A Child Database for Calculating Phonotactic Probability and Neighborhood Density

Jill R. Hoover1
Holly L. Storkel2
Douglas S. Kieweg3

1Child Language Doctoral Program, 2Speech-Language-Hearing: Sciences and Disorders, 3Digital and Electrical Engineering Core
University of Kansas

Form Characteristics

- Phonotactic Probability:
  - The likelihood of occurrence of sounds and sound sequences (common vs. rare)
  - Phonotactic probability affects memory (e.g., Munson, 2001), word learning (e.g., Storkel, 2001; Storkel, 2003) and grammatical morphemes (Lestard, Davis, & Devey, 2007) in children

- Neighborhood Density:
  - The number of phonologically similar sounding words (dense vs. sparse)
  - Neighborhood density affects word recognition (e.g., Garlock, Walley, & Metsala, 2001) production (e.g., Leonard, Davis, & Deevy, 2007) in children

Calculating Form Characteristics

- Adult-based databases have been used to calculate form characteristics of stimuli in child studies (e.g., Nusbaum, Pisoni, & Davis, 1984)
- It is unclear whether or not adult-based computations are similar to child-based computations

Purpose

1. Compare the form characteristics of words in a child database to the words in an adult database
2. Compare child-based form characteristic computations to adult-based form characteristic computations

Child Database Description

- The existing child databases of Kolson (1960) and Moe, Hopkins, & Rush (1982) were combined and edited to form one child database of 4,832 nonhomonymous root words, phonemic pronunciations, and spoken word frequency

Adult Database Description

- The Hoosier Mental Lexicon (Nusbaum et al., 1984) is an adult database of approximately 19,000 words with phonemic pronunciations and printed word frequency (Kucera & Francis, 1967)

Online Computerized Calculator Program

- Developed using the same algorithm used by the HML to calculate form characteristics

- Phonemic pronunciations of real words or nonwords are entered into the online calculator and phonotactic probability & neighborhood density are calculated

Phonotactic Probability

- Positional Segment Frequency: The likelihood of a given sound in a given word position
  - e.g., /k/ in /kit/
- Biphone Frequency: The likelihood of co-occurrence of two adjacent sounds
  - e.g., /ki/ in /kit/

Neighborhood Density

- Number of neighbors: the number of words that differ from a target word by one phoneme substitution, addition, or deletion
  - e.g., some of the neighbors of the word /fit/ include /fit/, /jfit/, /fj/...

Child-Based vs. Adult-Based Form Characteristic Computations

- Form characteristics were calculated for 3 sets of stimuli using both databases
- Child Database Words (n = 4, 832)
- CDI Nouns (n = 380)
- Word & Sound Learning Lab Nonword Stimuli (n = 310)

Summary & Conclusions

- Raw values of form characteristics obtained from the child vs. adult databases differ, but are highly correlated
- Phonotactic probability:
  - Child > Adult
- Neighborhood density:
  - Adult > Child

- High agreement for coding of dense vs. sparse and common vs. rare
- Child- and adult-based databases used to calculate form characteristics yield similar, but not exact, classification of stimuli
- Greater precision in stimulus classification may result from using a database that best matches participant characteristics

Correlations

<table>
<thead>
<tr>
<th>Variables</th>
<th>Adult &gt; Child</th>
<th>Adult &lt; Child</th>
<th>Child &gt; Adult</th>
<th>Child &lt; Adult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos. Seg. Frequency</td>
<td>r = .890 - .924</td>
<td>p &lt; .001</td>
<td>r = .871 - .922</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>Biphone Frequency</td>
<td>r = .784 - .867</td>
<td>p &lt; .001</td>
<td>r = .791 - .882</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>Neighborhood Density</td>
<td>r = .909 - .959</td>
<td>p &lt; .001</td>
<td>r = .890 - .924</td>
<td>p &lt; .001</td>
</tr>
</tbody>
</table>

Words were coded as high or low on phonotactic probability and neighborhood density for both sets of computations (i.e., child-based and adult-based) and the percent agreement for stimulus classification (i.e., common/rare and dense/sparse) was calculated

References


