Behav Res
DOI 10.3758/s13428-013-0356-8

Phi-square Lexical Competition Database (Phi-Lex): An online tool for quantifying auditory and visual lexical competition

Julia F. Strand

Abstract A widely agreed-upon feature of spoken word recognition is that multiple lexical candidates in memory are simultaneously activated in parallel when a listener hears a word, and that those candidates compete for recognition (Luce, Goldinger, Auer, & Vitevitch, *Perception* 62:615–625, 2000; Luce & Pisoni, *Ear and Hearing* 19:1–36, 1998; McClelland & Elman, *Cognitive Psychology* 18:1–86, 1986). Because the presence of those competitors influences word recognition, much research has sought to quantify the processes of lexical competition. Metrics that quantify lexical competition continuously are more effective predictors of auditory and visual (lipread) spoken word recognition than are the categorical metrics traditionally used (Feld & Sommers, *Speech Communication* 53:220–228, 2011; Strand & Sommers, *Journal of the Acoustical Society of America* 130:1663–1672, 2011). A limitation of the continuous metrics is that they are somewhat computationally cumbersome and require access to existing speech databases. This article describes the Phi-square Lexical Competition Database (Phi-Lex): an online, searchable database that provides access to multiple metrics of auditory and visual (lipread) lexical competition for English words, available at www.juliastrand.com/phi-lex.

Keywords Spoken word recognition · Lipreading · Lexical competition

When a speaker utters a word, the listener is tasked with matching that acoustic–phonetic input to a lexical item stored in memory. Given the size of the mental lexicon and the speed with which speech occurs, spoken word recognition is an impressive feat of human cognition. Although models of spoken word recognition differ in the details of how recognition occurs, there is general agreement about two features of the process: Input simultaneously activates multiple lexical items in memory, and these items then compete for recognition (see Weber & Scharenborg, 2012). The most common method for quantifying which items are activated in parallel has been to identify a group (a "neighborhood") of words that are perceptually similar to a stimulus word (Newman, Sawusch, & Luce, 1997; Tye-Murray, Sommers, & Spehar, 2007; but see Luce & Pisoni, 1998, and Strand & Sommers, 2011, for alternative approaches). These "neighbors" include any word that can be formed by a single phoneme addition, deletion, or substitution from the stimulus word. For example, neighbors of "cat" include "cot" (a substitution), "at" (a deletion), and "cast" (an addition). Words with many neighbors are subject to more lexical competition, and this competition comes at a cost: Stimulus words with many neighbors are recognized more slowly and less accurately than words with few neighbors (Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998; Strand & Sommers, 2011; Vitevitch & Luce, 1998).

Measures based on neighborhood size have proved to be very informative, but their binary nature has limitations. Words are classified as "neighbors" or "not neighbors" of a stimulus word, and therefore it is not possible to specify which neighbors are especially perceptually similar to the stimulus word and which are less similar.[1] Implicit in the neighborhood-based metric is the assumption that each neighbor provides an equivalent amount of competition (e.g., that "at" and "cut" both provide the same amount of competition for "cat"). Also implicit is the assumption that words that differ by more than one phoneme provide no

[1] In visual (written) word recognition, others (Yarkoni, Balota, & Yap, 2008) have identified limitations of categorical measures of competition and demonstrated that continuous measures account better for human word recognition.

J. F. Strand (✉)
Department of Psychology, Carleton College,
Northfield, MN 55057, USA
e-mail: jstrand@carleton.edu

Published online: 11 June 2013

competition for the stimulus word (e.g., although "cast" competes with "cat," "clad" does not). Because many models of recognition (e.g., Luce, Goldinger, Auer, & Vitevitch, 2000; Luce & Pisoni, 1998; McClelland & Elman, 1986) assume that activation is graded (with the degree of activation being based on the degree of perceptual similarity), these assumptions may not adequately explain the processes of lexical activation and competition. The models describe a structure in which perceptually salient differences between words create a continuum, but neighbor-based approaches ignore this by categorizing words as being either inside or outside the neighborhood boundary.

To overcome the limitations of categorical metrics of competition, other measures have been introduced that quantify perceptual similarity—and thus, lexical competition—continuously (Luce & Pisoni, 1998; Strand & Sommers, 2011; see also Strauss, Harris, & Magnuson, 2007). In these metrics, the perceptual similarity of a stimulus word and each of its competitors is quantified on the basis of the degree to which the phonemes of a stimulus word are perceptually confusable with the phonemes of a competitor word. Unlike the categorical measures, these continuous measures quantify the amount of perceptual similarity, and therefore, the expected competition that each word exerts on the stimulus word. For example, "pat" and "bat" are both neighbors of "cat," but /p/ and /k/ are more perceptually confusable than /b/ and /k/ (likely because /b/ differs from /k/ in both place of articulation and voicing, whereas /p/ differs only in place of articulation). Therefore, when "cat" is presented, "pat" will receive more activation, and therefore provide more competition than "bat" does. Instead of counting how *many* words compete with the stimulus word (as neighborhood-based metrics do), continuous measures quantify how *much* the competitors compete with the stimulus word. Continuous measures of lexical competition explain significant unique variance in spoken word recognition, beyond that accounted for by categorical, neighborhood-based metrics (Feld & Sommers, 2011; Strand & Sommers, 2011). This predictive power comes both from quantifying lexical competition continuously and from including a larger subset of the lexicon as possible competitors (Strand & Sommers, 2011).

Although models of word recognition were designed for auditory speech, they have also been adapted to describe visual spoken word recognition (lipreading; Feld & Sommers, 2011; Iverson, Bernstein, & Auer, 1998; Tye-Murray et al., 2007). Whereas auditory competitors are words that *sound* similar (e.g., "cat" and "cap"), visual competitors are those that *look* similar on a speaking face (e.g., "fat and "vat"; Binnie, Montgomery, & Jackson, 1974). Because of the nature of the auditory and visual signals, words that are highly perceptually similar in one modality may not be in another. For example, information about place of articulation is distorted by noise but is easy to identify on the face (Binnie et al., 1974). Therefore, "cat" and "pat" are easily confusable in the auditory domain but look very different on a talking face, so they are not likely to be confused while lipreading. Thus, the amounts of lexical competition that a word encounters in the auditory and visual domains are not correlated (Strand & Sommers, 2011). As a result, it is necessary to determine modality-specific measures of lexical competition in order to predict auditory and visual word recognition.

Although the composition of a stimulus word's competitors may differ in the auditory and visual domains, the effects that these competitors have are the same: Words that are visually similar to many other words are lipread less accurately than those that are similar to few other words (Mattys, Bernstein, & Auer, 2002; Tye-Murray et al., 2007). Lexical competition in visual word recognition has been modeled both categorically, by counting the number of visual neighbors (Iverson et al., 1998; Mattys et al., 2002), and continuously, following the process used for auditory competitors (Auer, 2002; Feld & Sommers, 2011; Strand & Sommers, 2011). As in the auditory domain, continuous measures account for significant unique variance in word recognition accuracy, beyond that accounted for by categorical, neighborhood-based metrics (Feld & Sommers, 2009; Strand & Sommers, 2011).

Although continuous measures are effective predictors of spoken word recognition in both the auditory and visual domains, they are somewhat computationally cumbersome. In addition, existing databases offer access to categorical measures of lexical competition (Balota et al., 2007; Sommers, 2002), but currently no publically accessible databases provide continuous measures of lexical competition in the auditory domain, nor do any databases include measures for lexical competition in the visual domain. This article describes the Phi-square Lexical Competition Database (Phi-Lex), an online, searchable database that contains multiple values for auditory and visual lexical competition for nearly 5,000 words. The goal of this database is to make continuous measures of auditory and visual lexical competition easily accessible to the scientific community.

Construction of the database

Phonological transcriptions and frequency-of-occurrence data for 40,481 English words were obtained from an existing database, the English Lexicon Project (ELP; Balota et al., 2007). The ELP contains words that range from one (e.g., "I") to 17 (e.g., "electrocardiogram") phonemes long. Longer words tend to be more perceptually distinct and to have less variability in distinctiveness (Storkel, 2004). For example, three-phoneme words in the ELP database have an average of 19.66 neighbors (*SD* = 10.18); four-phoneme words have an average of 8.65 neighbors (*SD* = 6.09); for five-phoneme words, the average number drops to 3.00 (*SD* = 3.27); and six-phoneme words have an average of just 1.0 neighbor (*SD* = 1.54). In addition, given that the majority of studies of spoken word recognition, and especially those concerned with lexical competition, are conducted on shorter words, the database is restricted to three- and four-phoneme words (referred to as the *target words*; $N = 4{,}864$).

For each of the target words, multiple measures of auditory and visual lexical competition were generated. Each of these measures makes somewhat different assumptions about the nature of lexical competition; including all of them would allow those assumptions to be tested. Most users of Phi-Lex will not wish to retrieve values for every metric of competition, but multiple measures are included in order to make the database as versatile as possible. Although categorical measures of lexical competition are available elsewhere (Balota et al., 2007; Sommers, 2002; Vaden, Halpin, & Hickok, 2009), they are included here because the number of neighbors that a word has depends on the idiosyncrasies and phonological transcriptions of the reference lexicon being used. In order to provide more direct comparisons between categorical and continuous measures, and between auditory and visual word recognition, all of the measures were calculated from the same reference lexicon ($N = 7{,}000$), which includes all two- to five-phoneme words with one vowel from the ELP. It is certainly possible that words with six or more phonemes could provide some competition for the target words. However, because of the data set used to calculate continuous measures of competition (see the auditory Continuous Measures section below), it was not possible to quantify the amount of competition between words with different numbers of vowels. Given that all of the target words have one vowel and words with six or more phonemes from the ELP have an average of 2.13 vowels, the reference lexicon was restricted to words with fewer than six phonemes. Brief descriptions of the measures are given in Table 1; a more complete discussion of each follows.

Notes on homophones

A large number of English words are pronounced the same way but are spelled differently (e.g., "write" and "right"; but see Gahl, 2008). Given this fact, the ELP—and therefore the reference lexicon—includes multiple entries for phonologically identical words. This complicates calculations of lexical competition, because it is unclear whether each separate spelling of a homophone should serve as a competitor (e.g., for the target word "tight," should "write" and "right" serve as one competitor or two?). Aurally perceived homophones provide activation for both semantic interpretations [e.g., /do/ primes both "bread" (*dough*) and "deer" (*doe*); Fleming, 1993], which may suggest multiple, separate entries for homophones in the mental lexicon. If homophonous words have multiple entries in the mental lexicon, we might expect measures of lexical competition that include all entries of homophones to account for more variance in word recognition than do measures that do not include all homophones. In the current database, measures of lexical competition were calculated using the 7,000-word reference lexicon, which lists homophones as separate entries. In addition, all measures of lexical competition were calculated using a subset of that lexicon ($N = 6{,}296$ words) that excluded multiple entries of homophones (but included the most frequently occurring homophone). For example, although the 7,000-word lexicon contains both "right" and "write," the 6,296-word lexicon excludes "write" (the less common member of the homophone pair). In the online database, measures calculated with this limited lexicon are labeled "no homophones" or "NH."

Measures of auditory lexical competition in the database

Categorical measures Auditory neighborhoods were built for each of the target words by counting the numbers of words in the reference lexicon that could be formed by the addition, deletion, or substitution of a single phoneme (Density B [a_denb][2]). In addition, a subset of those words that include only phoneme substitutions was also calculated for each target word (Density A [a_dena]; Large & Pisoni, 1998; Nusbaum, Pisoni, & Davis, 1984; Sommers, 2002). For example, the Density B neighbors of "cat" include "cast, "at," "mat," and "bat," but of those, only "mat" and "bat" would be included in Density A. The numbers of Density A and Density B values calculated here correlate highly with values available elsewhere (Sommers, 2002): $r = .88$ ($f^2 = 3.43$) for the two measures of Density A, $r = .89$ ($f^2 = 3.81$) for the two measures of Density B, $p < .01$. Although Density B values are more commonly used than Density A values, Density A values are included in order to more closely parallel analyses in the visual domain, which are based only on phoneme-substitution neighbors (see below). In addition to counting the numbers of Density B and Density A neighbors, the average log frequency of occurrence for the neighbors (adopted from Lund & Burgess, 1996) is included in the database for each measure [a_denb_freq and a_dena_freq]. Finally, for each target word, the numbers of higher-frequency neighbors were calculated for both Density B and Density A [a_denb_hfn and a_dena_hfn]. This measure may give additional information about the neighborhood composition. For example, "vet" and "ridge" have the same frequency (8.9) and number of neighbors (17), but 12 neighbors of "vet" are of higher frequency than it is, whereas only six neighbors of "ridge" are of higher frequency.

[2] Labels used in the database are indicated in brackets throughout the text.

Table 1 Summary of measures of lexical competition

Auditory	Categorical	a_denb	Number of neighbors
		a_denb_freq	Average frequency of neighbors
		a_dena	Number of substitution-only neighbors
		a_dena_freq	Average frequency of substitution-only neighbors
		a_denb_hfn	Number of higher frequency neighbors
		a_dena_hfn	Number of higher frequency substitution-only neighbors
	Continuous	a_psum	Phi-square density
		a_pskew	Phi-square skew
		a_pkurt	Phi-square kurtosis
		a_psd	Phi-square standard deviation
		a_pfwsum	Phi-square frequency-weighted density
		a_wt_psum	Phi-square density of words within type (e.g., CVCs for a CVC target)
		a_wt_pskew	Phi-square skew of words within type
		a_wt_pkurt	Phi-square kurtosis of words within type
		a_wt_psd	Phi-square standard deviation of words within type
		a_wt_pfwsum	Phi-square frequency weighted density of words within type
Visual	Categorical	v_lec	Size of lexical equivalence class (number of visual neighbors)
		v_lec_freq	Average frequency of words within the lexical equivalence class
	Continuous	v_psum	Phi-square density
		v_pskew	Phi-square skew
		v_pkurt	Phi-square kurtosis
		v_psd	Phi-square standard deviation
		v_pfwsum	Phi-square frequency-weighted density

All measures were also calculated including homophonous entries only once

Continuous measures To represent the assumption that lexical competition is graded (with more similar words providing more competition), Luce and Pisoni (1998) devised a metric called *neighbor word probabilities* (NWPs) that quantifies lexical competition continuously. NWPs quantify how perceptually similar two words are on the basis of the conditional probability that a target word's segments would be confused with the segments of a competitor word (Luce, 1986; Luce & Pisoni, 1998). For example, the probability that a listener would perceive "bid" when presented with "cat" was quantified as $\text{NWP}(\text{bId} \mid \text{kæt}) = p(\text{b} \mid \text{k}) * p(\text{I} \mid \text{æ}) * p(\text{d} \mid \text{t})$, where the p values are based on the likelihood that two phonemes would be confused for one another on a forced choice phoneme identification task. Therefore, words that have more confusable segments and more overlapping segments would be expected to provide greater competition for one another than do those that differ by multiple segments or that have dissimilar segments. For example, "cat" and "cap" would have a high NWP because "t" and "p" are reasonably confusable (Miller & Nicely, 1955), whereas "cat" and "lore" have a very low NWP because they differ by multiple phonemes, and those phonemes are not very confusable with one another.

Measures based on NWPs account for significant variance in word recognition accuracy (Luce, 1986; Luce & Pisoni, 1998). However, using probabilities of confusion as a metric for perceptual similarity has several limitations (see Iverson et al., 1998, and Strand & Sommers, 2011, for more detailed discussion of these issues). First, they are susceptible to response biases. If a participant has a tendency to respond with a certain phoneme for reasons unrelated to the task (e.g., regularly pressing the key closest to the resting position), the probabilities of confusion would contain artifacts. Second, the number of perceptually similar alternatives interacts with the probabilities. Hypothetically, if /m/ and /n/ are perceptually very similar to each other and not to any other phonemes, they should be expected to be confused on approximately 50% of trials. If /p/, /t/, /k/, and /θ/ are very similar to one another but not to any other phonemes, any given pair should be expected to be confused on 25% of trials. This leads to the incorrect impression that /m/ and /n/ are twice as confusable as /t/ and /p/. These limitations prompted Iverson et al. to use the phi-square statistic as an alternative measure of phoneme similarity (see also Strand & Sommers, 2011, for prior discussion of these issues). The phi-square statistic (a normalized version of the chi-square statistic) makes it possible to calculate the perceptual similarity of phoneme pairs while reducing the problems with confusion probabilities. The phi-square statistic is expressed as:

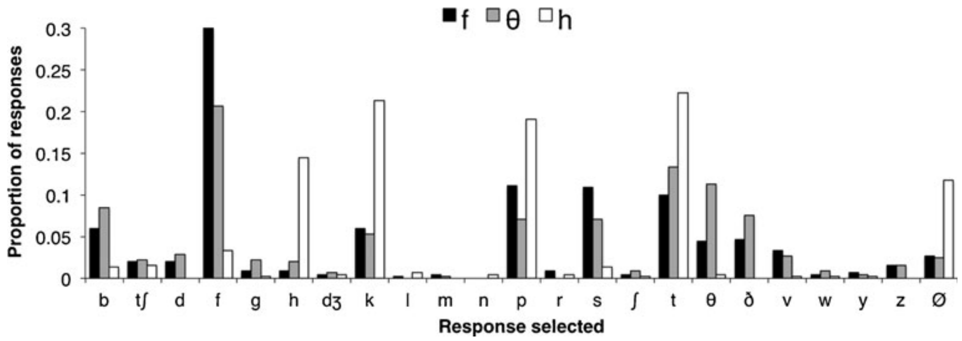

Fig. 1 The phonemes /f/ and /θ/ have more similar response distributions (and therefore a higher phi-square value) than do /f/ and /h/

$$\Phi^2 = 1-\sqrt{\frac{\sum\frac{(\mathrm{x_i}-\mathrm{E(x_i)})^2}{\mathrm{E(x_i)}}+\sum\frac{(\mathrm{y_i}-\mathrm{E(y_i)})^2}{\mathrm{E(y_i)}}}{\mathrm{N}}}.$$

Here, x_i and y_i are the frequencies with which the phonemes x and y were identified as phoneme i, $E(x_i)$ and $E(y_i)$ are the expected frequencies of response for x_i and y_i if the two phonemes are perceptually identical, and N is the total number of responses to phonemes x_i and y_i. The expected frequencies, $E(x_i)$ and $E(y_i)$, are determined by summing the frequency with which phoneme x was identified with category i and the frequency with which phoneme y was identified with category i, divided by 2. Therefore $E(x_i)$ and $E(y_i)$ are always equal. The rationale for this method is that if phonemes x and y are perceptually *identical*, they should be identified as members of a given category with equal frequency. Hypothetically, if /t/ and /p/ were perceptually identical, participants should choose evenly between them when they are presented with each other or with any other phoneme. The phi-square statistic reaches a value of 1 when the distributions of responses for two stimulus phonemes are identical (i.e., participants select each response alternative equally for both stimulus phonemes), and it reaches a value of 0 when the distributions have no overlap (i.e., participants do not use any of the same response categories for the two stimulus phonemes). Confusion probabilities quantify how regularly two phonemes are confused for one another; the phi-square value quantifies how similar the pattern of responses to the two phonemes are (see Fig. 1). A phoneme pair has a high phi-square value if the phonemes tend to be confused in similar ways.

To use phi-square values to assess the perceptual similarity of a word pair (e.g., to generate an NWP using phi-square values), the position-specific phi-square phoneme values for two words are multiplied (Feld & Sommers, 2011; Strand & Sommers, 2011; see also Luce & Pisoni, 1998). For example, the Φ^2NWP(bId | kæt) = Φ^2(b | k) * Φ^2(I | æ) * Φ^2(d | t). Word pairs with multiple shared phonemes and those with more perceptually similar phonemes render higher Φ^2NWPs. To quantify the overall competition that a target word encounters within the lexicon (called the *phi-square density*), the Φ^2NWPs of the target word and all other words in the lexicon are summed (e.g., Φ^2NWP(word$_1$ | target) + Φ^2NWP(word$_2$ | target) + Φ^2NWP(word$_N$ | target).[3] For example, to determine the total amount of competition the word "cat" encounters in a toy lexicon that consists only of four words, "cat," "cap," "bat," and "ten," Φ^2NWP values would be computed for cap | cat, bat | cat, and ten | cat. Of these, cap | cat has the highest value (because the only deviation comes from two phonemes with a high phi-square phoneme similarity). Bat | cat has a somewhat lower value, because /k/ and /b/ are less perceptually similar than /k/ and /p/. Although "cat" and "ten" share no position-specific phonemes, the phi-square similarity is still above zero, reflecting some similarity between at least one of the phoneme pairs (e.g., participants may occasionally confuse /ɛ/ and /æ/ for each other, or may occasionally choose /I/ when presented with /ɛ/ or /æ/). The phi-square density for "cat" in this toy lexicon is Φ^2NWP(cæp | cæt) + Φ^2NWP (bæt | cæt) + Φ^2NWP (tɛn | cat).

For the continuous measures of lexical competition, phi-square values were generated from an existing data set of forced choice phoneme confusions (Luce, 1986). This data set includes 22 consonants in the syllable-initial position (e.g., /dæ/), 21 consonants in the syllable-final position (e.g., /æd/), and 15 vowels from the corresponding syllable-initial and syllable-final positions. Participants (N = 122) made forced choice identifications amidst a background of white noise at 75 dB. In the original data set, phoneme confusions were obtained at three signal-to-noise ratios (SNRs): –5, +5, and +15. The rates at which phonemes were confused at all three SNRs were combined to generate phi-square values for the present analysis. The rationale behind collapsing across SNRs is two-fold. First, combining SNRs provides a general estimate of how confusable phoneme pairs are (as well as increasing the number of observations for each phoneme pair), which may then be applied to other SNRs, and to predict confusability for

[3] Because phi-square values are generated from probability-of-confusion values, the two are highly correlated, but measures that use Φ^2NWPs predict significant unique variance in word recognition accuracy after controlling for the influence of NWPs on the basis of probability of confusion (Strand & Sommers, 2011).

stimuli that are presented without masking noise or in writing (e.g., lexical decision or reading tasks). Therefore, this combined SNR is likely to be most useful to other researchers in a variety of settings. In addition, the confusions made at each SNR are highly correlated (*r* values range from .62 to .99; $f^2 >$ 0.66 and $p < .001$ for all), indicating that the patterns of confusions are relatively consistent across SNRs, even though overall accuracy differs (see Miller & Nicely, 1955, for additional evidence that the types of confusions made are relatively stable across SNRs).

Although there are many published data sets of phoneme confusions, these data were chosen as the basis for the development of the auditory-only measures of lexical competition for several reasons. First, identifications at multiple SNRs allow for measures of lexical competition that are more versatile. Second, the data set includes identification data for both consonants and vowels from the same participants in the same conditions. This consistency makes combining the consonant and vowel data to form NWPs less problematic. Third, it includes syllable-initial and syllable-final consonant identifications separately, which reflect any differences in phoneme confusability based on position. Finally, this data set has the benefit of having included a "null response" category in the consonant identification trials. Although participants heard a consonant-and-vowel pair (e.g., /bæ/) on the majority of trials, on some trials, the consonant was omitted (e.g., /æ/), but participants were still forced to make a response about what consonant was presented (or to indicate that no consonant was presented). As a result, the data set contains information about how likely participants were to report hearing each consonant when none had been presented [i.e., *p*(b | Ø)], and how likely participants were to report that no consonant was presented when one in fact was [i.e., *p*(Ø | b)]. This procedure is very helpful in the present analysis, because it allows a target word to be compared to stimulus words of differing lengths. For example, it is possible to compare "brat" and "rat" by aligning the vowels and including the null-response category for the extra phoneme, Φ^2NWP(bræt | ræt) = Φ^2(Ø | b) * Φ^2 (r | r) * Φ^2 (æ | æ) * Φ^2 (t | t). Therefore, the word similarities are multiplied by the likelihood that a participant would report hearing /b/ when no consonant was presented. Words that differ by many segments may be compared, as well. For example, a consonant–vowel–consonant (CVC) word ("cat") may be compared to a VCC word ("ant") by lining up the vowels as follows: Φ^2(Ø | k) * Φ^2(æ | æ) * Φ^2(n | t) * Φ^2(t | Ø). This method works well for consonants, because the vowel context can be presented without a consonant (e.g., /æ/ in place of /bæ/). However, it is more problematic for vowels, because it would be difficult to occasionally omit a vowel and ask a listener to identify the consonant (e.g., /b/ in place of /bæ/). Therefore, the Luce (1986) data set did not include a null response option in the vowel task, making it impossible to compare words that differed in numbers of vowels (e.g., "bait" and "abate").

The Luce (1986) confusion matrices were converted to phi-square values using SPSS (version 19).[4] For each of the 4,864 target words, Φ^2NWPs were calculated by comparing the target word to the 7,000 other words in the reference lexicon. Next, all of the Φ^2NWPs for each target word were summed in order to generate the auditory phi-square density [a_psum]. This measure represents the total amount of competition that each target word encounters. A frequency-weighted measure of phi-square density was also calculated [a_fwpsum], in which each Φ^2NWPs was weighted by the competitor word's log frequency of occurrence (from Lund & Burgess, 1996). For example, when compared with "cat," the words "catch" and "hath" both have a Φ^2NWP of .27, but because "catch" is much more frequent than "hath," it contributes a larger value in the frequency-weighted phi-square density of "cat." Words that have perceptually similar, high-frequency competitors should be expected to undergo greater lexical competition than those with perceptually similar, low-frequency competitors (Luce & Pisoni, 1998).

Although phi-square density describes the total amount of competition that a word encounters, it does not describe whether the bulk of that competition comes from few, highly perceptually similar competitors, or many, less perceptually similar competitors. For example, "seat" and "dose" have very similar phi-square densities (39.8 for both), but for "seat," much of the phi-square density comes from a cluster of highly similar words (e.g., "feet," and "seep"), whereas "dose" has few highly similar competitors, so the phi-square density comes from a greater number of less-similar words (e.g., "dope," "dough," "deuce," "boat," and "both"). Models of word recognition do not make predictions about how these differences should affect word recognition, but including values that pertain to dispersion, as well as the total summed density, addresses the issue. Therefore, for each target word, values for the standard deviation [a_psd], kurtosis [a_pkurt], and skew [a_pskew] of the Φ^2NWPs were also calculated. The difference in the distributions of "seat" and "dose" is reflected in the standard deviation: "Dose" has a relatively low standard deviation (.014), reflecting the fact that the competition is dispersed, whereas "seat" has a relatively high one (.022).

Finally, all values (sum, standard deviation, skew, and kurtosis) were calculated by comparing each target word only to words with the same consonant/vowel pattern type (e.g., a CVC compared with only CVCs). These values are designated "_wt" (within type) in the database. This method is somewhat akin to Density A, because only words that are

[4] This is a standard command in SPSS, and can be completed using the following syntax: PROXIMITIES VAR00001 . . . /VIEW = CASE /MEASURE = PH2/STANDARDIZE = NONE. This generates a matrix of dissimilarities, so the values are then subtracted from 1 to create a matrix in which higher values correspond to greater similarity.

formed from phoneme substitutions are included as competitors (but not words that require phoneme additions or deletions). The within-type measures allow for more direct comparisons with some of the continuous measures of vowel-only perception, which are limited to within-pattern analyses (see below).

Significant correlations emerged between many measures of lexical competition, reflecting the fact that these metrics are attempting to quantify the same underlying variables. For example, Density B is significantly correlated with phi-square density, $r(4862) = .68$, $f^2 = 0.86$, $p < .001$. However, this correlation is partially an artifact, driven by the fact that different word pattern types (e.g., CVC) have very different ranges for both measures of lexical competition. For example, CCCV words have an average of 4.55 Density B neighbors, whereas CVCs have an average of 26.89. Therefore, the correlation between phi-square density and Density B that includes both CVCs and CCCVs partially reflects the relationship between phi-square density and Density B, and partially reflects the different distributions of values of the two words types. When the word type is limited to just CVCs, for example, the correlation between phi-square density and Density B drops considerably, $r(1634) = .34$, $f^2 = 0.13$, $p < .001$, which is very similar to the values reported in earlier investigations (Strand & Sommers, 2011). This moderate correlation indicates that words that have a high value in one metric may have a low value in another. For example, "both" has 15 Density B neighbors (well below the mean of 26.9 for CVCs) but has a phi-square density of 65.8 (well above the mean of 42.4 for CVCs). Inversely, "nail" has a high value for Density B (47) but a relatively low phi-square density (26.1).

Measures of visual lexical competition in the database

Categorical measures A parallel to the categorical measures of auditory lexical competition (Density A and Density B) has also been created for the visual domain. Visual neighborhoods (also called *lexical equivalence classes*, or LECs) were constructed using the method described in Iverson et al. (1998) and Owens and Blazek (1985). First, clusters of phonemes that appear similar on their faces (phoneme equivalence classes, or PECs[5]) were identified (Iverson et al., 1998; Owens & Blazek, 1985; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). To calculate PECs, an existing set of visually presented phoneme confusions (derived from 72 participants identifying 24 consonants and 14 vowels) were employed (Strand & Sommers, 2011). These confusion matrices, which display the frequency with which each phoneme was identified as every other phoneme, were converted to phi-square values and entered in a hierarchical cluster analysis. The analysis generated a tree structure in which all phonemes are in a unique PEC at the lowest level of the structure and all phonemes are in a single PEC at the highest level. PEC groupings are defined as the level at which 70% of responses fall within the PEC (see Strand & Sommers, 2011, for more computational details). On average, 80% of vowel responses and 88% of consonant responses fall within PEC. For each target word, LECs were defined as any word that could be formed by substitutions within PEC. For example because /p/, /m/, and /b/ are part of the same PEC, and because /t/ and /d/ are part of another PEC, "bad," "pat," and "mad" would be in the same LEC. For each word, the size of the LEC was calculated [v_lec], as well as the average frequency of occurrence (Lund & Burgess, 1996) of the LEC members [v_lec_freq]. As expected, auditory and visual neighborhoods differed for the same target, reflecting modality-specific patterns of confusability. For example, for the target "cat," "gaze" and "get" are visual (but not auditory) neighbors, "pat" and "vat" are auditory (but not visual) neighbors, and "cad" and "hat" are in both the auditory and visual neighborhoods (for information on how information is integrated from auditory and visual neighborhoods during audiovisual speech perception, see Tye-Murray et al., 2007).

Continuous measures The calculations for visual phi-square density follow those of auditory phi-square density, but substitute the visual phoneme confusion matrices. In parallel to the auditory analysis, these phoneme confusions were converted to phi-square values, which were used to calculate Φ^2NWPs comparing the target word to every other word in the reference lexicon. From these Φ^2NWPs, the phi-square density [v_psum], frequency-weighted density [v_fwpsum], standard deviation [v_psd], skew [v_pskew], and kurtosis [v_pkurt] were calculated. In parallel to the analysis in the auditory domain, visual phi-square density and LEC size were significantly correlated, $r(4862) = .84$, $f^2 = 2.40$, $p < .001$, and dropped very slightly to $r(1634) = .82$, $f^2 = 2.05$, $p < .001$, for CVCs, reflecting the fact that CVCs have highly variable LEC sizes.

Unlike the auditory analysis, however, target words in the visual domain were only compared with words of the same pattern type (e.g., CVC targets compared with CVC competitors). Although the method of including a null response category [e.g., $p(\text{b} \mid \text{Ø})$] is easy to implement in auditory phoneme identifications, it is difficult to adapt this procedure for visual phoneme identifications. The challenge of visual phoneme identification is not determining *whether* something is present (a detection task), but rather determining *what* is present (an identification task). Although detection might be a challenge in the auditory domain (in which phonemes are presented masked in noise), the fact that a speech movement had been made is visually very salient, even if identification of that movement is difficult. In fact, it is difficult to imagine a scenario in which a

[5] The terms *viseme group* and *PEC* are synonymous, as are *LEC* and *homophene group*.

participant could watch a speaking face and respond that the face did not move. Therefore, the method of comparing words to competitors of different lengths is not applicable. Because words are compared only to words within a pattern type (e.g., CVCs are compared with other CVCs), some caution should be exercised in comparing the visual phi-square density values of words of different patterns. For example, many more CVCs are present in the reference lexicon than are CCCVs, so CVC target words will generally have higher values than CCCV target words simply as a function of there being more of them. This poses a limitation in the quantification process: Although it is not possible to quantify how words of different lengths compete in the visual domain, participants do occasionally confuse words that differ in length (e.g., respond "cast" for "cat"). Therefore, future work should seek to develop methods that parallel those allowed by the "null response" category in the auditory domain.

How to use the online interface

The online database may be accessed at www.juliastrand.com/phi-lex. First, enter the orthographic form of the word or words for which to generate data (phonological transcriptions are available using the ELP; Balota et al., 2007). Multiple words may be entered at once, separated by commas, spaces, or line breaks. Next, select the values to generate, either by selecting individual measures (e.g., v_lec) or classes of measures (e.g., all visual categorical measures). Place the cursor over each value attribute for more information. The output will be displayed as a table within the browser. Values for the pattern (i.e., CVC) and number of phonemes may also be generated.

Demonstrations of possible analyses

To demonstrate some uses of the values in the database, several analyses were conducted, using recognition accuracy in auditory word identification as the dependent measure (see Feld & Sommers, 2011; Strand & Sommers, 2011, for comparisons of categorical and continuous measures of lexical competition at predicting visual word recognition). Recognition data on 400 CVC words (randomly selected from the ELP; Balota et al., 2007) were collected from students at Washington University in St. Louis [$N = 50$; mean age = 21 years ($SD = 2.3$); 34 women, 16 men]. All of the participants had better-ear hearing thresholds below 25 dB HL at frequencies of 500, 1000, 2000, and 4000 Hz and reported English as their native language. They received $10 or course credit for their participation.

Participants read an information sheet, gave verbal consent, and were seated in a sound-attenuating booth (IAC 120A) approximately 0.5 m from a 17-in. Touchsystems monitor (ELO-170C) running Superlab (Version 4.0.7b, Cedrus Corporation, 2009) software. They were presented with audio clips of the 400 words in the carrier phrase "Say the word ______," which they identified by typing their responses on a keyboard. The speech materials were recorded by a Midwestern female speaker in a sound-attenuating chamber (IAC 120A) at 44100 Hz, 32 bits. The stimuli were recorded, edited, and equated for root-mean square amplitude using Adobe Audacity. All stimuli were presented via headphones (Beyerdynamic DT 770 Pro) at approximately 68 dB SPL amidst six-talker background babble at a –2 SNR. Prior to the analysis, recognition responses were hand-checked for obvious entry errors, such as a superfluous punctuation mark (e.g., "soup["). Entry corrections accounted for fewer than 1% of the responses. No other deviations from the stimulus words (plurals, inflected forms) were counted as correct. Eleven words were never identified correctly, so these were removed to exclude the possibility of faulty stimuli. Due to experimental error, one word was never presented to participants, so the analyses are on the 388 remaining words. For each word, values from Phi-Lex were obtained.

Comparing density B and phi-square density

A hierarchical multiple regression compared the efficacy of a categorical metric (Density B) and a continuous one (phi-square density) at predicting word recognition. Given the established influences of target word frequency, length (in milliseconds), and phonotactic probability (see, e.g., Vitevitch, Luce, Pisoni, & Auer, 1999), these values were entered into the model first, followed by Density B. Phi-square density was added in the final step, and it explained a significant proportion of unique variance in word recognition accuracy, beyond that accounted for by the other metrics (see Table 2).

One other study (Strand & Sommers, 2011) directly compared phi-square measures of competition with neighborhood-based metrics and found very similar results, with phi-square measures accounting for an additional 7% of unique variance beyond that explained by frequency and Density B. Critically, when phi-square density was entered in the second step, adding Density B in the third step failed to account for significant unique variance in word recognition accuracy. This analysis highlights the advantage of using a continuous rather than a categorical measure of competition.

Homophones as competitors

To test whether multiple entries of homophones act as separate competitors for the target word, another hierarchical multiple regression was conducted, again predicting

Table 2 Hierarchical multiple regression comparing the efficacy of Density B and phi-square density at predicting spoken word recognition

Variables	R^2	ΔR^2	β	f^2
Step 1:	.04	.04**		.04
Frequency			.21	
PhonProb			.02	
Length			–.02	
Step 2:	.06	.03**	–.09	.02
DensityB				
Step 3:	.16	.09**	–.33	.12
Phi-square density				

f^2 values represent the effect size when adding each step to the model, and βs reflect values at the final step. $^{**}p < .01$

recognition accuracy for the 387 CVC words. After controlling for target word frequency, word length, phonotactic probability, and phi-square density while *excluding* homophones, phi-square density *including* homophones accounted for a small but significant amount of unique variance in word recognition accuracy (see Table 3). If phi-square density including homophones was entered first, phi-square density excluding homophones did not explain significant additional variance. Although the amount of additional variance explained was small, this implies that multiple homophonous entries might provide competition for a stimulus word.

Dispersion

A final analysis explored how the dispersion of the distribution of Φ^2NWP values influenced word recognition accuracy. After controlling for frequency, phonotactic probability, word length, and phi-square density, a step was included in which phi-square *SD*, skew, and kurtosis were entered in a stepwise regression. Phi-square *SD* accounted for a small but significant additional 2% of the variance (see Table 4).

Table 3 Hierarchical multiple regression comparing the efficacy of measures including and excluding homophones at predicting spoken word recognition

Variables	R^2	ΔR^2	β	f^2
Step 1:	.04	.04**		.04
Frequency			.21	
PhonProb			–.01	
Length			.01	
Step 2:	.14	.10**	–.74	.12
A_Psum_NH				
Step 3:	.16	.02*	–1.1	.02
A_Psum				

f^2 values represent the effect size when adding each step to the model, and βs reflect values at the final step. $^{*}p < .05$, $^{**}p < .01$

Table 4 Hierarchical multiple regression comparing the efficacy of measures including and excluding homophones at predicting spoken word recognition

Variables	R^2	ΔR^2	β	f^2
Step 1:	.04	.04**		.04
Frequency			.21	
PhonProb			.02	
Length			–.03	
Step 2:	.15	.11**	–.08	.13
A_Psum				
Step 3:	.17	.02*	–.32	.02
A_PSD				

f^2 values represent the effect size when adding each step to the model, and βs reflect values at the final step. $^{*}p < .05$, $^{**}p < .01$

The beta weight of *SD* predicting accuracy was negative, indicating that words with more dispersion in the distribution of Φ^2NWPs were identified less accurately than words with NWPs more clustered around the mean. In the earlier example, words like "dose," whose phi-square densities come from many, less-similar competitors, are recognized more accurately than words like "seat," whose competition is derived from fewer, more-similar competitors.

Suggestions for future research Future studies might explore whether the novel findings demonstrated here (e.g., the effects of homophones and dispersion on auditory word recognition) also influence visually identified words. Analyses of this kind would further test the claim that the processes underlying word recognition are similar in the auditory and visual domains (Mattys et al., 2002), Future research might also investigate how auditory and visual phi-square density predict audiovisual word recognition. Tye-Murray et al. (2007) demonstrated that audiovisual word recognition depends jointly on auditory and visual neighborhood size, but no studies to date have attempted to model the integration of auditory and visual lexical competition using continuous measures.

Conclusions and notes on use

A large number of metrics are included in the present database so that researchers may test specific hypotheses about the nature of lexical competition, like those described above. For example, if one were exploring how multiple homophones provide independent competition, it would be useful to have the _NH measures in addition to the standard measures. If comparing auditory and visual competition, it would be more appropriate to use the _WT (within-type) measures of auditory competition, because all visual measures are calculated within pattern type. However, the large number of values might lead to concerns about the potential for bias: Researchers may opt to use a metric of competition that would produce a desired effect. Unless there is a specific theoretical justification for using another measure, researchers should opt for auditory or visual phi-square density (including homophones) as the default measures of lexical competition. This includes researchers whose main interests lie elsewhere and who are using Phi-Lex to control for measures of lexical competition (e.g., to match lists of words on relevant variables). In addition, because different perceptual features are salient in auditory and visual speech perception, measures of competition should only be applied to the appropriate modality (i.e., one should use auditory measures to predict auditory word recognition).

Measures of lexical competition based on the phi-square statistic have demonstrated success in predicting both auditory and visual spoken word recognition and overcoming the limitations of categorical measures (Feld & Sommers, 2011; Strand & Sommers, 2011). Because lexical competition influences many cognitive and linguistic processes, studies of reading and memory (e.g., Delattre, Bonin, & Barry, 2006; Perre, Pattamadilok, Montant, & Ziegler, 2009; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002; Ziegler & Ferrand, 1998) have often relied on measures of neighborhood size to equate stimulus lists while exploring other variables. Making metrics based on the phi-square statistic readily accessible to the scientific community will allow others to more effectively control for the influence of lexical competition when exploring other topics, and to test assumptions about the processes underlying spoken word recognition.

Author note I am grateful to Jon Strand for creating the online database, to Paul Mains and Sarah Meerts for insightful suggestions, and to Allison Cooperman, Susanne Gahl, and Brent Spehar for providing helpful comments on a previous version. Correspondence concerning this article should be addressed to J.F.Strand, Department of Psychology, Carleton College, Northfield, MN 55057 (email: jstrand@carleton.edu).

References

Auer, E. T. (2002). The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review, 9,* 341–347. doi:10.3758/BF03196291

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17,* 619–630.

Delattre, M., Bonin, P., & Barry, C. (2006). Written spelling to dictation: Sound-to-spelling regularity affects both writing latencies and durations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1330–1340. doi:10.1037/0278-7393.32.6.1330

Feld, J., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research, 52,* 1555–1565. doi:10.1044/1092-4388(2009/08-0137)

Feld, J., & Sommers, M. S. (2011). There goes the neighborhood: Lipreading and the structure of the mental lexicon. *Speech Communication, 53,* 220–228. doi:10.1016/j.specom.2010.09.003

Fleming, K. K. (1993). Phonologically mediated priming in spoken and printed word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 272–284. doi:10.1037/0278-7393.19.2.272

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language, 84,* 474–496. doi:10.1353/lan.0.0035

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. doi:10.1016/0749-596X(89)90009-0

Iverson, P., Bernstein, L. E., & Auer, E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26,* 45–63. doi:10.1016/S0167-6393(98)00049-1

Large, N., & Pisoni, D. B. (1998). *Subjective familiarity of words: Analysis of the Hoosier Mental Lexicon (Research on Spoken Language Processing, Progress Report No. 22)*. Bloomington, IN: Indiana University, Psychology Department, Speech Research Laboratory.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon (Research on Spoken Language Processing, Progress Report No. 6)*. Bloomington, IN: Indiana University, Psychology Department, Speech Research Laboratory.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception, 62,* 615–625. doi:10.3758/BF03212113

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36. doi:10.1097/00003446-199802000-00001

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203–208. doi:10.3758/BF03204766

Mattys, S. L., Bernstein, L. E., & Auer, E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics, 64,* 667–679. doi:10.3758/BF03194734

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86. doi:10.1016/0010-0285(86)90015-0

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352. doi:10.1121/1.1907526

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology. Human Perception and Performance, 23,* 873–889. doi:10.1037/0096-1523.23.3.873

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, IN: Indiana University, Psychology Department, Speech Research Laboratory.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28,* 381–393.

Perre, L., Pattamadilok, C., Montant, M., & Ziegler, J. C. (2009). Orthographic effects in spoken language: On-line activation or phonological restructuring? *Brain Research, 1275,* 73–80. doi:10.1016/j.brainres.2009.04.018

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1019–1034. doi:10.1037/0278-7393.28.6.1019

Sommers, M. S. (2002). Washington University Speech & Hearing Lab Neighborhood Database. Retrieved from neighborhoodsearch.wustl.edu/Home.asp

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research, 47,* 1454–1468. doi:10.1044/1092-4388(2004/108)

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *Journal of the Acoustical Society of America, 130,* 1663–1672. doi:10.1121/1.3613930

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39,* 19–30. doi:10.3758/BF03192840

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11,* 233–241.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. Retrieved from www.iphod.com

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325–329. doi:10.1111/1467-9280.00064

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. doi:10.1006/brln.1999.2116

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130–145.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *WIREs Cognitive Science, 3,* 387–401. doi:10.1002/wcs.1178

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15,* 971–979. doi:10.3758/PBR.15.5.971

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review, 5,* 683–689. doi:10.3758/BF03208845