Methods for Minimizing the Confounding Effects of Word Length in the Analysis of Phonotactic

Probability and Neighborhood Density

Holly L. Storkel

University of Kansas

Abstract

Recent research suggests that phonotactic probability, the likelihood of occurrence of a sound sequence, and neighborhood density, the number of words phonologically similar to a given word, influence spoken language processing and acquisition across the lifespan in both normal and clinical populations. The majority of research in this area has tended to focus on controlled laboratory studies, rather than naturalistic data such as spontaneous speech samples or elicited probes. One difficulty in applying current measures of phonotactic probability and neighborhood density to more naturalistic samples is the significant correlation between these variables and word length. The current paper examines several alternative transformations of phonotactic probability and neighborhood density as a means of reducing or eliminating this correlation with word length. Computational analyses of the words in a large database and re-analysis of archival data supported the use of z-scores for the analysis of phonotactic probability as a continuous variable and the use of median transformation scores for the analysis of phonotactic probability as a dichotomous variable. Neighborhood density results were less clear with the conclusion that analysis of neighborhood density as a continuous variable warrants further investigation to differentiate the utitility of z-scores in comparison to median transformation scores. Furthermore, balanced dichotomous coding of neighborhood density was difficult to achieve, suggesting that analysis of neighborhood density as a dichotomous variable should be approached with caution. Recommendations for future application and analyses are discussed.

Key Words: children, adults, phonology, vocabulary selection

Methods for Minimizing the Confounding Effects of Word Length in the Analysis of Phonotactic

Probability and Neighborhood Density

The influence of language structure on the recognition, production, and acquisition of spoken

language has received attention recently as a means of understanding how adults and children use

regularities in the environment to process and acquire spoken language. One area of emerging interest

is regularities related to form, or sound, characteristics of language. Two form characteristics that have

been shown to affect language processing and acquisition across the lifespan are phonotactic

probability and neighborhood density. The majority of research on phonotactic probability and

neighborhood density has tended to focus on controlled laboratory studies (e.g., Aslin, Saffran, &

Newport, 1998, 1999; Beckman & Edwards, 1999; Edwards, Beckman, & Munson, in press;

Gathercole, Frankish, Pickering, & Peaker, 1999; Hollich, Jusczyk, & Luce, 2002; Jusczyk, Luce, &

Charles-Luce, 1994; Luce & Pisoni, 1998; Munson, Edwards, & Beckman, 2003; Saffran, Aslin, &

Newport, 1996; Storkel, 2001, 2002, 2003; Storkel & Rogers, 2000; Vitevitch, 2002a; Vitevitch,

Armbruster, & Chu, 2004; Vitevitch & Luce, 1998, 1999; Vitevitch, Luce, Charles-Luce, &

Kemmerer, 1997), rather than more naturalistically collected samples such as probe samples or

conversational speech (but see German & Newman, in press; Gierut & Morrisette, 1998; Gierut &

Storkel, 2002; Gordon, 2002; Morrisette, 1999; Newman & German, 2002; Storkel, 2004; Storkel &

Gierut, 2002; Vitevitch, 1997, 2002b). One stumbling block to analyzing the effect of phonotactic

probability and neighborhood density on naturalistic samples is the correlation between these variables

and other form characteristics, such as word length. The goals of this paper are to review previous

studies of phonotactic probability and neighborhood density, to consider the problems associated with

applying these constructs to the analysis of naturalistic data, and to examine potential solutions to this

problem.

Phonotactic Probability

Phonotactic probability refers to the likelihood of occurrence of a sound sequence such that certain sound sequences are considered common and others rare. A common sound sequence consists of individual sounds and adjacent sounds that co-occur in many other words of the language in the same word position. In contrast, a rare sound sequence consists of individual sounds and adjacent sounds that co-occur in few other words of the language.

Phonotactic probability has been shown to influence language processing across the lifespan. Eight-month-old infants appear to readily learn the phonotactic probabilities of an artificial language after minimal exposure (Aslin et al., 1998, 1999; Saffran et al., 1996). In addition, by 9-months of age, infants have extracted the phonotactic probabilities of their native language, demonstrating a listening preference for common sound sequences (Jusczyk et al., 1994). This knowledge of phonotactic probability appears to support segmentation of continuous speech into component words (e.g., Jusczyk, 1999; Mattys & Jusczyk, 2001; Mattys, Jusczyk, Luce, & Morgan, 1999; Thiessen & Saffran, 2003). Older children also appear to be sensitive to phonotactic probability and this seems to affect language processing and acquisition. In particular, phonotactic probability influences nonword repetition. Children repeat common sound sequences more accurately and with greater stability than rare (Beckman & Edwards, 1999; Edwards et al., in press; Munson, 2001). This has been documented in both typically developing children and children with phonological delays (Beckman & Edwards, 1999; Munson et al., 2003). In addition, children recall more nonwords composed of common sound sequences than those composed of rare sound sequences (Gathercole et al., 1999). In terms of acquisition, children are able to learn the phonotactic probabilities of an artificial language with minimal incidental exposure (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997). Moreover, phonotactic probability influences word learning with common sound sequences being learned more rapidly than rare for typically developing children (Hollich et al., 2002; Storkel, 2001, 2003; Storkel & Rogers, 2000). Interestingly, the reverse pattern, learning of rare sound sequences more rapidly than common, has been observed in children with functional phonological delays (Storkel, in press).

This influence of phonotactic probability continues into adulthood, with adults being faster to recognize nonwords composed of common sound sequences than those composed of rare sound sequences (Vitevitch & Luce, 1998, 1999; Vitevitch et al., 1997). In terms of production, adults are faster to name pictures of words composed of common sound sequences than pictures of words composed of rare sound sequences (Vitevitch et al., 2004). Furthermore, adults have better recognition memory for common over rare sound sequences (Frisch, Large, & Pisoni, 2000). Taken together, across the lifespan common sound sequences appear to entail a processing advantage over rare, except perhaps for children with phonological delays in certain processing tasks (e.g., word learning).

Neighborhood Density

Neighborhood density is often defined as the number of words that are similar to a given word based on a one phoneme substitution, deletion, or addition (Luce & Pisoni, 1998). Thus, some words reside in dense neighborhoods as defined by phonological similarity to many other words in the language, whereas other words reside in sparse neighborhoods because they are phonologically similar to only a few other words in the language.

Neighborhood density appears to affect language processing and acquisition in children. In particular, children require less acoustic information to identify words from sparse neighborhoods than words from dense neighborhoods (Garlock, Walley, & Metsala, 2001; Metsala, 1997). A similar sparse neighborhood advantage has been observed in children with hearing impairments who are more accurate recognizing words from sparse neighborhoods than words from dense neighborhoods (Kirk, Hay-McCutcheon, Sehgal, & Miyamoto, 2000; Kirk, Pisoni, & Osberger, 1995). In terms of speech production, typically developing children and children with word-finding difficulties appear to be more accurate naming words from sparse neighborhoods than words from dense neighborhoods (Newman & German, 2002, but see German & Newman, in press). Neighborhood density also has been shown to influence sound learning in children with phonological delays. Specifically, treatment of sounds in words from sparse neighborhoods promotes greater sound change than treatment in words from dense

neighborhoods (Gierut, Morrisette, & Champion, 1999; Morrisette & Gierut, 2002). In addition, neighborhood density differentiates which words are likely to change from incorrect to correct sound production, although the direction of this effect varies across studies depending on the type of sound change observed (i.e., phonetic versus phonemic), the word position studied (i.e., prevocalic versus postvocalic), and the sounds examined (e.g., fricatives versus affricates, Gierut & Morrisette, 1998; Gierut & Storkel, 2002; Morrisette, 1999; Morrisette & Gierut, 2001; Storkel & Gierut, 2002). Turning to lexical acquisition, typically developing children learn words from dense neighborhoods at earlier ages than words from sparse neighborhoods (Storkel, 2004, but see Hollich et al., 2002). Furthermore, it has been suggested that neighborhood density may influence the quality of mental representations of words, with words in dense neighborhoods having more detailed representations than words in sparse neighborhoods (Garlock et al., 2001; Metsala & Walley, 1998; Storkel, 2002).

Neighborhood density continues to influence language processing into adulthood. Young adults are faster to recognize words from sparse neighborhoods than words from dense neighborhoods (Cluff & Luce, 1990; Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998; Luce, Pisoni, & Goldinger, 1990; Vitevitch & Luce, 1998, 1999). Likewise, adults are less likely to misperceive words from sparse neighborhoods than words from dense neighborhoods (Vitevitch, 2002b). This advantage of sparse neighborhoods in spoken word recognition also has been observed in adults with hearing impairments (Dirks, Takayanagi, & Moshfegh, 2001; Kirk, Pisoni, & Miyamoto, 1997) and in older adults (Sommers & Danielson, 1999). Relative to speech production, the opposite pattern is observed. Here, words in dense neighborhoods are less likely to induce speech errors and word retrieval failures than words in sparse neighborhoods (Vitevitch, 1997, 2002a; Vitevitch & Sommers, 2003). Likewise, words in dense neighborhoods seem to be named more quickly than words in sparse neighborhoods (Vitevitch, 2002a). Adults with aphasia demonstrate a similar influence of neighborhood density on speech production. Words from dense neighborhoods are more resistant to phonological speech errors than words from sparse neighborhoods in adults with aphasia (Gordon, 2002). Likewise, words from

dense neighborhoods appear to be more resistant to phonological speech errors resulting in nonwords than words from sparse neighborhoods in adults with aphasia (Gordon, 2002).

Across the lifespan and normal and clinical populations, neighborhood density influences language processing and acquisition, although the direction of this effect varies by age and processing. For both normal and clinical populations of children, sparse neighborhoods seem to entail a processing advantage in recognition, production, and sound acquisition. In contrast, dense neighborhoods seem to facilitate word learning and the development of detailed representations of words. For both normal and clinical populations of adults, sparse neighborhoods entail a processing advantage in spoken word recognition, whereas dense neighborhoods promote a processing advantage in production.

One important extension in this area of inquiry is the application of these form characteristics to more naturalistically collected speech samples, such as spontaneous speech or elicited probes (cf., German & Newman, in press; Gierut & Morrisette, 1998; Gierut & Storkel, 2002; Gordon, 2002; Morrisette, 1999; Newman & German, 2002; Storkel, 2004; Storkel & Gierut, 2002; Vitevitch, 1997, 2002b). That is, re-analysis of data collected for other purposes may shed new light on the influence of form characteristics on language processing and acquisition. For example, given that phonotactic probability and neighborhood density have been shown to influence production, one might want to analyze the phonotactic probability and neighborhood density characteristics of speech errors obtained in a spontaneous speech sample or on an elicited naming task. In this case, the investigator may wish to use a continuous measure of phonotactic probability or neighborhood density to examine the degree of association between form characteristics and proportion of errors or types of errors. Alternatively, an investigator may want to compute proportion of errors for dichotomous categories of common/rare or dense/sparse words for analysis. One difficulty in applying these measures of form characteristics to more naturalistic samples is that correlations among variables must be considered so that appropriate conclusions can be drawn.

Correlations among form characteristics are well documented (Bard & Shillcock, 1993; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985; Vitevitch, Luce, Pisoni, & Auer, 1999; Zipf, 1935). In particular, word length is correlated with both phonotactic probability and neighborhood density because of the computational formulas typically used to measure these variables. This correlation is important in the analysis of naturalistic speech samples or in the re-analysis of data collected for other purposes because the words are likely to vary in length, leading to problems in analysis and interpretation. In past empirical studies of phonotactic probability and neighborhood density, word length was controlled explicitly to avoid these problems. Word length is positively correlated with phonotactic probability due to the computational formula used to quantify phonotactic probability in a number of studies. Specifically, the frequency of each sound or each sound combination in the word is summed. In this way, short words will tend to have lower phonotactic probability than long words because fewer values are being summed. This is counter to other studies of the lexicon which have shown that short words actually occur more frequently than long words (e.g., Zipf, 1935). In contrast, neighborhood density, measured by a one phoneme change metric, is negatively correlated with word length such that short words tend to have more neighbors than long words (Bard & Shillcock, 1993; Pisoni et al., 1985). Also of importance, when word length is controlled, phonotactic probability is positively correlated with neighborhood density with common sound sequences tending to reside in dense neighborhoods and rare sound sequences tending to reside in sparse neighborhoods (Vitevitch et al., 1999). Notably, this correlation is not a by-product of the computational formulas used to measure phonotactic probability and neighborhood density but rather is attributable to the intrinsic relationship between phonotactic probability and neighborhood density. For this reason, differentiating effects of phonotactic probability and effects of neighborhood density can be difficult.

The impact of these correlations among form characteristics on statistical analysis will depend, in part, on the analyses being used. One approach might be to analyze the items. In this case, one might compute the length, phonotactic probability, and neighborhood density of each word in the sample or

probe and also compute the proportion of errors for each word collapsed across participants. The variables related to form characteristics could then be entered in a regression analysis as potential predictors of error proportions. If all form variables are entered in the regression analysis, then the results for each variable will reflect the unique contribution of that variable to the prediction of error proportions when all other variables are controlled. In this way, the potential confounding effect of word length with phonotactic probability or neighborhood density may be resolved by the statistical analysis selected. However, this type of analysis may be problematic when the correlation among the predictor variables is high, resulting in inflated standard errors for the regression coefficients and inaccurate interpretation of these coefficients (e.g., Kleinbaum, Kupper, & Muller, 1988; Myers & Well, 1995). Thus, methods that reduce the correlation among the predictor variables are desirable.

A second analysis approach to this example might be to code each word into a dichotomous category of common versus rare or dense versus sparse. Error proportions could be computed for each participant for common versus rare or dense versus sparse words. Then, results could be analyzed using a paired t-test. In this type of analysis, it is difficult to disentangle the effects of phonotactic probability/neighborhood density and word length because the correlation between these variables is not incorporated into the analysis. It is likely that word length would be confounded with these category assignments such that, for example, the words coded as dense may be shorter in length than the words coded as sparse. For this reason, it would be useful to be able to sort words into dichotomous categories of common versus rare or dense versus sparse using a measure that avoided this confound.

Finally, in certain types of analysis it may be of interest to determine whether a particular speech sample or elicited probe is equally representative of common versus rare or dense versus sparse words. For example, in interpreting the previously described error analysis, it may be useful to know whether the participants produced an equivalent number of common versus rare or dense versus sparse words. If the majority of the words produced were common words with only a minority of rare words, then the error data for rare words may not be representative of the number and types of errors that are

made in response to rare words. The same argument can be made for imbalances related to neighborhood density. In addition, production of primarily one type of word (e.g., common words) in a token-based analysis may be of theoretical importance. To draw this type of conclusion, another desirable quality of a dichotomous measure of phonotactic probability or neighborhood density is that it divides a type-based sample of the lexicon into an equivalent number of common versus rare or dense versus sparse words so that asymmetries in token-based analyses will be interpretable.

The purpose of this database analysis was to explore possible methods to decrease the correlation between word length and phonotactic probability and word length and neighborhood density. A database of words representative of English was identified to explore the relationship between word length and phonotactic probability/neighborhood density in American English. Multiple measures of phonotactic probability and neighborhood density were computed in an attempt to minimize the correlation with word length. Subsequent descriptive and statistical analyses were performed to evaluate (1) the association between each measure of phonotactic probability or neighborhood density and word length; (2) the relative balance of common versus rare or dense versus sparse words; (3) the association between phonotactic probability and neighborhood density; (4) prediction of reaction time in a single nonword shadowing task.

<div align="center">Method</div>

Database

The *Hoosier Mental Lexicon* (HML) was used to examine the relationship between word length and phonotactic probability/neighborhood density (Nusbaum, Pisoni, & Davis, 1984). This database was chosen because it contains a large number of words (i.e., over 19,000 words), providing sufficient power to detect even small correlations; provides a transcription of each word based on American English; and supplies the word frequency of each word (Kucera & Francis, 1967). The HML has been used in numerous other studies of phonotactic probability and neighborhood density (e.g., Garlock et al., 2001; Gierut et al., 1999; Hollich et al., 2002; Jusczyk et al., 1994; Luce & Pisoni, 1998; Metsala,

1997; Morrisette & Gierut, 2002; Newman & German, 2002; Storkel, 2001, 2003, 2004; Storkel &

Rogers, 2000; Vitevitch, 1997, 2002a; Vitevitch et al., 2004; Vitevitch & Luce, 1998, 1999; Vitevitch

et al., 1997; Vitevitch & Sommers, 2003).

Form Variables

Three variables were computed for each word in the HML: (1) word length; (2) phonotactic

probability; (3) neighborhood density.

Word length. Word length was computed for each word by counting the number of sounds in

the phonetic transcription.

Phonotactic probability. Table 1 provides a sample phonotactic probability calculation for the

word "historiographer." Note that the values displayed in Table 1 do not exactly match those analyzed

due to rounding error in hand calculation of phonotactic probability for this illustration. As in past

research, two measures were computed as an index of phonotactic probability: (1) positional segment

frequency; (2) biphone frequency. *Positional segment frequency* is the likelihood of occurrence of a

given sound in a given word position. To compute the positional segment frequency for a given word,

the positional segment frequency for each sound in the word was computed and then summed. To

compute the positional segment frequency for a given sound in a given word, the log frequency of the

words in the HML containing the target sound in the target word position were summed and then

divided by the sum of the log frequency of the words in the HML containing any segment in the target

word position. *Biphone frequency* is the likelihood of co-occurrence of two adjacent sounds. To

compute the biphone frequency for a given word, the biphone frequency for each pair of sounds in the

word was computed and then summed. To compute the biphone frequency for a given pair of adjacent

sounds in a given word, the log frequency of the words in the HML containing the target pair of sounds

in the target word position were summed and then divided by the sum of the log frequency of the

words in the HML containing any pair of sounds in that word position. This denominator also includes

words containing a sound in the first position of a pair followed by a null sound (i.e., no following

sound). This algorithm has been used in numerous other studies of phonotactic probability (e.g., Jusczyk et al., 1994; Storkel, 2001, 2003; Storkel & Rogers, 2000; Vitevitch et al., 2004; Vitevitch & Luce, 1998, 1999; Vitevitch et al., 1997). Alternative algorithms are attested in the literature (e.g., Coleman & Pierrehumbert, 1997; Edwards et al., in press; Frisch et al., 2000). One important finding related to these alternatives is that wordlikeness judgments are more closely related to measures of overall phonotactic probability than measures of the individual probability of a part of a word (Coleman & Pierrehumbert, 1997; Edwards et al., in press; Frisch et al., 2000). Thus, measures of overall phonotactic probability, such as the current algorithm (as well as others), are preferred.

As previously noted, these measures of phonotactic probability that are based on sums are correlated with word length. In an attempt to decrease or eliminate this correlation with word length, three alternative measures were computed. First, the *average* segment frequency and biphone frequency were computed by dividing the previously described sums by the number of segments or biphones in the word. Second, a *z-score* based on the mean and standard deviation for words of a given length was computed, z-score = (obtained value – M)/SD. That is, the mean and standard deviation of the positional segment average and the biphone average were computed for all the words in the HML of a given length. These values are shown in Table 2. Then, for each word of a given length, these means were subtracted from the obtained positional segment or biphone average. The resulting value then was divided by the standard deviation, yielding a length sensitive z-score. Third, a *median transformation score* was computed for words of a given length, median transformation score = (obtained value – Mdn)/ ½ IQR. This score is similar to the z-score but uses the median and interquartile range (i.e., 25th – 75th percentile), rather than the mean and standard deviation. The median positional segment average and biphone average as well as the interquartile ranges were computed for all the words in the HML of a given length (see Table 2). Then, for each word of a given length, these medians were subtracted from the obtained positional segment or biphone average. The resulting value then was divided by half the interquartile range (i.e., a quartile).

One advantage of the z-score and median transformation score are that these scores have both a continuous and dichotomous interpretation. Relative to a continuous interpretation, incremental differences among the scores can be used to make fine-grain distinctions among words. In terms of a dichotomous interpretation, positive scores indicate that a sound sequence is common relative to other words of the language of that same length, whereas negative scores indicate that a sound sequence is rare relative to other words of the language of the same length. In terms of the difference between the z-score and median transformation score, z-scores have a known distribution, namely a z-distribution, with a mean of 0 and standard deviation of 1. It is possible, however, that z-scores may not equally divide the lexicon into common versus rare words because the mean may be influenced by extreme values and scores may not be normally distributed. Measures of phonotactic probability may be prone to this due to the use of word frequency in the computation. The median is less influenced by extreme values and may result in a more equivalent division of words into dichotomous categories.

Neighborhood density. Neighborhood density was computed for each word in the HML by counting the number of words that differed from a given word by a one phoneme substitution, deletion or addition. This algorithm has been used in numerous other studies of neighborhood density (e.g., Garlock et al., 2001; Gierut et al., 1999; Gordon, 2002; Hollich et al., 2002; Kirk et al., 2000; Kirk et al., 1997; Luce & Pisoni, 1998; Metsala, 1997; Morrisette & Gierut, 2002; Newman & German, 2002; Storkel, 2004; Vitevitch, 1997, 2002a; Vitevitch & Luce, 1998, 1999; Vitevitch & Sommers, 2003). Z-scores and median transformation scores were computed for neighborhood density using the same procedure previously described for phonotactic probability. Means, standard deviations, medians, and interquartile ranges for the words in the HML by word length are shown in Table 2.

Results

Each measure of phonotactic probability and neighborhood density (raw values, z-scores, and median transformation scores) was analyzed independently. Specifically, a correlation analyses was performed to determine whether word length was significantly associated with phonotactic probability

or neighborhood density. In cases where a significant correlation was not obtained, words were coded into dichotomous categories of common versus rare or dense versus sparse. A chi square test was then performed to determine whether this coding resulted in a balanced split of the lexicon into the categories of common/dense versus rare/sparse. Finally, the relationship between the measures of phonotactic probability and neighborhood density were examined via correlation analysis to examine degree of association and via McNemar's test to examine differences based on dichotomous coding.

Raw Values

Average positional segment frequency (top panel), average biphone frequency (middle panel), and neighborhood density (bottom panel) are shown for words of differing lengths in Figure 1. Word length was significantly correlated with average positional segment frequency and average biphone frequency, r (19290) = 0.28, $p < 0.001$, $r^2 = 0.08$, and r (19285) = 0.45, $p < 0.001$, $r^2 = 0.20$, respectively. Note that the degrees of freedom for the biphone frequency analysis are less than the degrees of freedom for the positional segment frequency analysis. This reduction is due to the exclusion of words containing only one segment because these words contain no biphones. As seen in Figure 1, as word length increased both measures of phonotactic probability also increased. Likewise, as reported in previous studies, word length was significantly correlated with neighborhood density, r (19290) = -0.60, $p < 0.001$, $r^2 = 0.36$. As word length increased, neighborhood density decreased.

The significant correlation between average phonotactic probability and word length was not anticipated. A priori, it was hypothesized that dividing by the number of segments would eliminate the effect of word length. This continued correlation was further explored by examining raw positional segment frequency and raw biphone frequency for each sound or sound pair by position. In this way, the individual values that were entered into the calculation of the average could be examined. Figure 2 displays the positional segment frequency (top panel) and biphone frequency (bottom panel) by word position. As can be seen from this figure, both segment frequency and biphone frequency tend to increase across word position. That is, sounds at the ends of words tended to obtain higher positional

segment frequencies and higher biphone frequencies than sounds at the beginnings of words. This, in turn, increases the average phonotactic probability for longer words relative to shorter words.

In terms of the correlation among form characteristics, positional segment average was significantly correlated with biphone average, r (19285) = 0.66, $p < 0.001$, $r^2 = 0.44$. Thus, as positional segment average increased, so too did biphone average. Likewise both positional segment average and biphone average were correlated with neighborhood density, r (19290) = 0.03, $p < 0.001$, $r^2 = 0.0009$ and r (19285) = -0.13, $p < 0.001$, $r^2 = 0.017$ respectively. Importantly, these correlations are in the opposite direction. Specifically, as positional segment average increased neighborhood density also increased. In contrast, as biphone average increased, neighborhood density decreased. Note that both effects are small ($r^2 < 0.02$).

Z-scores

Figure 3 shows the positional segment z-score (top panel), biphone z-score (middle panel), and neighborhood density z-score (bottom panel) for words varying in length. Word length was not significantly correlated with positional segment z-scores or biphone z-scores, r (19290) = -0.001, p =0.91, $r^2 = 0.000001$, and r (19285) = 0.000, p =0.98, $r^2 = 0.000$, respectively. Likewise, word length was not significantly correlated with neighborhood density, r (19290) = -0.001, p =0.91, $r^2 = 0.000001$.

Words were divided into dichotomous categories of common versus rare or dense versus sparse. Common/dense was defined as a positive z-score, and rare/sparse was defined as a negative z-score or a z-score of 0. Dichotomous coding based on positional segment z-scores resulted in a balanced split of the lexicon with approximately equal words classified as common (9,663 words) or rare (9,627 words), X^2 (1, 19290) = 0.07, $p = 0.80$. In contrast, dichotomous coding based on biphone z-scores resulted in significantly fewer words being classified as common (8,051 words) than rare (11,234 words), X^2 (1, 19285) = 525.36, $p < 0.001$. Likewise, classification based on density z-scores yielded fewer dense words (5,890 words) than sparse words (13,400 words), X^2 (1, 19290) = 2923.80, $p < 0.001$. As shown in Figure 3, there is minimal variability in z-scores for words with 7 or more

phonemes. When these words were excluded from the analysis, the imbalance remained with fewer

dense words (4,672 words) than sparse words (6,187 words), X^2 (1, 10859) = 211.37, $p < 0.001$.

In terms of the relationship among form characteristics, both measures of phonotactic

probability were significantly correlated with one another, r (19285) = 0.066, $p < 0.001$, $r^2 = 0.44$.

Although correlated, dichotomous coding based on each variable did not yield the same classification,

X^2 (1, 19285) = 531.059, $p < 0.001$, 14,404 concordant pairs (common/common; rare/rare) versus

4,881 discordant pairs (common/rare; rare/common). Both measures of phonotactic probability were

significantly correlated with neighborhood density, r (19290) = 0.26, $p < 0.001$, $r^2 = 0.07$ for positional

segment z-score; r (19285) = 0.34, $p < 0.001$, $r^2 = 0.12$ for biphone z-score. Dichotomous coding based

on phonotactic probability was significantly different from coding based on neighborhood density, X^2

(1, 19290) = 1763.727, $p < 0.001$, 11,223 concordant pairs (common/dense, rare/sparse) and 8,067

discordant pairs (common/sparse, rare/dense) for positional segment z-score; X^2 (1, 19285) = 684.602,

$p < 0.001$, 12,451 concordant pairs and 6,834 discordant pairs for biphone z-score.

Median transformation scores

Figure 4 shows the median transformation scores for positional segment average (top panel),

biphone average (middle panel), and neighborhood density (bottom panel) for words differing in

length. Word length was not significantly correlated with positional segment median transformation

scores or biphone median transformation scores, r (19290) = -0.012, $p = 0.08$, $r^2 = 0.00014$, and r

(19285) = -0.013, $p = 0.06$, $r^2 = 0.00017$, respectively. In contrast, word length was significantly

correlated with neighborhood density median transformation scores, r (19290) = -0.055, $p < 0.001$, $r^2 =$

0.003. To examine whether this correlation resulted from decreased variability in longer words, the

correlation analysis was performed using only words with fewer than 6 sounds. Words with fewer than

6 sounds were selected for analysis because the median neighborhood density is 0 for words of 6 or

more sounds, thereby restricting the median transformation scores to 0 or positive values. The

correlation remained significant in this subset analysis, r (7820) = 0.15, $p < 0.001$, $r^2 = 0.02$.

Words were divided into dichotomous categories of common versus rare using the same cut-off

scores as previously described for the z-score analysis. Coding based on positional segment median

transformation scores or biphone median transformation scores yielded approximately equivalent

numbers of common (9,655 based on segment; 9648 based on biphone) versus rare words (9,635

segment; 9637 biphone), X^2 (1, 19290) = 0.021, p = 0.89 for segment scores; X^2 (1, 19285) = 0.006, p

= 0.94 for biphone scores. Coding based on neighborhood density median transformation scores was

not analyzed due to the significant correlation with word length.

Turning to the association among the form variables, positional segment median transformation

scores were significantly correlated with biphone median transformation scores, r (19285) = 0.64, p <

0.001, r^2 = 0.41, and coding based on positional segment transformation scores did not differ

significantly from that based on biphone transformation scores, X^2 (1, 19285) = 0.002, p = 0.965,

14,639 concordant pairs versus 4,646 discordant pairs. Both measures of phonotactic probability were

significantly correlated with neighborhood density, r (19290) = 0.28, p < 0.001, r^2 = 0.08 for positional

segment transformation scores; r (19285) = 0.31, p < 0.001, r^2 = 0.10 for biphone transformation

scores. However, dichotomous coding of phonotactic probability did not yield the same classification

as neighborhood density, X^2 (1, 19290) = 1778.937, p < 0.001, 11,275 concordant pairs, 8,015

discordant pairs for positional segment transformation scores; X^2 (1, 19285) = 1938.446, p < 0.001,

11,949 concordant pairs, 7,336 discordant pairs for biphone transformation scores.

Prediction of Reaction Time

The previous analyses indicated that the correlation between word length and phonotactic

probability could be minimized using either z-scores or median transformation scores. In addition, an

equivalent division of the HML into common and rare words also was obtained using transformed

scores, but not raw values. Correlations between word length and neighborhood density also were

minimized using either z-scores or median transformation scores, but only the z-scores yielded a non-

significant correlation. Neither score resulted in an equivalent division of the HML into dense and

sparse words. These analyses suggest that z-scores and median transformation scores may be preferred over raw scores of phonotactic probability and neighborhood density. Before making a final judgment, it is important to determine which of these measures, if any, predict actual behavior. To address this issue, raw data were obtained from the single nonword shadowing experiments of Vitevitch and Luce (1998; 1999). In Vitevitch and Luce (1998), adult participants repeated single syllable common/dense and rare/sparse words and nonwords. In Experiment 4 of Vitevitch and Luce (1999), adult participants repeated two syllable common/dense and rare/sparse words and nonwords. Reaction times from the onset of the auditory stimulus to the onset of the participant's response were recorded. Only the nonword stimuli were analyzed here because interpretation of the data from the two syllable real words was complex, with reaction time being influenced by the probability/density of each syllable.

Mean reaction time, word length, raw phonotactic probability/neighborhood density, z-scores of phonotactic probability/neighborhood density, and median transformation scores of phonotactic probability/neighborhood density were computed for each nonword. For each form characteristic (positional segment frequency, biphone frequency, and neighborhood density), a separate forward linear regression analysis was completed. For each regression, the outcome variable was the mean reaction time and the predictor variables were word length and the three measures of each form characteristic (i.e., raw value, z-score, median transformation score). Across all three regression analyses, length was the first variable that entered the regression analysis, $B = 85.92$, $pr = 0.80$, $t(357) = 24.76$, $p < 0.001$ for positional segment frequency analysis, $B = 81.72$, $pr = 0.67$, $t(357) = 17.16$, $p < 0.001$ for biphone frequency analysis, $B = 94.87$, $pr = 0.81$, $t(357) = 25.63$, $p < 0.001$ for neighborhood density analysis. In the second step of the regression analysis, one of the measures of each form characteristic entered as a significant predictor of reaction time. The specific measure varied across form characteristics. Specifically, z-scores were significant predictors of reaction time for both measures of phonotactic probability $B = -12.74$, $pr = -0.16$, $t(357) = -2.97$, $p = 0.003$ for positional segment frequency, $B = -26.33$, $pr = -0.12$, $t(357) = -2.31$, $p = 0.02$ for biphone frequency.

In contrast, median transformation scores were significant predictors of reaction time for neighborhood density, $B = -13.79$, $pr = -0.15$, $t(357) = -2.77$, $p = 0.006$. It is important to note that neighborhood density was constant for the two syllable nonwords with a density of 0 for each nonword. Vitevitch and Luce (1999) computed neighborhood density for each syllable in the nonword; whereas, in the current study neighborhood density was computed for the nonword as a whole. Thus, extrapolating the current results to two syllable words or nonwords should be approached with caution. None of the remaining measures of each form characteristic significantly accounted for additional variance.

<div align="center">Discussion</div>

The goal of this study was to identify a measure of phonotactic probability and neighborhood density that was not significantly correlated with length and was a significant predictor of behavioral data so that this measure could be used in statistical analyses appropriate for continuous variables, such as correlations and linear regressions. An additional goal was to determine a dichotomous measure of phonotactic probability and neighborhood density that resulted in a balanced division of the HML into common versus rare or dense versus sparse words so that this measure could be used in statistical analyses appropriate for binary or categorical variables, such as t-tests and ANOVAs. Further, the inter-relationship between these measures was explored.

For phonotactic probability, results provided a relatively clear indication that z-scores are the preferred measure for analyses of continuous variables, and median transformation scores are the preferred measure for analyses of dichotomous variables. In the correlation and regression analyses, which rely on continuous variables, positional segment and biphone z-scores yielded a non-significant correlation with length and were the strongest predictors of reaction time. In the chi square analysis, which relies on dichotomous variables, positional segment and biphone median transformation scores yielded an equivalent split of common versus rare sound sequences in the HML. In addition, dichotomous coding based on positional segment median transformation scores was largely in agreement with coding based on biphone median transformation scores.

For neighborhood density, the results were less clear. In terms of analyzing neighborhood density as a continuous variable, z-scores showed a non-significant correlation with word length. In contrast, median transformation scores were the strongest predictor of reaction time. In future work, it may be important to consider both z-scores and median transformation scores of neighborhood density to shed further light on this issue. In terms of analyzing neighborhood density as a dichotomous variable, none of the measures resulted in a balanced division of the HML into dense versus sparse words. For this reason, analysis of neighborhood density as a dichotomous variable should be approached with caution because differences in the proportion of dense versus sparse words could arise due to problems in coding, rather than effects of neighborhood density on processing. One means of addressing this problem might be to randomly select control words either from the language sample being analyzed or from an existing database that are matched in word length to the words being analyzed, and then compare the proportion of dense and sparse words across the control and experimental words (see Gordon, 2002; Vitevitch, 1997, 2002b).

Importantly, these recommendations to use transformed scores of phonotactic probability and neighborhood density rather than raw values do not invalidate past studies of phonotactic probability and neighborhood density. The majority of these studies have controlled word length either by holding word length constant within a given study or matching items on word length. In this case, z-scores and median transformation scores represent a simple arithmetic transformation of the raw value, yielding the same rank ordering of stimuli. For example, in a study of consonant-vowel-consonant (CVC) stimuli, raw phonotactic probability values would be transformed to z-scores or median transformation scores using the mean, standard deviation, median and interquartile range of three phoneme words in the HML (refer to Table 2). Thus, all stimuli would be transformed using the same values. In this way, CVCs previously classified as common based on raw phonotactic probability would still obtain higher z-scores or median transformation scores of phonotactic probability than CVCs previously classified as rare based on raw scores. This also would be true of neighborhood density classifications.

Measures of phonotactic probability were always significantly correlated with neighborhood density. Of interest, the correlations obtained in this study where lower than those previously reported. Vitevitch and colleagues (1999) reported an r^2 of 0.37, whereas the current study obtained r^2s between 0.009 and 0.12. Vitevitch et al. focused only on CVCs whereas the current study analyzed words varying in length. This suggests that the robustness of the relationship between phonotactic probability and neighborhood density diminishes as word length increases. This change in association is most likely attributable to the reduced range of neighborhood density observed for longer words. That is, density is almost constant for words containing 7 or more phonemes. In contrast, phonotactic probability continues to vary across all word lengths. For this reason, examination of longer words may aid in disentangling effects of phonotactic probability from effects of neighborhood density. Although phonotactic probability and neighborhood density were correlated, dichotomous coding of the HML did not result in the exact same classification of words. That is, not all common words were classified as dense and not all rare words were classified as sparse. Although these variables are related, they are not identical constructs (see also Luce & Large, 2001).

One fortuitous finding from this study is that average phonotactic probability increased as word length increased. One possible interpretation of this finding is that the ends of longer words are more predictable than either the ends of shorter words or the beginnings of both long and short words. This seems possible given that long words tend to end in derivational or inflectional morphemes. For example, in the HML words containing 15 sounds ended in 'tion,' 'ity,' 'ible,' 'ed,' 'tal', and 'tive.' This points to the possibility that it may be useful to consider morphological relationships when computing phonotactic probability or neighborhood density. It is also the case that the denominator used in the phonotactic probability calculations decreases by word position. Thus, the pool of words serving as a comparison point reduces as words become longer. Both factors contribute to the greater predictability of the endings of long words.

The current study focused on eliminating effects of word length; however, in some cases it may be desirable to capture these trends across word length by using raw values for phonotactic probability or neighborhood density. For instance, it may be important to capture the greater predictability at the ends of long words because this may influence how the ends of long words are processed when compared to the ends of short words. Likewise, additional computations may be necessary to describe other aspects of phonotactic probability of long words. In particular, it may be useful to take into account the likelihood of a given word length, canonical structure, or stress pattern. Edwards and colleagues (in press) describe a method of computing phonotactic probability that is sensitive to canonical structure frequency. In addition, Frisch et al (2000) describe a method of computing phonotactic probability that is sensitive to stress patterns. Finally, Bailey and Hahn (2001) offer several additional variants of phonotactic probability and neighborhood density metrics that are sensitive to syllable structure. The Bailey and Hahn metrics only were examined in monosyllabic words in British English, but also might be useful in capturing patterns in multi-syllabic words in American English. Other means of accounting for the probability of word length or canonical structure await exploration.

In summary, use of transformed scores appeared to successfully reduce the correlation between phonotactic probability and word length and between neighborhood density and word length. More importantly, transformed scores were stronger predictors of reaction time than raw values. For this reason, transformed scores may be preferable to raw scores when analyzing the effects of form characteristics. In particular, z-scores are recommended for the analysis of phonotactic probability as a continuous variable; whereas median transformation scores are recommended for the analysis of phonotactic probability as a dichotomous variable. Analysis of neighborhood density as a continuous variable warrants further investigation to differentiate the utitility of z-scores in comparison to median transformation scores. When analyzing neighborhood density as a dichotomous variable, it is important to examine whether dense and sparse words are equally represented because balanced dichotomous coding may be difficult to achieve. Finally, although phonotactic probability and

neighborhood density were significantly correlated, the magnitude of this correlation was lower than previously reported. For this reason, investigations of form characteristics should attempt to measure both variables to discern the independent contribution of each to the behavior being studied.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321-324.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 359-380). Mahwah, N.J.: Lawrence Erlbaum Associates.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language, 44*, 568-591.

Bard, E., & Shillcock, R. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing* (pp. 235-275). Hove, Sussex: Erlbaum.

Beckman, M. E., & Edwards, J. (1999). Lexical frequency effects on young children's imitative productions. In J. B. Pierrehumbert (Ed.), *Papers in laboratory phonology V* (pp. 208-218). Cambridge, UK: University Press.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 551-563.

Coleman, J., & Pierrehumbert, J. (1997). *Stochastic phonological grammars adn acceptability.* Paper presented at the SIGPHON, Madrid.

Dirks, D. D., Takayanagi, S., & Moshfegh, A. (2001). Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners. *Journal of the American Academy of Audiology, 12*, 233-244.

Edwards, J., Beckman, M. E., & Munson, B. (in press). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language and Hearing Research.*

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*, 481-496.

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, 45*, 468-492.

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 84-95.

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

Gierut, J. A., & Morrisette, M. L. (1998). Lexical properties in implementation of sound change. In A. Greenhill & M. Hughes & H. Littlefield & H. Walsh (Eds.), *Proceedings of the 22nd Annual Boston University Conference on Language Development* (pp. 257-268). Somerville, MA: Cascadilla Press.

Gierut, J. A., Morrisette, M. L., & Champion, A. H. (1999). Lexical constraints in phonological acquisition. *Journal of Child Language, 26*, 261-294.

Gierut, J. A., & Storkel, H. L. (2002). Markedness and the grammar in lexical diffusion of fricatives. *Clinical Linguistics & Phonetics, 16*, 115-134.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501-518.

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and Language, 82*, 113-145.

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In B. Skarabela & S. Fish & A. H.-J. Do (Eds.), *Proceedings of the 26th annual*

Boston University Conference on Language Development (Vol. 1). Sommerville, MA: Cascadilla.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends Cognitive Science, 3*, 323-328.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630-645.

Kirk, K. I., Hay-McCutcheon, M., Sehgal, S. T., & Miyamoto, R. T. (2000). Speech perception in children with cochlear implants: Effects of lexical difficulty, talker variability, and word length. *Annals of Otology, Rhinology and Laryngology, 12*, 79-81.

Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research, 40*, 1395-1405.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Belmont, CA: Duxbury Press.

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

Luce, P., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16*, 565-581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1-36.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91-121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465-494.

Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory and Cognition, 25*, 47-56.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word Recognition in Beginning Literacy* (pp. 89-120). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Morrisette, M. L. (1999). Lexical characteristics of sound change. *Clinical Linguistics & Phonetics, 13*, 219-238.

Morrisette, M. L., & Gierut, J. A. (2001). Variables governing diffusion in phonological acquisition. In A. H.-J. Do & L. Dominguez & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development* (Vol. 2, pp. 534-541). Somerville, MA: Cascadilla Press.

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.

Munson, B., Edwards, J., & Beckman, M. E. (2003). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Manuscript submitted for publication.*

Myers, J. L., & Well, A. D. (1995). *Research design and statistical analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Newman, R. S., & German, D. J. (2002). Effects of lexical factors on lexical access among typical language-learning children and children with word-finding difficulties. *Language and Speech, 45*(3), 285-317.

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

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4*, 75-95.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological-Science, 8*, 101-105.

Sommers, M. S., & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging, 14*, 458-472.

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. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language, 29*, 251-274.

Storkel, H. L. (2003). Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language and Hearing Research, 46*(6), 1312-1323.

Storkel, H. L. (2004). Do children acquire dense neighbourhoods? An investigation of similarity neighbourhoods in lexical acquisition. *Journal of Applied Psycholinguistics, 25*, 201-221.

Storkel, H. L. (in press). The emerging lexicon of children with phonological delays: Phonotactic

constraints and probability in acquisition. *Journal of Speech, Language and Hearing Research*.

Storkel, H. L., & Gierut, J. A. (2002). Lexical influences on interword variation. In B. Skrabela & S.

Fish & A. H. J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on*

Language Development (Vol. 2, pp. 665-676). Sommerville, MA: Cascadilla Press.

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition.

Clinical Linguistics & Phonetics, 14, 407-425.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word

boundaries by 7- to 9-month-old Infants. *Developmental Psychology, 39*(4), 706-716.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech, 40*,

211-228.

Vitevitch, M. S. (2002a). The influence of phonological similarity neighborhoods on speech

production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 735-

747.

Vitevitch, M. S. (2002b). Naturalistic and experimental analyses of word frequency and neighborhood

density effects in slips of the ear. *Language and Speech, 45*, 407-434.

Vitevitch, M. S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech

production: Effects of phonotactic probability and onset density. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 30(2), 1-16.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of

spoken words. *Psychological Science, 9*, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in

spoken word recognition. *Journal of Memory of Language, 40*, 374-408.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40*, 47-62.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*, 306-311.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition, 31*(4), 491-504.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: Houghton Mifflin Co.

Author Note

Holly L. Storkel, Department of Speech-Language-Hearing, University of Kansas

Contact author: Holly Storkel, Ph.D., Assistant Professor, Department of Speech-Language-Hearing: Sciences and Disorders, University of Kansas, 3001 Dole Human Development Center, 1000 Sunnyside Avenue, Lawrence, KS 66045-7555. E-mail: hstorkel@ku.edu. Website: www.ku.edu/~wrdlrng/.

Table 1

Sample phonotactic probability calculation for "historiographer."

Position	Sound(s)	Numerator	Denominator	Description of Denominator	Frequency
			Positional Segment Frequency		
S1[a]	h	1132	28649	All words with ≥ 1 segment	0.0395
S2	ɪ	2740	28631	All words with ≥ 2 segments	0.0957
S3	s	2219	28119	All words with ≥ 3 segments	0.0789
S4	t	2221	24930	All words with ≥ 4 segments	0.0891
S5	o	400	20132	All words with ≥ 5 segments	0.0199
S6	r	705	15436	All words with ≥ 6 segments	0.0457
S7	i	570	11175	All words with ≥ 7 segments	0.0510
S8	ɑ	33	7477	All words with ≥ 8 segments	0.0045
S9	g	10	4549	All words with ≥ 9 segments	0.0022
S10	r	55	2502	All words with ≥ 10 segments	0.0219
S11	ə	46	1180	All words with ≥ 11 segments	0.0391
S12	f	2	510	All words with ≥ 12 segments	0.0039
S13	ɚ	1	182	All words with ≥ 13 segments	0.0055

Positional Segment Frequency Sum = 0.50

Average Positional Segment Frequency = Sum/Number of Segments = 0.50/13 = 0.04

Biphone Frequency

B1[b]	hɪ	115	28649	All words with ≥ 0 biphones	0.0040
B2	ɪs	477	28631	All words with ≥ 1 biphone	0.0167
B3	st	653	28119	All words with ≥ 2 biphones	0.0232
B4	to	47	24930	All words with ≥ 3 biphones	0.0019
B5	or	112	20132	All words with ≥ 4 biphones	0.0055
B6	ri	139	15436	All words with ≥ 5 biphones	0.0090
B7	iɑ	2	11175	All words with ≥ 6 biphones	0.0002
B8	ɑg	1	7477	All words with ≥ 7 biphones	0.0001
B9	gr	3	4549	All words with ≥ 8 biphones	0.0007
B10	rə	7	2502	All words with ≥ 9 biphones	0.0028
B11	əf	1	1180	All words with ≥ 10 biphones	0.0008
B12	fɚ	1	510	All words with ≥ 11 biphones	0.0020

Biphone Frequency Sum = 0.07

Average Biphone Frequency = Sum/Number of Biphones = 0.07/12 = 0.006

Note. Values in this table do not exactly match those used in the analysis due to rounding in this example during hand computation.

[a]"S" refers to segment position and "1" refers to the beginning of the word. [b]"B" refers to biphone position and "1" refers to the beginning of the word.

Table 2

Mean (M), standard deviation (SD), median (Mdn), and interquartile range (IQR) for each variable by word length.

Word Length	Positional Segment Average		Biphone Average		Neighborhood Density	
	M	*Mdn*	*M*	*Mdn*	*M*	*Mdn*
	(SD)	*(IQR)*	*(SD)*	*(IQR)*	*(SD)*	*(IQR)*
1	0.0113	0.0075			21.40	25.00
	(0.0118)	(0.0170)			(8.62)	(14.00)
2	0.0305	0.0282	0.0015	0.0010	22.89	23.00
	(0.0170)	(0.0248)	(0.0028)	(0.0017)	(8.05)	(13.25)
3	0.0449	0.0453	0.0030	0.0023	18.21	18.00
	(0.0166)	(0.0222)	(0.0026)	(0.0025)	(8.47)	(12.75)
4	0.0466	0.0463	0.0040	0.0033	6.87	6.00
	(0.0135)	(0.0173)	(0.0027)	(0.0034)	(4.53)	(7.00)
5	0.0462	0.0459	0.0042	0.0037	1.78	1.00
	(0.0121)	(0.0168)	(0.0025)	(0.0035)	(1.96)	(3.00)
6	0.0474	0.0473	0.0046	0.0043	0.53	0.00
	(0.0118)	(0.0162)	(0.0025)	(0.0035)	(0.86)	(1.00)
7	0.0483	0.0485	0.0049	0.0047	0.25	0
	(0.0111)	(0.0152)	(0.0023)	(0.0032)	(0.54)	(0)
8	0.0504	0.0507	0.0057	0.0053	0.16	0
	(0.0108)	(0.0152)	(0.0026)	(0.0035)	(0.43)	(0)
9	0.0521	0.0524	0.0069	0.0061	0.15	0

	(0.0104)	(0.0146)	(0.0036)	(0.0042)	(0.43)	(0)
10	0.0570	0.0571	0.0093	0.0071	0.12	0
	(0.0114)	(0.0161)	(0.0059)	(0.0070)	(0.34)	(0)
11	0.0564	0.0564	0.0082	0.0069	0.06	0
	(0.0111)	(0.0159)	(0.0043)	(0.0070)	(0.24)	(0)
12	0.0603	0.0608	0.0094	0.0070	0.03	0
	(0.0118)	(0.0171)	(0.0054)	(0.0105)	(0.17)	(0)
13	0.0563	0.0558	0.0077	0.0072	0	0
	(0.0108)	(0.0143)	(0.0036)	(0.0058)	(0)	(0)
14	0.0563	0.070	0.0080	0.0074	0	0
	(0.0113)	(0.0153)	(0.0034)	(0.0065)	(0)	(0)
15	0.0629	0.0593	0.0099	0.0096	0	0
	(0.0114)	(0.0120)	(0.0024)	(0.0044)	(0)	(0)

Note. Word length in number of sounds.

Figure Captions

Figure 1. Positional segment average (top panel), biphone average (middle panel), and neighborhood density (bottom panel) by word length (i.e., number of sounds). The line within the box represents the median or 50th percentile. The ends of the box represent the 25th and 75th percentiles (i.e., the interquartile range). Whiskers show the range of the data (i.e., 1.5 times the interquartile range), excluding extreme values and outliers.

Figure 2. Positional segment frequency for individual sounds by segment position (top panel) and biphone frequency for pairs of sounds by biphone position (bottom panel). The line within the box represents the median or 50th percentile. The ends of the box represent the 25th and 75th percentiles (i.e., the interquartile range). Whiskers show the range of the data (i.e., 1.5 times the interquartile range), excluding extreme values and outliers.

Figure 3. Z- scores by word length in number of sounds for positional segment average (top panel), biphone average (middle panel), and neighborhood density (bottom panel). The line within the box represents the median or 50th percentile. The ends of the box represent the 25th and 75th percentiles (i.e., the interquartile range). Whiskers show the range of the data (i.e., 1.5 times the interquartile range), excluding extreme values and outliers.

Figure 4. Median transformation scores by word length in number of sounds for positional segment average (top panel), biphone average (middle panel), and neighborhood density (bottom panel). The line within the box represents the median or 50th percentile. The ends of the box represent the 25th and 75th percentiles (i.e., the interquartile range). Whiskers show the range of the data (i.e., 1.5 times the interquartile range), excluding extreme values and outliers.