study is called Psylex [9], which contains about 88000 words. All word entries in the database are taken from a popular dictionary [10] which includes accentual information. Various information about each word are added to dictionary entries in Psylex. In the current study, word frequencies (calculated from 10 years of newspaper issues), word familiarities (rated by about 30 subjects), and prosodic structure (moraic boundaries) are of interest among others.
NEIGHBORHOOD CALCULATION

The original Greenberg-Jenkins calculation, namely, deletion, addition and substitution of one phoneme were first carried out for all the entries. Table 1 shows a sample of the calculation. The target word for this sample set is “aAto” where “A” stands for the latter half of a long vowel. S-, D-, and A-NEIGHBOR in the first column stand for substitution, deletion, and addition, respectively. The third column gives a word length, and the fourth column gives the accent location from the head of the word, both counted by mora. The familiarity score in the next column is on a scale of 1 (least familiar) to 7 (most familiar).

<table>
<thead>
<tr>
<th>Entry</th>
<th>ID#</th>
<th>Length</th>
<th>Accent</th>
<th>Frequency</th>
<th>Familiarity</th>
<th>Seg</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>~TARGET</td>
<td>000110</td>
<td>3</td>
<td>1</td>
<td>1438</td>
<td>6.094</td>
<td>aAto</td>
<td>art</td>
</tr>
<tr>
<td>-S-NEIGHBOR</td>
<td>001920</td>
<td>3</td>
<td>1</td>
<td>785</td>
<td>6.062</td>
<td>auto</td>
<td>out</td>
</tr>
<tr>
<td>-D-NEIGHBOR</td>
<td>009890</td>
<td>2</td>
<td>1</td>
<td>188862</td>
<td>5.781</td>
<td>ato</td>
<td>later</td>
</tr>
<tr>
<td>-D-NEIGHBOR</td>
<td>009910</td>
<td>2</td>
<td>1</td>
<td>7694</td>
<td>5.781</td>
<td>ato</td>
<td>trace</td>
</tr>
<tr>
<td>-A-NEIGHBOR</td>
<td>426620</td>
<td>3</td>
<td>0</td>
<td>424</td>
<td>6.250</td>
<td>haAto</td>
<td>heart</td>
</tr>
<tr>
<td>-A-NEIGHBOR</td>
<td>426640</td>
<td>3</td>
<td>0</td>
<td>4065</td>
<td>5.906</td>
<td>paAto</td>
<td>part</td>
</tr>
</tbody>
</table>

ACCENTUAL IDENTITY

Based on the basic neighborhood calculation in Table 1, accentual identity was calculated in three different ways: mora-based counting from the head; mora-based counting from the tail; syllable based counting from the tail. The reason for trying these three is that, counting from the head and tail are both useful in the analysis of Japanese accentuation. Generalizations over “initial”, “penultimate” or “ante-penultimate” accent are all common in the literature. Moreover, it is widely known that the default accent location is “the syllable containing the third mora from the tail” [11]. For example, “atakkku” (attack) has 4 moras (boundary as “.”) and 3 syllables (grouped by []) with an accent on “ta”: [a].[ta*.k].[ku], while “appuru” (apple) also has 4 moras and 3 syllables with an accent on “a”: [a*.p].[pu].[ru]. In the latter case, the accent is relocated to the preceding mora within the syllable because the second mora in a syllable cannot usually bear an accent in standard Japanese. This generalization suggests that counting by syllable from the tail is worth exploring.

Descriptive statistics of the number of neighboring words for target word are shown in Table 2. It is expected to shrink by incorporating the accentual identity due to the fact that accentually non-identical pairs are simply removed from the set of neighboring words in calculation. Compared to the baseline condition which does not consider any accentual identity, the median, the mean, and the maximum value of neighborhood size decrease in all counting methods. Note that the neighborhood size is a property of each target word. We will see the effect of accentual identity for the number of oppositions (as a property of each phoneme) in the next section.

<table>
<thead>
<tr>
<th>Counting method</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>1.00</td>
<td>15.00</td>
<td>23.33</td>
<td>180.00</td>
</tr>
<tr>
<td>from head by mora</td>
<td>1.00</td>
<td>10.00</td>
<td>18.23</td>
<td>160.00</td>
</tr>
<tr>
<td>from tail by syllable</td>
<td>1.00</td>
<td>10.00</td>
<td>17.90</td>
<td>162.00</td>
</tr>
<tr>
<td>from tail by mora</td>
<td>1.00</td>
<td>10.00</td>
<td>17.60</td>
<td>158.00</td>
</tr>
</tbody>
</table>
FUNCTIONAL LOAD

The number of oppositions for a given pair of phonemes can be obtained from summing the numbers over all the neighborhoods. For example, “A-u, A-0, 0-h, 0-p” are the opposing phoneme pairs in the sample data in Table 1. “A-0” pair occurs twice between “aAto” and “ato” and other pairs occur only once in this particular set. Summing those occurrences across all the neighborhoods, “A-0” actually occurs 174 times in the baseline, i.e., those including accentually non-identical pairs. To check the effect of different counting methods for accent, the number of oppositions sorted in an increasing order is plotted in Figure 1.

![Figure 1: The number of oppositions for all the opposing pairs in the baseline condition and in 3 different accent counting conditions](image)

The overall trend in Figure 1 is that the accentual identity decreases the number of oppositions in the lower area, but it is reversed in the higher area. The solid line (baseline condition) and other dotted lines (accent counted) cross at about Index 220. This suggests that the decrease in neighborhood size shown in Table 2 affects in such a way that oppositions of the fewer numbers are trimmed more than that of the larger numbers. The difference among accent counting methods is not salient, however. Following the tradition of accent research in Japanese, counting from head by mora is tentatively adopted in the following analyses.

The number of oppositions is a property of a pair of phonemes. A particular phoneme is involved in many different pairs. Summing the number of oppositions in terms of a particular phoneme tells us how useful that particular phoneme is. Suppose /k/ in oppositions like /k-p/, /k-t/, /k-g/...and so on, the sum of the occurrences of those pairs gives us the first step to estimate the functional load of the phoneme. The results of such calculation are presented elsewhere [12].

At this point, we should ask whether oppositions are equally distinct at any position. For example, in a triplet “top–pop–pot”, the /t-p/ opposition at the first position divides the triplet into “top” – “pop, pot” while the same distinction at the final position distinguishes the latter two words. In other words, oppositions at a later position have less words to distinguish. More generally, the leftward context of an opposition narrows down the search space.
Table 3 shows a more detailed example of the leftward context obtained from Psylex database. “toriQpu – toriQku” is the pair where these are the only words left for the /k-p/ opposition to distinguish. When a listener is presented to one of this pair, the role of /k-p/ opposition is minimal because the search space is maximally narrowed down by the leftward context “toriQ”. On the contrary, the bottom example has many other pairs than “ope – oke”. Thus the role of /k-p/ is still important for the rest of the word to be heard.

### Table 3: Minimal pairs for /k-p/ opposition with respect to its leftward context (LC)

<table>
<thead>
<tr>
<th>Minimal pair</th>
<th>LC</th>
<th># of words sharing LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>toriQpu “trip” – toriQku “trick”</td>
<td>toriQ</td>
<td>2</td>
</tr>
<tr>
<td>toraQpu “trap” – toraQku “truck”</td>
<td>toraQ</td>
<td>3</td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
</tr>
<tr>
<td>supaato “spurt” – sukaato “skirt”</td>
<td>su</td>
<td>1611</td>
</tr>
<tr>
<td>ope “operation” – oke “orchestra”</td>
<td>o</td>
<td>3155</td>
</tr>
</tbody>
</table>

The above steps are formalized in (1) where the functional load is redefined as the property of segment within a particular lexicon.

\[
LDS(\alpha) = \frac{\sum_{i} \sum_{j} |/\alpha/:/\beta_{i}/|_{j}}{L} \tag{1}
\]

- **LDS(\alpha)**: Lexical distinctiveness of a target segment \( \alpha \)
- **\( \beta_{i} \)**: opposing segments
- **\(|/\alpha/:/\beta_{i}/|_{j}\)**: number of words sharing LC of a particular opposition
- **\( L \)**: Size of the lexicon

Incorporating all the entries in Psylex database \((L = 88574)\), a sample calculation for the functional load of /k/ is as follows:

\[
\sum_{j}^{n} \frac{|/k/:/p/|_{j}}{L} = \frac{2}{88574} + \frac{3}{88574} + \cdots + \frac{3155}{88574} = 6.0738
\]

\[
\sum_{j}^{n} \frac{|/k/:/t/|_{j}}{L} = 61.2414
\]

\[
\sum_{j}^{n} \frac{|/k/:/b/|_{j}}{L} = 41.6202
\]

\[
\text{LDS}(/k/) = 6.0738 + 61.2411 + 41.6202 + \cdots + \sum_{j}^{n} \frac{|/k/:/\beta_{m}/|_{j}}{L} = 1728.9011 \tag{2}
\]

### RESULTS AND DISCUSSION

Figure 2 shows the plot of **LDS** for all phonemes in Japanese against their occurrence frequency. Based on the discussion on accentual identity in previous sections, neighborhood was trimmed by the accent counted from the head by mora.

Figure 2 means that the order of segments in **LDS** and that in frequency are sometimes different. One insight drawn from this fact is an issue of (um)markedness ranking. Taking just one example, substitution errors in children’s speech show sometimes conflicting results [13]. In their study, error patterns suggest the partial hierarchy of “unmarkedness” as follows: /t/ > /k/ and /s/ > /ʃ/ in English, while /k/ > /t/ and /ʃ/ > /s/ in Japanese. However, the plain segmental frequency is in conflict with this hierarchy: /t/ > /s/ > /k/ > /ʃ/ in English, and /k/ > /s/ > /ʃ/ > /t/ in Japanese. The proposed **LDS** measure gives a better fit on the contrary: /t/ > /k/ > /s/ > /ʃ/ in
English, and /k/ > /t/ > /s/ in Japanese. Although the data for English is still very limited and tentative, we foresee more interesting results to come in the line of research on LDS and functional load.

ACKNOWLEDGMENTS

This work is supported by JSPS Grant No. 23520474.

REFERENCES


