Behavior Research Methods, Instruments, & Computers
2004, 36 (3), 481-487

A Web-based interface to calculate phonotactic probability for words and nonwords in English

MICHAEL S. VITEVITCH
University of Kansas, Lawrence, Kansas

and

PAUL A. LUCE
State University of New York, Buffalo, New York

Phonotactic probability refers to the frequency with which phonological segments and sequences of phonological segments occur in words in a given language. We describe one method of estimating phonotactic probabilities based on words in American English. These estimates of phonotactic probability have been used in a number of previous studies and are now being made available to other researchers via a Web-based interface. Instructions for using the interface, as well as details regarding how the measures were derived, are provided in the present article. The Phonotactic Probability Calculator can be accessed at http://www.people.ku.edu/~mvitevit/PhonoProbHome.html.

Crystal (1992, p. 301) defined phonotactics as "The sequential arrangements of phonological units that are possible in a language. In English, for example, initial /spr-/ is a possible phonotactic sequence, whereas /spm-/ is not." Although phonotactics has traditionally been thought of in dichotomous terms (legal vs. illegal), the sounds in the legal category of a language do not all occur with equal probability. For example, the segments /s/ and /j/ are both legal as word-initial consonants in English, but /s/ occurs word initially more often than /j/. Similarly, the word-initial sequence of segments /s^/ is more common in English than the word-initial sequence /ji/. The term *phonotactic probability* has been used to refer to the frequency with which legal phonological segments and sequences of segments occur in a given language (Jusczyk, Luce, & Charles-Luce, 1994).

A comprehensive review of the studies that have demonstrated influences of phonotactic probability on the processing of spoken words is beyond the scope of this article, but it is worth noting a few examples to illustrate the breadth of processes that rely on this probabilistic information. For example, Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk (1993) found that sensitivity to phonotactic information occurs very early in life. They found that by 9 months of age, infants were able to discriminate among the sounds that were and were not part of their native language. Jusczyk et al. (1994) further demonstrated that infants of the same age could discriminate between nonwords that contained sounds that were more common or less common within their native language. Adults are also sensitive to the probability with which sounds occur in their native language. Ratings of the word-likeness of specially constructed nonwords by adults were influenced by phonotactic probability such that nonwords comprised of high-probability segments and sequences of segments were rated as being more like words in English than nonwords comprised of low-probability segments and sequences of segments (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Vitevitch, Pisoni, Kirk, Hay-McCutcheon, & Yount, 2002; see also Eukel, 1980; Messer, 1967; Pertz & Bever, 1975).

Phonotactic probability also appears to influence several on-line language processes. For example, phonotactic probability is one of several cues that enable infants (Mattys & Jusczyk, 2001) and adults (Gaygen, 1997; Pitt & McQueen, 1998) to segment words from fluent speech. Once a word has been segmented from fluent speech, phonotactic probability also influences how quickly children acquire novel words (Storkel, 2001, 2003; see also Storkel & Rogers, 2000), as well as how quickly normal hearing adults (Vitevitch & Luce, 1998, 1999; see also Vitevitch, 2003, and Luce & Large, 2001) and hearing-impaired adults who use cochlear implants (Vitevitch et al., 2002) recognize spoken words. Recent work also suggests that phonotactic probability influences the production, in addition to the comprehension of spoken language (Dell, Reed, Adams, & Meyer, 2000; Vitevitch, Armbrüster, & Chu, 2004). Clearly, phonotactic probability affects many spoken language processes.

The present article describes a computer program with a Web-based interface that provides estimates of phono-

This research was supported in part by Grants R03 DC 04259 and P30 HD 002528 (University of Kansas), and R01 DC 0265801 (SUNY Buffalo) from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. We thank the staff at Media Services in the Schiefelbusch Institute for Life Span Studies at the University of Kansas for their assistance in developing the Web-based interface. Correspondence concerning this article should be addressed to M. S. Vitevitch, Spoken Language Laboratory, Department of Psychology, 1415 Jayhawk Blvd., University of Kansas, Lawrence, KS 66045-7556 (e-mail: mvitevit@ku.edu).

tactic probability on the basis of a sample of words in an American English dictionary. Consistent with the goals of the current issue of *Behavior Research Methods, Instruments, & Computers* and the National Institutes of Health (NIH-NOT-OD-02-035, 2002), sharing data—like that contained in the database described in the present article—reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, and enables the exploration of topics that were not initially envisioned. Furthermore, by making this database and the methodology used to create it available to the scientific community, we hope to spare multiple researchers the difficulties inherent in developing such a database individually, and give researchers interested in studying other populations (e.g., children, languages other than English) a common starting point from which to make comparable databases.

Creation of the Database

Two measures are used in the database to estimate phonotactic probability: (1) positional segment frequency (i.e., how often a particular segment occurs in a certain position in a word) and (2) biphone frequency (i.e., segment-to-segment co-occurrence probability of sounds within a word). Both estimates were derived from the approximately 20,000 words in the Merriam-Webster Pocket Dictionary of 1964. Note that the entries in the database are from the Merriam-Webster Pocket Dictionary; however, the pronunciations were derived, checked, and edited by several researchers at the Massachusetts Institute of Technology, including Dennis Klatt, Dave Shipman, Meg Withgott, and Lori Lamel. The pronunciations in the database are in a computer-readable phonemic transcription developed by Dennis Klatt (see Luce & Pisoni, 1998; Nusbaum, Pisoni, & Davis, 1984). The computer-readable transcription is commonly referred to as "Klattese" and uses the following characters: @ a b C c D d E e f G g h I i J k L l M m N n O o p R r S s T t U u v W w X x Y y Z z ^ and |. A table containing the International Phonetic Alphabet (IPA) equivalents of the Klattese transcription is included in the Appendix. Note that stress and syllabification markers are not included in this database (nor in the Appendix) and therefore should not be included in the transcription entered into the Phonotactic Probability Calculator. Other conventions of the Klattese transcription are also described in the Appendix.

This database has been used in previous research to estimate neighborhood density for spoken words (e.g., Luce & Pisoni, 1998). *Neighborhood density* refers to the number of words that sound similar to a target word. Words with few neighbors, or few similar sounding words, are said to have a *sparse neighborhood* and are recognized more quickly and accurately than words with many neighbors, or words with a dense neighborhood. By using the same dictionary to estimate phonotactic probability and neighborhood density, one can rule out the possibility that the estimates of these two correlated variables (Vitevitch, Luce, Pisoni, & Auer, 1999) have been influenced by different sampling procedures used to create different dictionaries or lexicons.

Positional segment frequency was calculated by searching the computer readable transcriptions for all of the words in the dictionary (regardless of word length) that contained a given segment in a given position. The log (base 10) values of the frequencies with which those words occurred in English (based on the counts in Kučera & Francis, 1967) were summed together and then divided by the total log (base 10) frequency of all the words in the dictionary that have a segment in that position to provide an estimate of probability. Log-values of the Kučera and Francis word frequency counts were used because log values better reflect the distribution of frequency of occurrence and better correlate with performance than with raw frequency values (e.g., Balota, Pilotti, & Cortese, 2001; Zipf, 1935). Thus, the estimates of position-specific segment frequencies are token- rather than type-based estimates of probability.

By way of illustration, consider word-initial /s/. All the words in the dictionary that contained /s/ in the initial position were examined (e.g., *space*, *sat*, *siphon*, but not *infest*, *kiss*, etc.). The log-values of the Kučera and Francis (1967) frequency counts for those words with word-initial /s/ were summed and then divided by the summed log-values of the Kučera and Francis frequency counts for all the words in the dictionary that have a segment in that position to produce a token-based probability estimate that /s/ will occur in the initial position of a word in English. To further illustrate, consider /s/ in the second position of a word. Again, the log-values of the Kučera and Francis frequency counts for those words with /s/ in the second position would be summed and then divided by the summed log-values of the Kučera and Francis frequency counts for all the words in the dictionary that have a segment in the second position to produce a token-based probability estimate that /s/ will occur in the second position of a word in English. In this instance, all words that are one phoneme in length will not be included in the denominator because they do not contain a phoneme in the second position of the word. Similarly, when calculating the probability of a segment in the third position of a word, words that are only one or two phonemes long are not included in the denominator because they do not contain a phoneme in the third position of the word.

Position-specific biphone frequency was calculated in a similar way. That is, all instances in which a sequence of two phonemes occurred together in specific adjacent positions in a word were counted. The log-value of the frequency of occurrence for the words in which the (position-specific) two phoneme sequences were found were summed and then divided by the summed log-values of the Kučera and Francis (1967) frequency counts for all words in the dictionary that contained phonemes in those two adjacent positions. This yielded a token-based estimate of position-specific biphone probability.

Again, by way of illustration, consider the phoneme sequence /st/ found in the third and fourth positions of a word (e.g., *best*, *test*, *abstain*, *chest*, etc.). The (log-value) frequency of occurrence of all the words that contained /st/ in the third and fourth positions were summed and then divided by the summed log-values of the Kučera and Francis (1967) frequency counts for all words in the dictionary that contained biphones in the third and fourth positions. This produced a token-based probability estimate that /st/ would occur in the third and fourth positions of a word in English.[1] Note that the biphone probability that is calculated in this program is based on the co-occurrence of segments *within* words and not on the transitional probabilities of sounds *between* words as they might occur in fluent speech (cf. Gaygen, 1997).

How to Use the Web-Based Interface

A link to the Web-based interface of the Phonotactic Probability Calculator is located at http://www.people.ku.edu/~mvitevit. Users wishing to calculate the phonotactic probabilities of real English words or specially constructed nonwords should click on the link for "Phonotactic Probability Calculator." Note that the Phonotactic Probability Calculator is best viewed with the Internet Explorer Web browser. Analyses can be performed on words or nonwords up to 17 phonemes in length and must be phonemically transcribed into the computer-readable Klattese transcription, as described above. On the page that appears when following the link for "Phonotactic Probability Calculator" there is an electronic version of this article, an electronic version of the computer-readable transcription found in the Appendix of this article, the proper citation for using the Phonotactic Probability Calculator, and a request to send the first author a copy of any articles using or referencing the Phonotactic Probability Calculator.

After downloading and reading the instructions of how to use the Phonotactic Probability Calculator and the computer-readable transcription, click on the link to "Connect to the Phonotactic Probability Calculator." By doing so, a screen like that depicted in Figure 1 will appear in your Web browser. The left portion of the window contains a field to enter the list of phonemically transcribed words (or nonwords) whose phonotactic probabilities you wish to calculate. The user may either type the items directly in the field (one [non]word per line, using <return> or <enter> to move to the next line), or "copy and paste" the items that have already been typed (one word per line, separated by a hard return) from a simple word processing file. Clicking on the "Calc your Entry" button will result in the phonotactic probabilities for the items you entered appearing in the field on the right side of your browser. Note that the number of items you have entered, as well as the amount of traffic on the network, will determine how quickly your calculations are completed. To ease further analyses of these items, the user should copy and paste the results displayed in the output field to a simple word processing or spreadsheet file. Recall that the entries in the left field are treated as phonemic transcriptions. Thus, the word *cut* should be entered in the left

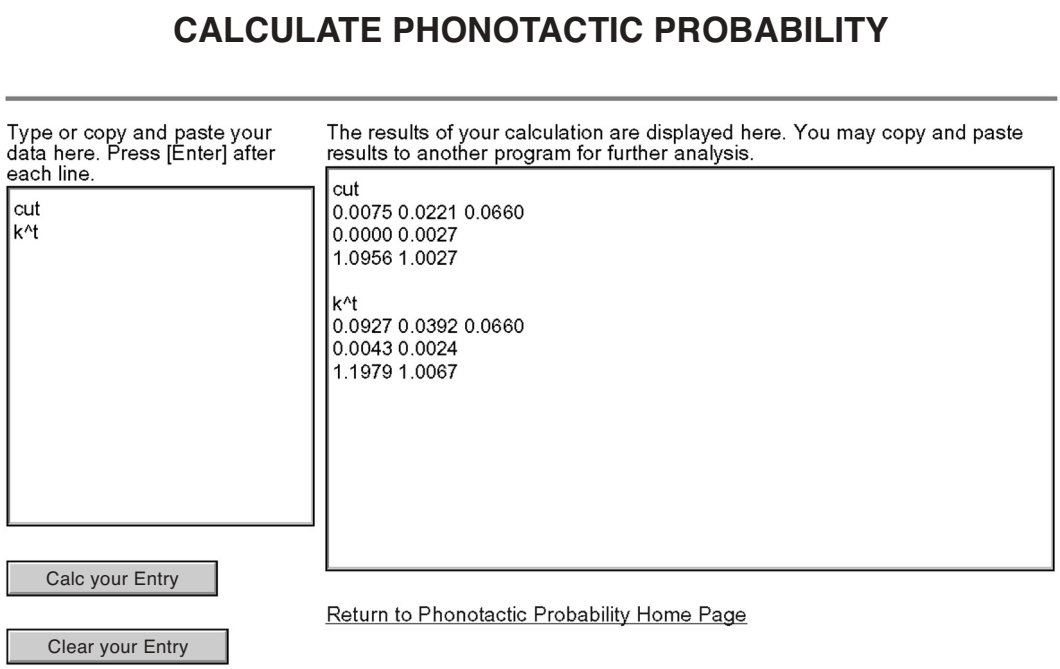

Figure 1. Depiction of the Phonotactic Probability Calculator Web page. The field on the left side of the page is where the computer-readable phonemic transcription is entered. The field on the right side of the page is where the output is displayed. Explanations of the output can be found in the text.

field in Klattese as k^t. If *cut* were entered in the left field, phonotactic calculations would be performed, but the output would be for /ɔut/ ("aw-oot"), not *cut* (see the computer-readable transcription in the Appendix). Note that some words in English do start with the phoneme /ɔ/, but no words in English start with the sequence /ɔu/; we will return to this example below.

The output of the Phonotactic Probability Calculator (which appears in the field on the right side of your browser) typically consists of four lines of information for each item entered in the left field. The first line contains the phonemic transcription you originally entered. The second line contains the position-specific probability for each segment in the item entered by the user. If the (non)word you entered contains more than seven phonemes, the position-specific probabilities for each phoneme in the item will wrap around to the next line.

The third line (assuming the entry is seven phonemes or less) contains the position-specific biphone probabilities for each of the biphones in the word. For example, a word with three phonemes, like *cut* (/k^t/), will have two biphones (/k^/ and /^t/). If the (non)word contains more than eight phonemes, the position-specific probabilities for the seven biphones in the item will wrap around to the next line. Entering segments or sequences of segments (e.g., /ɔu/) that are not attested in English (i.e., are illegal) in a given position will result in a probability of zero for that segment or sequence of segments (see Figure 1). Note that certain segments and sequences of segments may be legal in English but may not be legal in a given position. Consider the phoneme /ŋ/, as in *sing*. Although this segment does occur in English words, it does not appear in the word-initial position in English words. Thus, /ŋ/ would have a position-specific probability of zero if it were in the initial position of a stimulus item, but a nonzero position-specific probability if it were in some other position of a stimulus item.

The final line in the output for each entry contains the sum of all the phoneme probabilities and the sum of all the biphone probabilities. Note that the value 1 has been added to each of these sums to aid in locating these values when you cut and paste the output in the right field to another program (such as a spreadsheet or word processor). Users should remember to subtract this value (i.e., 1) from the sums of the phonemes and the sums of the biphones when reporting these values.

Previous studies (e.g., Vitevitch & Luce, 1998) using these estimates of position-specific phoneme and biphone probability ordered the sums of the phoneme probabilities and the sums of the biphone probabilities and found the median value for both measures in that set of potential stimuli. Items that were greater than the median values for both measures were operationally defined as (non)words with high phonotactic probability, and items that were less than the median values for each measure were operationally defined as (non)words with low phonotactic probability. Although the median value of a prospective stimulus set has been used in previous studies examining phonotactic probability to operationally categorize stimuli, other criteria may also be used. For example, the 33rd and 66th percentiles might be used to categorize stimuli as those with low, medium, or high phonotactic probability.

Considerations to Keep in Mind

There are a few issues that users of the Phonotactic Probability Calculator should keep in mind. For example, phonotactic probability in the present database is calculated on the frequency of occurrence information contained in Kučera and Francis (1967). Some have argued that the Kučera and Francis word counts are somewhat dated, or that these word counts should not be used in studies of spoken language because they were based on the occurrence of words in written texts, not from samples of spoken language. Although the entries in the English version of the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) are from a slightly more recent source, the frequency of occurrence counts contained in the CELEX database are still predominantly based on the frequency of occurrence of words from written sources. Furthermore, the orthographic and phonological entries in CELEX conform to standards of British English, whereas the entries in Kučera and Francis conform to the standards of American English. More important, several studies have found significant correlations between estimates of spoken and written language use (e.g., Pisoni & Garber, 1991), so it is unlikely that our estimates of phonotactic probability would change significantly if frequency of occurrence counts were obtained from a different corpus. Thus, a common word (or segment, or sequence of segments) will still be relatively common, and a rare word (or segment, or sequence of segments) will still be relatively rare, regardless of the source.

Although frequency of occurrence counts of written English words were used to provide estimates of phonotactic probability, the Phonotactic Probability Calculator does not provide estimates of orthotactics, or probability estimates of letters or sequences of letters occurring in a given position in a word. Although some languages (like Spanish) have a shallow orthography, or close to a one-to-one mapping of phonemes to letters, English has a much deeper orthography. This means that in languages with deeper orthographies, like English, a phoneme like /f/ might be realized orthographically as several different letters or letter combinations, as in *fig*, *cough*, *telephone*, *cuff*, and so forth. Alternative sources should be consulted for information regarding the frequency of occurrence of letters or letter sequences (e.g., Mayzner & Tresselt, 1965).

Also note that the Kučera and Francis (1967) corpus contains adult-directed language. Although some have argued that using adult-directed corpora in research with children is inappropriate, Jusczyk et al. (1994) demonstrated that the phonotactic probabilities of the specially constructed nonwords they employed in their study of 9-month-old infants remained relatively unchanged when the probabilities were based on an adult-directed corpus

(i.e., Kučera & Francis, 1967) or a child-directed corpus (i.e., the Bernstein [1982] corpus available in MacWhinney, 1991). As they stated (Jusczyk et al., 1994, p. 634): ". . . regardless of whether one calculates frequency estimates from adult- or child-directed speech corpora, the phoneme and biphone probabilities are substantially higher for items in the high-probability than for the low-probability phonotactic lists." Note that this relationship was found for the nonword stimuli employed in Jusczyk et al. (1994) and may not necessarily generalize to all stimulus sets, so some caution should still be exercised when using an adult-directed corpus to study language processing in children.

Another important consideration to keep in mind is that the method of calculating phonotactic probability that we employed was relatively neutral with regard to linguistic theory. Other linguistically relevant factors such as metrical stress pattern (Cutler & Butterfield, 1992), onset-rime syllable structure (Treiman, 1986), or even relationships between phonology and grammatical part of speech (Kelly, 1992) were not included as constraints in the calculations that estimated the probability of position-specific segments and sequences of segments. We view the neutrality toward linguistic theory as a feature that allows researchers to examine the influence of phonotactic probability on processing with minimal assumptions. Including constraints based on certain linguistic theories would require a new computational metric each time linguistic theory changed or developed. Furthermore, by starting out with a simple metric other assumptions and constraints could be added, depending on the needs and goals of individual researchers. For example, in studies seeking to identify the exact determinates of "wordlikeness" (Bailey & Hahn, 2001), additional assumptions and constraints may need to be included to account for increasing amounts of variability in the dependent measure of "wordlikeness."

Moreover, we believe that phonotactic probability is one of many factors that influences spoken language processing. Processing representations of many types of information—sublexical, lexical, semantic, contextual, visual, nonverbal, and so forth—influences spoken language comprehension (and production). By keeping the phonotactic metric as simple and as neutral to linguistic theory as possible, the interaction of phonotactic probability with other factors can be examined more easily (e.g., see Vitevitch et al., 1997, for a study investigating phonotactic probability and metrical stress).

The neutrality toward linguistic theory inherent in the calculations used to estimate phonotactic probability does not mean, however, that these metrics cannot be used to assess claims derived from linguistic theory. In a study examining onset-density, or the proportion of lexical neighbors that share the same initial phoneme as the target word, Vitevitch (2002a) ruled out the possibility that the rime portion of the monosyllabic CVC words used as stimuli was driving the difference in response times he observed in that study. Vitevitch (2002a) used a regression analysis with the measure of onset-density and the transitional probability of the VC sequence in all the stimulus words entering into the equation. The results of that analysis suggested that onset-density, rather than the frequency of the rime structure, was producing the observed differences in response times.

The use of (non)words with a number of other lexical characteristics explicitly controlled can also rule out potential confounds (see also Storkel, in press). For example, by using monosyllabic words, metrical stress is controlled for all the stimuli. Using monosyllabic words as stimuli also increases confidence that the researcher is using words that are relatively common in the language (see Zipf, 1935, for evidence that shorter words are more common than longer words in English). More important, users of the Phonotactic Probability Calculator should also be aware of the significant correlation ($r = +.61$) between phonotactic probability and neighborhood density that was found among CVC content words (Vitevitch et al., 1999; see also Landauer & Streeter, 1973). Words comprised of common segments and sequences of segments tend to have many lexical neighbors (i.e., dense neighborhoods). However, with careful stimulus selection, it is possible to control one variable while manipulating the other (e.g., Vitevitch, 2002b; Vitevitch et al., 2004), or to independently manipulate both variables (e.g., Luce & Large, 2001; Vitevitch et al., 2004).

Depending on the type of research being conducted, the issues discussed above may have implications for the conclusions one wishes to draw from a study that employed the metric of phonotactic probability described in the present article (Mook, 1983). It is our hope that the database that is described here and that is now available to the scientific community will stimulate new research questions and will facilitate the transition from basic research findings to practical or clinical applications.

REFERENCES

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX-2* [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

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

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, **29**, 639-647.

Bernstein, N. (1982). *Acoustic study of mothers' speech to language-learning children: An analysis of vowel articulatory characteristics*. Unpublished doctoral dissertation, Boston University.

Crystal, D. (1992). *An encyclopedic dictionary of language and languages*. Middlesex, U.K.: Blackwell.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory & Language*, **31**, 218-236.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of experience in language production. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1355-1367.

Eukel, B. (1980). Phonotactic basis for word frequency effects: Implications for lexical distance metrics [Abstract]. *Journal of the Acoustical Society of America*, **68**, S33.

Francis, W. N., & Kučera, H. (1984). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Gaygen, D. E. (1997). *The effects of probabilistic phonotactics on the segmentation of continuous speech*. Unpublished doctoral dissertation, SUNY, Buffalo.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory & Language*, **32**, 402-420.

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

Kelly, M. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignment. *Psychological Review*, **99**, 349-364.

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

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning & Verbal Behavior*, **12**, 119-131.

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

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

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

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

Mayzner, M. S., & Tresselt, M. E. (1965). Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, **1**, 13-32.

Messer, S. (1967). Implicit phonology in children. *Journal of Verbal Learning & Verbal Behavior*, **6**, 609-613.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, **38**, 379-387.

NIH-NOT-OD-02-035 (2002). NIH announces draft statement on sharing research data. Retrieved February 14, 2003 from http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-02-035.html.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words* (Research on Speech Perception, Progress Report No. 10). Bloomington: Indiana University, Psychology Department, Speech Research Laboratory.

Pertz, D. L., & Bever, T. G. (1975). Sensitivity to phonological universals in children and adolescents. *Language*, **39**, 347-370.

Pisoni, D. B., & Garber, E. E. (1991). Lexical memory in visual and auditory modalities: The case for a common lexicon. In H. Fujisaki (Ed.), *Proceedings of the 1990 International Conference of Spoken Language Processing*. Kobe, Japan.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory & Language*, **39**, 347-370.

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

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

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

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics*, **14**, 407-425.

Treiman, R. (1986). The division between onsets and rimes in English syllables. *Journal of Memory & Language*, **25**, 476-491.

Vitevitch, M. S. (2002a). Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 270-278.

Vitevitch, M. S. (2002b). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 735-747.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, **6**, 487-499.

Vitevitch, M. S., Armbrüster J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 514-529.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science*, **9**, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and spoken word recognition. *Journal of Memory & 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 & 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 & Language*, **68**, 306-311.

Vitevitch, M. S., Pisoni, D. B., Kirk, K. I., Hay-McCutcheon, M., & Yount, S. L. (2002). Effects of phonotactic probabilities on the processing of spoken words and nonwords by postlingually deafened adults with cochlear implants. *Volta Review*, **102**, 283-302.

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

NOTE

1. In the present database—as with the orthographic entries in the Kučera and Francis (1967) word counts—homographic homophones (e.g., *saw* and *saw*) had a single phonological entry; see Francis and Kučera (1984) for separate counts of homographic homophones based on the syntactic class of the words. The frequency of occurrence for the single phonological entry was the summed frequency of occurrence for the different forms of the homographic homophones. In the present database, the frequency of occurrence of heterographic homophones (e.g., *bare* and *bear*) were also summed under a single phonological word form; in Kučera and Francis (1967), heterographic homophones had unique orthographic entries.

APPENDIX

Table A1
International Phonetic Alphabet (IPA) and Computer-Readable "Klattese" Transcription Equivalents

IPA	Klattese
Stops	
p	p
t	t
k	k
b	b
d	d
ɡ	g
Affricates	
tʃ	C
dʒ	J
Sibilant Fricatives	
s	s
ʃ	S
z	z
ʒ	Z
Nonsibilant Fricatives	
f	f
θ	T
v	v
ð	D
h	h
Nasals	
n	n
m	m
ŋ	G
Syllabic Consonants	
n̩	N
m̩	M
l̩	L
Glides and Semivowels	
l	l
ɹ	r
w	w
j	y
Vowels	
i	i
ɪ	I
ɛ	E
e	e
æ	@
ɑ	a
ɑu	W
aɪ	Y
ʌ	^
ɔ	c
oɪ	O
o	o
ʊ	U
u	u
ɝ	R
ə	x
ɨ	|
ɚ	X

Klattese Transcription Conventions

Repeated phonemes. The only situation in which a phoneme is repeated is in a compound word. For example, the word *homemade* is transcribed in Klattese as /hommed/. All other words with two successive phonemes that are the same just have a single segment. For example, *shrilly* would be transcribed in Klattese as /SrIli/.

X/R alternation. /X/ appears only in unstressed syllables, and /R/ appears only in stressed syllables.

Schwas. There are four schwas: /x/, /|/, /X/, and unstressed /U/. The /U/ in an unstressed syllable is taken as a rounded schwa.

Syllabic consonants. The transcriptions are fairly liberal in the use of syllabic consonants. Words ending in *-ism* are transcribed /IzM/ even though a tiny schwa typically appears in the transition from the /z/ to the /M/. However, /N/ does not appear unless it immediately follows a coronal. In general, /xl/ becomes /L/ unless it occurs before a stressed vowel. Words that end in the suffix *-ly* are exceptions. For example, *bodily* is /badxli/ not /badLi/.

Vowels preceding /r/. Nine of the vowels appear before /r/. In some cases, the differences are subtle, as between /cr/ and /or/, or /@r/ and /Er/.

ar	as in *aardvark*	Ir	as in *fear*
cr	as in *horse*	Ur	as in *tour*
or	as in *hoarse*	Yr	as in *fire*
@r	as in *hairy* or *Mary*	Wr	as in *hour*
Er	as in *herring* or *merry*		

Diphthongs. /yu/ and /wa/ are considered by some to be diphthongs. In the database, /yu/ in stressed syllables is /yu/ and in unstressed syllables it is /yU/. /wa/ is transcribed as /wa/ in all stress environments.

Alternate pronunciations. Only the most common pronunciation for each word is included in the database (e.g., *tomato*, *potato*).

The information in this Appendix is also contained in a document that accompanied the computerized database, prepared by several researchers at MIT including Dennis Klatt, Dave Shipman, Meg Withgott, and Lori Lamel. The information is included here because the Phonotactic Probability Calculator uses information contained in this database, and therefore, the same conventions.

(Manuscript received April 30, 2002;
revision accepted for publication October 14, 2003.)