A Child Database for Calculating Phonotactic Probability and Neighborhood Density

Jill R. Hoover[1] Holly L. Storkel[2] Douglas S. Kieweg[3]

[1]Child Language Doctoral Program, [2]Speech-Language-Hearing: Sciences and Disorders, [3]Digital 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 (Leonard, Davis, & Deevy, 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., German & Newman, 2004) word learning (e.g., Storkel, 2001; Storkel, 2003) and sound learning (e.g., Morrisette & Gierut, 2002) 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

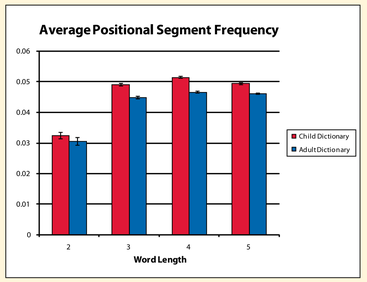
- Positional Segment Frequency: The likelihood of a given sound in a given word position
  - e.g., / k i w i /
- Biphone Frequency: The likelihood of co-occurrence of two adjacent sounds
  - e.g., / k i w i /

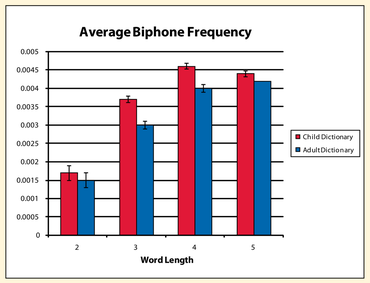
Neighborhood Density

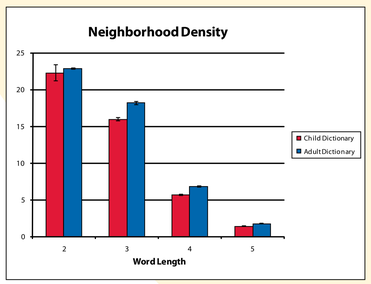
- Number of neighbors: the number of words that differ from a target word by a one phoneme substitution, addition, or deletion
  - e.g., some of the neighbors of the word / f i t / include / f l i t /, / tʃ i t /, / f i /

Child Database vs. Adult Database

- Phonotactic probability and neighborhood density were calculated for all of the words in the child database and then compared to the mean values of phonotactic probability and neighborhood density of all the words in the HML (available in Storkel, 2004)







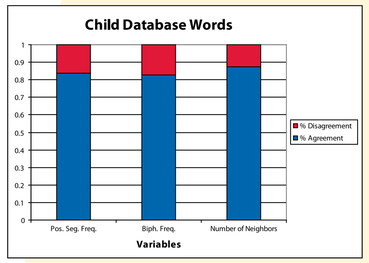
Child-Based vs. Adult-Based Form Characteristic Computations

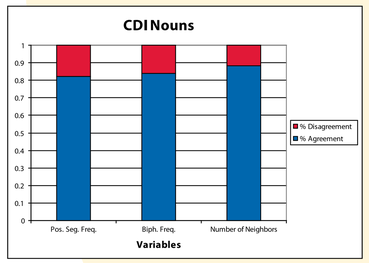
- Form characteristics were calculated for 3 sets of stimuli using both databases
- Child Database Words (n = 4, 832)
- CDI Nouns (n = 380)
- Word & Sound Learning Lab Nonword Stimuli (n = 310)

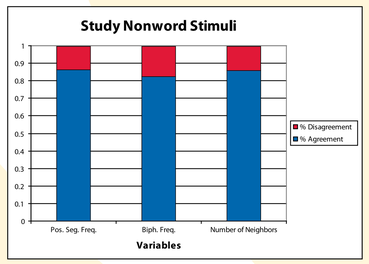
Correlations

	Positional Segment Frequency Adult	Biphone Frequency Adult	Neighborhood Density Adult
Positional Segment Frequency Child	r = .890 - .924 *p*s < .001		
Biphone Frequency Child		r = .784-.867 *p*s < .001	
Neighbohood Density Child			r = .909 - .959 *p*s < .001

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







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

References

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language, 44*, 1-25.

German, D. J., & Newman, R. S. (2004). The impact of lexical factors on children's word-finding errors. *Journal of Speech, Language, and Hearing Research, 47*, 624-636.

Kolson, C. J. (1960). The vocabulary of kindergarten children.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

Leonard, L. B., Davis, J., & Deevy, P. (2007). Phonotactic probability and past tense use by children with specific language impairment and their typically developing peers. *Clinical Linguistics & Phonetics, 21*, 747-758.

Moe, A. J., Hopkins, K. J., & Rush, R. T. (1982). *The vocabulary of first grade children*. Springfield, IL: Thomas.

Morrisette, M. L., & Gierut, J. A. (2002). Lexical organization and phonological change in treatment. *Journal of Speech, Language, and Hearing Research, 45*, 143-159.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research, 44*, 778-792.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon. In *Research on Spoken Language Processing Report No. 10* (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44*, 1321-1337.

Storkel, H. L. (2003). Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research, 46*, 1312-1323.

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research, 47*, 1454-1468.

This research is supported by DC009135, DC08095, DC005803 & the Word and Sound Learning Lab

For more information please contact:
jrhoover@ku.edu
www.ku.edu/~wrdlrng/