Journal of Memory and Language **44,** 568–591 (2001)
doi:10.1006/jmla.2000.2756, available online at http://www.academicpress.com on IDEAL®

Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?

Todd M. Bailey

University of Oxford, Oxford, United Kingdom

and

Ulrike Hahn

University of Wales, Cardiff, United Kingdom

Wordlikeness, the extent to which a sound sequence is typical of words in a language, affects language acquisition, language processing, and verbal short-term memory. Wordlikeness has generally been equated with phonotactic knowledge of the possible or probable sequences of sounds within a language. Alternatively, wordlikeness might be derived directly from the mental lexicon, depending only on similarity to known words. This paper tests these two cognitively different possibilities by comparing measures of phonotactic probability and lexical influence, including a new model of lexical neighborhoods, in their ability to explain empirical wordlikeness judgments. Our data show independent contributions of both phonotactic probability and the lexicon, with relatively greater influence from the lexicon. The influence of a lexical neighbor is found to be an inverted-U-shaped function of its token frequency. However, our results also indicate that current measures are limited in their ability to account for sequence typicality.

Key Words: wordlikeness; phonotactics; token frequency; lexical neighborhood; sequence typicality.

Any speaker of English can tell that Zbigniew is not an English name. It is not that Zbigniew merely happens to not be in a mental dictionary of English names; there is a strong intuition that it could not be included in any such dictionary. Even among actual words of English, some words sound more typical than others, for example *rat* as opposed to *sphere*, *splurge*, or *flail*. Speakers have consistent intuitions about whether a sequence of speech sounds could be a word of their native language and how typical it would sound as a word. Moreover, speech processing is highly sensitive to the typicality of sound sequences, as we will review. These effects can be seen in speech perception and production, lexical development, lexical access, and verbal memory. However, it is far from clear what makes a sequence of sounds more or less typical, more or less *wordlike*. Our own desire to understand the knowledge structures underlying wordlikeness was sparked initially by unsuccessful attempts to use common measures of wordlikeness for the principled construction of experimental stimuli. We were seeking an objective method of choosing sets of one-syllable words and nonwords covering a range of typicality, as would be desirable in a wide variety of psycholinguistic studies. Contrary to our expectations, standard measures of typicality did not generally correspond to our intuitions of wordlikeness in any real way. This prompted us to examine more closely the determinants of wordlikeness.

Sound sequence typicality is most often thought of as phonotactic probability, that is, the frequency with which a particular phoneme or phoneme sequence occurs in a language. Phono-

Todd Bailey was supported by grants from the McDonnell–Pew Centre for Cognitive Neuroscience, Oxford, and a grant to Kim Plunkett from the Biotechnology and Biological Sciences Research Council, UK. We thank John Coleman, Gordon Brown, and Dan Jurafsky for valuable discussions, and we are grateful to our anonymous reviewers for helpful comments. A preliminary analysis of Experiment 1 appeared in the Proceedings of the 20th Annual Meeting of the Cognitive Science Society.

Address correspondence and reprint requests to Todd Bailey or Ulrike Hahn, School of Psychology, Cardiff University, P.O. Box 901, Cardiff CF10 3YG, United Kingdom. E-mail: baileytm1@cardiff.ac.uk or hahnu@cardiff.ac.uk.

0749-596X/01 $35.00

tactic probability has been shown to correlate with performance on a variety of speech processing tasks. These include nonword repetition (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Vitevitch & Luce, 1998), naming (Levelt & Wheeldon, 1994), recall from verbal short term memory (e.g., Gathercole, Hitch, Service, & Martin, 1997; see Gathercole & Martin, 1996, for a review), phoneme identification (Pitt & McQueen, 1998), and wordlikeness ratings for nonwords (Coleman & Pierrehumbert, 1997; Dankovicova, West, Coleman, & Slater, 1998; Frisch, Large, & Pisoni, 2000; Gathercole & Martin, 1996; Vitevitch et al., 1997). In addition, Jusczyk, Luce, and Charles-Luce (1994) found that 9-month-old infants preferred to listen to nonwords with high phonotactic probability over nonwords containing low probability sequences. Indeed, knowledge of phonotactic probabilities may be crucial for infants to segment continuous speech into appropriate word-size chunks (e.g., Cairns, Shillcock, Chater, & Levy, 1997; Morgan & Saffran, 1995).

An entirely different kind of sequence typicality is the extent to which a sequence overlaps with individual known words. One measure of this overlap is neighborhood density, that is, the number of real words within a fixed phonological radius of the sequence. Neighborhood density has been reported to affect performance in a variety of tasks, including phoneme identification (Newman, Sawusch, & Luce, 1996), auditory lexical decision (Luce & Pisoni, 1998), and speech production (Vitevitch, 1997). An unpublished study by Martin and Gathercole (reported in Gathercole & Martin, 1996) found neighborhood density ratings to predict wordlikeness judgments. Similarly, Greenberg and Jenkins (1964) found that whether or not nonwords had a real word neighbor at a particular distance was a predictor of wordlikeness judgments.

In cognitive terms, lexical influences and disembodied, statistical knowledge of subword probabilities are two very different kinds of explanation for sequence typicality and its effects on linguistic processing (see also Treiman, 1988). The studies above manipulated either phonotactic probabilities *or* lexical neighborhoods. However, phonotactic probabilities and lexical neighborhoods may be confounded. It is widely recognized that words in high-density neighborhoods tend to have high-probability phonotactic patterns, and words in low-density neighborhoods tend to have low-probability phonotactic patterns (Charles-Luce & Luce, 1995; Frauenfelder & Schreuder, 1992; Gathercole & Martin, 1996; Jusczyk et al., 1993; Landauer & Streeter, 1973; Luce, Pisoni, & Goldinger, 1990; Vitevitch & Luce, 1998, 1999). This raises the question of which effects are genuinely due to phonotactic probabilities and which are due to lexical neighborhoods.

Moreover, it is conceivable that phonotactic knowledge is entirely implicit, that is, contained in our knowledge of individual words (Vitevitch et al., 1997; but see Vitevitch & Luce, 1998, 1999). For example, McClelland and Elman (1986) argued that apparent effects of phonotactic rules in phoneme identification were really the result of neighborhood effects in which similar words conspire to activate individual shared phonemes. Whether knowledge of phonotactic probabilities is genuinely distinct from lexical knowledge and stored separately is an issue which has also been given attention in the context of memory (Gathercole & Martin, 1996).

Whether the typicality of sound sequences is a reflection of subword phonotactic probabilities or lexical neighborhoods is a question which can only be answered by directly comparing them on the same task. The only study examining this question is Bailey and Hahn (1998), a preliminary investigation which is a precursor to the material presented here. Frisch et al. (2000) separately considered both phonotactics and lexical influences, but did not examine whether one might in fact subsume the other. Thus, the interpretation of previous research on sound sequence typicality remains unclear. Given the suspected correlations between phonotactics and lexical knowledge, any experimental evidence for the role of phonotactics in determining sequence typicality might actually be due to lexical effects or vice versa. To see whether typicality reduces to phonotactics or to lexical neighborhoods when both alternatives are considered was a central goal of the present study. Also, given our suspicion that phonotactic

measures might explain rather less about sequence typicality than desired, we wondered whether lexical neighborhoods might be more influential. We thus sought to estimate the amount of variance in wordlikeness which could be accounted for by phonotactic knowledge or lexical effects, or the combination of both.

These goals require a departure from the methodology of previous research. Past studies have generally sought to establish that high and low typicality sequences give rise to differential effects. For this, it is sufficient to choose a measure of typicality, select two groups of sound sequences which differ according to this predictor, and demonstrate that the groups lead to significant differences in some experimental task. This method is insufficient for the questions we are pursuing here. First, lexical and sublexical explanations refer to two cognitively distinct types of account, each of which might be instantiated in many different ways. Consequently, a general contrast between these two requires that a range of measures, both of phonotactics and of lexical neighborhood, is considered. In principle, one might try to identify a single stimulus set which factorially controls all predictor variables of interest. However, in practice it is very difficult to find orthogonal subsets of pronounceable nonwords to control a large number of correlated variables. Second, evaluating the comprehensiveness of lexical and sublexical explanations requires an estimation of the unique contribution of the competing measures to the prediction of typicality on a representative sample of sound sequences. In other words, one must examine the predictive power of lexical and phonotactic measures, both individually and jointly, on a random sample in order to draw conclusions about the sound sequences of a language more generally.

The research reported in this paper begins by identifying a set of measures of sequence typicality, including a new model of lexical neighborhood. Using wordlikeness judgments from two experiments, we examine whether ratings of sequence typicality are best explained by phonotactic probabilities or by similarity to known words. We find independent effects of both, but lexical measures, particularly our new measure of lexical neighborhood, prove to be better predictors than the phonotactic measures. However, none of the measures fully explain participants' wordlikeness judgments. This points to limitations in present conceptions of either phonotactics or lexical influence. Consequently, we proceed with an exploration of possible avenues along which better measures might be sought. Although modest improvements can be made in phonotactic measures, we conclude that more important contributions will likely come from better models of lexical neighborhoods.

CONTRASTING PHONOTACTICS AND LEXICAL KNOWLEDGE

The relationship between phonotactics and lexical knowledge will be sensitive to the way in which each is measured. To clarify the relationship generally, it is thus necessary to consider the widest possible range of phonotactic and lexical neighborhood measures. We first introduce our central phonotactic measures.

Phonotactics

Phonotactic measures compile frequency profiles and co-occurrence statistics for sounds, which can be used to evaluate the probability of a novel sequence. We computed several measures which made different assumptions about the size and number of units to be taken into account (two or three phonemes or syllable onsets, nuclei, and codas). The most popular phonotactic measure by far is bigram probability. There are two ways of calculating this: simple co-occurrence of two sounds within a body of speech or transition probabilities between sounds. In practice the two are highly correlated (Gaygen, 1997), but important structure is typically reflected more accurately in transition probabilities (see Aslin, Saffran, & Newport, 1998). Also, transition probabilities reflect the temporal nature of speech. For these reasons, we focus our analyses on transition probabilities. A bigram transition probability is based on the number of times, say, /t/ occurs after /o/, etc., across all English words and results in a conditional probability for each sequence of two sounds. We treated word boundaries as segments in order

to capture potential differences in probabilities at different positions within words (Nelson & Nelson, 1970; Sendlmeier, 1987). To calculate a composite value for an entire word or nonword, we took the geometric mean of conditional segment probabilities across the whole item, giving a single average bigram probability.

Of course, sensitivity to co-occurrence statistics need not be limited to pairs of sounds. We computed trigram transition probabilities analogous to the bigram probabilities by calculating the conditional probability of each segment given the preceding two segments. Again, word boundaries were included in the computation.

Bigram and trigram probabilities assume that the phoneme is the relevant phonological unit over which sequence probabilities are determined. In order to test for effects of probabilities computed across larger phonological units, we computed a syllable part probability metric, taking syllable onset, nucleus, and coda to be the basic units of analysis. The syllable part probability for a whole word was the geometric mean of the conditional probabilities of its onset in word-initial position, its nucleus following its word-initial onset, its coda following its nucleus, and the word-end following its coda.

Because we were interested in computing phonotactic probabilities for monosyllables only, our phonotactic measures are not concerned with issues such as stress and word position.

Orthotactics

Since one of our tasks was a written one, we also calculated orthotactic probabilities, that is transition probabilities for letter sequences rather than for sounds. However, given that a substantial amount of our language processing involves reading, influences of orthotactic patterns cannot be ruled out even in purely oral tasks. It is therefore an interesting empirical question whether effects of orthotactic probabilities are observed in oral as well as written tasks. To measure orthotactic probabilities, we computed orthotactic bigram and trigram probabilities which were directly analogous to the phonotactic measures described above. Because of the difficulties involved in identifying appropriate larger orthographic units for all English words automatically, we did not compute orthotactic syllable part probabilities.

Lexical Neighborhoods

Our other family of measures conceives of sequence typicality as based on the extent to which a sound sequence overlaps with words in the mental lexicon.

Neighborhood density. The standard measure of lexical neighborhoods is based on the single phoneme edit distance (Luce, 1986). By this metric, a neighbor is any word that can be derived by substituting, deleting, or inserting a single phoneme. For example, the English word *cat* has among its neighbors *mat*, *at*, and *scat*. The number of such neighbors (NNB) is the neighborhood density of an item. Most studies of neighborhood effects have used this simple notion of neighborhood density (e.g., Charles-Luce & Luce, 1990; Luce & Pisoni, 1998; Metsala, 1997; Newman et al., 1996; Vitevitch, 1997; Vitevitch & Luce, 1998). Although it is universally recognized as a crude approximation, NNB has been surprisingly successful in the study of lexical processing. This success may be due, in part, to the fact that the application of NNB has centered on words with a simple CVC pattern.[1] Relatively few works have applied NNB in the context of stimuli including consonant clusters, although clusters are not at all uncommon in natural language (some 67% of English monosyllables in the CELEX online dictionary of English contain at least one cluster of consonants; cf. www.kun.nl/celex). The NNB measure fails to take similarity *between* phonemes into account; replacing /b/ with /p/ (a change involving only voicing) yields a neighbor just the same as does replacing /b/ with /s/ (involving place and manner of articulation in addition to voicing). Furthermore, NNB has a sharp cutoff, simply ignoring all words outside the single phoneme edit distance, whether they differ by a single feature on each of two phonemes or by the insertion of several syllables.

As a starting point, we used a variant NNB measure based on a two-phoneme edit distance,

[1] We are grateful to an anonymous reviewer for calling this to our attention.

which is better suited to the stimuli we used. Our stimuli had many consonant clusters and many had no real word neighbors just one phoneme different.[2] Given the limitations of NNB, however, more sophisticated measures of lexical influence are required if we are to obtain a valid comparison with knowledge of phonotactic probabilities, and if we wish to acquire more than a superficial understanding of neighborhood effects.

Generalized neighborhood model. Although lexical neighborhoods play a prominent role in psycholinguistics, there is no obvious alternative to NNB in the literature which could be used directly as a predictor of sequence typicality. A number of competing conceptions of neighborhood interactions have been proposed in the context of word recognition (e.g., Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Morton, 1979). Implicit in these models is some notion of neighbor. However, few provide an analytic measure of neighborhood influence which can be factored out of the model of word recognition and applied to other tasks. An exception is the Neighborhood Activation Model of Luce (1986; Luce, Pisoni, & Goldinger, 1990). In Luce's model of word recognition, neighbor similarity is based on experimentally derived phoneme confusability. However, because the set of phonemes and the contrasts in which they participate differ in British pronunciations from those in American pronunciations, Luce's confusability data could not be directly applied in our study.

To fill this gap, we developed the Generalized Neighborhood Model (GNM), an adaptation of the Generalized Context Model (GCM; Nosofsky, 1986) of classification based on similarity to exemplars. The GCM is based on a large literature relating generalization to distance in a mental space, for a variety of tasks including identification, old–new recognition, and categorization (e.g., Lamberts, 1998; Nosofsky, 1986, 1988, 1990; Shepard, 1987). The GCM has also been applied to inflectional morphology (Hahn & Nakisa, 2000; Nakisa & Hahn, 1996) and artificial grammar learning (Bailey & Pothos, 1998; Pothos & Bailey, 1997, 2000). In the GCM, neighbors vary on a continuous scale of similarity. Instead of imposing a sharp distinction between neighbors and nonneighbors, all relevant items are neighbors to some degree. Specifically, if the psychological distance between instances i and j is $d_{i,j}$, then the GCM model of perceived similarity of a probe i to a *set* of instances stored in memory is

$$S_i = \sum_j e^{-D \cdot d_{i,j}}$$

(this is the exponential form of the GCM; other variants adopt a gaussian function by squaring the distance term). The term D in the GCM similarity equation is a sensitivity parameter, determined by regression. This parameter determines how quickly similarity decreases as a function of distance from the exemplars.

The GNM is an adaptation of the GCM similarity component which is aimed specifically at words. Since token frequency has been shown to be an important variable in a wide range of lexical processing tasks (e.g., Luce, 1986; Whaley, 1978), we incorporated a frequency-weighting term into the similarity equation. Frequency effects are commonly modeled as a linear function of token frequency or log token frequency (cf. Luce et al., 1990; Nosofsky, 1988; Vitevitch & Luce, 1998). This treatment assumes that the effects in question change monotonically with increasing frequency. However, nonmonotonic frequency effects have been reported in the context of inflectional morphology (Bybee, 1995; Hahn et al., 1998; Moder, 1992). To allow for possible nonmonotonic effects, the GNM adopts a quadratic frequency weighting term. The complete GNM neighborhood similarity equation, with frequency weighting, is

$$S_i = \sum_j \left(Af_j^2 + Bf_j + C\right) e^{-D \cdot d_{i,j}},$$

where f_j is the log token frequency of neighboring word j, as given in the CELEX database (to include words listed in CELEX with

[2] An alternative way of extending the metric to longer words is to base it on a certain fraction of the phonemes in a word (e.g., a proportional change of 1/3 would be equivalent to a single phoneme change when applied to CVC words; cf. Frisch et al., 2000).

zero token frequency, a constant 2 was added to the citation frequency counts before taking logarithms).[3]

The GNM requires measures of the psychological distances between individual items (words or nonwords). Our primary model combines a linguistic measure of phoneme similarity with a standard edit distance metric to estimate these psychological distances (two other distance metrics will also be considered). Edit distance is a weighted sum of the substitutions, insertions, and deletions required to transform one representation (e.g., a word or a nonword) into another (e.g., a neighboring word). Although edit distance is usually used with equal weightings or costs for all substitutions, a more realistic measure of phonological differences should vary depending on the phonemes involved. The initial consonants of *bat* and *sat*, for example, are more different than the initial consonants of *bat* and *pat*.

In order to take phonological differences into account, we assessed the relative cost of substituting one phoneme for another based on the natural class lattice distance metric (Frisch, 1996). This measure of phoneme dissimilarity has been related to behavioral data including English speech errors, phonotactic constraints in Arabic, and acceptability judgments of Arabic pseudo-words (Frisch, 1996; Frisch, Broe, & Pierrehumbert, 1997). The natural class distance metric counts the number of shared and different natural classes of two phonemes (a natural class is a group of sounds sharing one or more linguistically significant phonetic characteristic). If S, D_1, and D_2 are the number of natural classes shared by two phonemes, the number of natural classes which the first phoneme has distinct from the second, and the number of natural classes which the second phoneme has distinct from the first, then the natural class distance between the phonemes is $d_{1,2} = (D_1 + D_2)/(S + D_1 + D_2)$. This metric ranges from 0 for identical phonemes to 1 for phonemes which have no features in common. Because the natural class distance metric is based on a theory of natural classes defined in terms of phonological features, it can be applied to any language whose phonemes have been classified and for which a pronunciation dictionary is available.

The relative cost of insertions and deletions compared to substitutions was determined empirically. Several different insertion–deletion costs were computed, and 0.7 was chosen because it gave the best fit to the oral wordlikeness data of Experiment 2 (a broad plateau in overall fit was observed, with fairly similar values across a range of insertion–deletion costs from 0.6 to 1.0).[4] In order to limit the number of free parameters, these preliminary fits were obtained without the frequency weighting term of the GNM. The costs for phoneme insertion and deletion were constrained to be of equal value in order to keep distances symmetrical (i.e., the distance between *cat* and *scat* would equal the distance between *scat* and *cat*).

THE WORDLIKENESS TASK

The comparison between phonotactic and lexical influences requires a suitable task within which to observe sequence typicality. As discussed above, many online processing tasks are influenced by sequence typicality, including nonword repetition, recall, and lexical decision. All of these are possible tasks in which the relative contribution of phonotactics and lexical measures might be assessed. The ideal starting point, however, is wordlikeness judgments, that is, direct ratings of sequence typicality (cf. Coleman & Pierrehumbert, 1997; Dankovicova et al.,

[3] GNM neighborhood similarity is similar in spirit to the frequency-weighted neighborhood confusability term in Luce's (1986) Neighborhood Probability Rule, but differs in using an exponential transformation from psychological distances instead of using confusion probabilities and in allowing for quadratic as well as linear frequency effects.

[4] The relative cost of insertion–deletion constitutes an additional free parameter in the edit distance version of the GNM model. Note that the mere presence of free parameters does not distinguish the GNM from phonotactic probabilities. Each metric of phonotactic probability operationalizes a theory that there is a linear relationship between that metric and wordlikeness, and we test that theory by fitting a model with one free parameter to find the best-fitting straight line. The GNM has not just one, but several free parameters. The extra complexity of the GNM is justified to the extent that it provides a better explanation for empirical data.

1998; Frisch et al., 2000; Gathercole & Martin, 1996; Vitevitch et al., 1997). This task is likely to be dominated by sequence typicality; by contrast, sequence typicality is likely to be just one of many contributing factors in a task such as auditory lexical decision. This means that a far greater part of the item variance can be expected to depend on sequence typicality for wordlikeness judgments than for reaction times in lexical decision, making the former more powerful to empirically distinguish the two different, but correlated, types of knowledge which potentially underlie sequence typicality. Given previous findings which relate wordlikeness ratings to recall from short term memory (Gathercole & Martin, 1996), in addition to Frisch et al.'s (2000) finding that predictors of wordlikeness ratings were also good predictors of recognition performance, we can reasonably expect our key findings to carry over to at least some online processing tasks.

EXPERIMENT 1

Experiment 1 compared phonotactic and lexical accounts in their ability to explain ratings of wordlikeness for a representative sample of monosyllabic nonwords. We used nonword stimuli in an effort to avoid potentially confounding attributes of real words, including semantic representations, age of acquisition, imageability, and other nonphonological characteristics which participants might inadvertently include in their judgments of typicality. We restricted our stimuli to monosyllables in order to avoid the complexities of stress placement and syllabification. Finally, we chose nonwords in clusters around a random sample of isolates to be able to generalize our conclusions beyond our stimulus set and infer properties of the wider population of nonwords from which our sample was drawn.

In this study, stimuli were presented in written form, but participants were asked to focus on how wordlike the stimuli sounded.

Method

Participants. Participants included 22 first-year psychology students from Warwick University who took part in the experiment as part of their Methods course, plus 1 other undergraduate and 1 graduate student who volunteered, for a total of 24 participants (17 female, 7 male).

Materials. Pronounceable wordforms (including real words as well as nonwords) were generated by a syllable formation grammar. This was based on onset–nucleus and nucleus–coda combinations found in one-syllable words in the CELEX English database. Each wordform was classified as a *word*, *near miss*, or *isolate*. Near misses were nonwords (e.g., *drump*) which differed from the nearest real word neighbor by a single phoneme. Isolates (e.g., *drolf*) differed from their nearest neighbor by exactly two phonemes.

A set of 22 isolates were chosen at random. In addition, for each isolate, we identified all neighboring near misses, that is, nonwords which were one phoneme different from the isolate and also one phoneme different from the nearest real word. By using near misses around the isolates, we can be sure that the distance to the nearest neighbor really is greater for isolates than for near misses, which is important for testing measures of neighborhood similarity. If near misses were chosen independent of the isolates, then potential differences in the importance of onsets versus rimes, for example, might be obscured if the isolates and near misses happened to differ in their distributions of various phonemes. Our method of stimulus selection eliminates this uncertainty by ensuring that the near misses lie midway on the edit path between our isolates and their nearest neighbors. Moreover, by using *all* the near misses around our isolates, we sample the greatest possible variety of edit differences and include both high and low probability phoneme sequences. Excluding items which could not be assigned a plausible and unambiguous spelling, 250 near misses were identified.

In addition to the isolates and near misses, 69 real word fillers were included in order to encourage participants to process the stimuli as linguistic entities and not as disembodied sound sequences. The complete set of 341 probes was randomized to produce four different orderings. Four additional orderings were produced by reversing the first four.

Procedure. We assigned participants to the stimulus lists at random, three participants per

list. Participants filled in a written questionnaire, which took about 20 minutes. The instructions asked participants to judge the sounds of the words, not their spellings, and to indicate the perceived typicality of each probe on a scale from 1 to 9. For example,

> Does minth sound like a typical word of English?
> Very non-typical = 1 2 3 4 5 6 7 8 9 = Very typical

Results

Real word fillers were excluded from analysis. Thirteen misspelled near misses were identified, and these data were dropped. This left 259 probes for analysis (22 isolates and 237 near misses), which are listed in the Appendix. Results from all eight stimulus lists were pooled, and raw ratings were scaled to the interval 0 to 1. Initial checks revealed that variance was related to mean item ratings, so that variance was highest for items toward the middle range of the rating scale. In order to reduce the dependency between mean and variance, individual scaled ratings were transformed using the arcsine transformation

$$x' = \frac{2 \arcsin \sqrt{x}}{\pi}.$$

The resulting transformed ratings ranged from 0 (nonwordlike) to 1 (maximally wordlike).

Across individual responses, differences between items were large relative to noise in the data, as measured by omega-squared (or, equivalently, intraclass correlation), $\omega^2 = .22$ (values greater than .138 are considered to be large effects; see, e.g., Kirk, 1995). This statistic means that about 22% of the variance in individual responses (as distinct from variance among items' ratings averaged across participants) was due to real differences in wordlikeness between items.[5] The remaining variance is due to sampling error or nuisance effects.

[5] To put ω^2 in perspective, it may be helpful to imagine how much variance among reaction times for individual key presses in an online task one would expect to be due to differences between items as opposed to nuisance noise. Unfortunately, since such ω^2 statistics are seldom reported, the exercise must be left to the imagination.

On average, the transformed ratings for near misses ($M = .44$, $SE = .006$, $n = 237$) were somewhat higher than those for isolates ($M = .38$, $SE = .015$, $n = 22$). Additional statistical analyses were performed to assess the effectiveness of sequence probabilities and neighborhood measures in explaining differences in ratings among items and to test whether the effects of those predictors were consistent across subjects. Given the design of our study, the way to assess the effectiveness of a particular metric in predicting differences among items is to compute R^2 values for item ratings, averaged across subjects. The way to test whether the effect of a metric is consistent across subjects is to do repeated measures ANOVA on individual ratings, not on means computed across subjects (see Lorch & Myers, 1990, for a very readable discussion of the statistical pitfalls associated with the common practice of performing significance tests on mean item ratings, including high Type I error rates and the implicit treatment of subjects as a fixed effect). In repeated measures ANOVA, the variance attributed to each predictor is evaluated against the corresponding subject by predictor interaction (i.e., the F statistic is the ratio between these two MS terms). Significant results therefore reflect effects which are consistent across participants.

Sequence probabilities. The simple effect of each of the phonotactic and orthotactic measures on its own was significant, Fs(1,23) > 27, ps < .001. Together, phonotactic and orthotactic probabilities accounted for about 18% of the variance between items (see Table 1). Because the probability metrics are correlated with each other (see Table 2), combinations account for less than the sum of the parts. The combination of phonotactic and orthotactic probabilities (R^2 = 18%) was only a little better than orthotactics alone (R^2 = 16%), and the combination of orthotactic bigrams and trigrams was little better than trigrams alone (R^2 = 15%). Among phonotactic probabilities, all three measures together (R^2 = 10%) were only slightly better than phonotactic trigrams on their own (R^2 = 8%).

Lexical neighborhoods. Probes varied in number of neighbors from 14 to 408 ($M = 67$, $Mdn = 50$, $SD = 51$). To take all of these neigh-

TABLE 1

Variance in Item Ratings Accounted for by Phonotactic and Orthotactic Probabilities in Experiments 1 and 2

	Experiment 1			Experiment 2		
Measure	Phonotactics	Orthotactics	Both	Phonotactics	Orthotactics	Both
Bigram	5	8		15	2	
Trigram	8	15		4	2	
Syl Part	4			5		
Combined	10	16	18	16	3	17

Note. R^2 statistics (× 100) based on item ratings averaged across subjects.

bors into account in the GNM as well, we applied it to the 408 nearest monosyllabic neighbors for each probe, using the edit distance measure described above as an estimate of psychological distance between probes and their neighbors.[6] Referring to 408 neighbors ensured, as a minimum, that all words differing by no more than two phonemes from any probe were taken into account. These 105,672 neighbors (259 probes times 408 neighbors) and their token frequencies were used in the GNM as predictors of wordlikeness.[7]

By itself, NNB (R^2 = 8.4%) accounted for much less item variance than did the GNM (R^2 = 24%). Neighborhood effects increased the total item variance explained from 18% (sequence probabilities alone) to 23% (with NNB) and 29% (with the GNM). The unique contribution of each neighborhood model was significant: R^2 = 5.4 and 12% for NNB and GNM, respectively, $F(1,23) = 27$ and $F(4,92) = 25$, ps < .001. Clearly, wordlikeness ratings were influenced by lexical neighborhoods and were not determined just by disembodied phonotactic or orthotactic probabilities.

In order to test whether sequence probabilities had any effect beyond what could be attributed to neighborhoods, we tested for unique effects of sequence probabilities. After partialling out neighborhood effects, the total remaining effect of sequence probabilities was significant, R^2 = 15 (partialling out NNB) and 6% (partialling out the GNM), Fs(5,115) = 26 and 12, ps < .001. These significant effects indicate that sequence probabilities were not subsumed by neighborhood effects. Finally, phonotactic and orthotactic probabilities both had significant unique effects after partialling out the other in addition to NNB, R^2 = 3.6 and 4.7%, respectively, $F(3,69) = 20$ and $F(2,46) = 15$, ps < .001.

To investigate the role of token frequency in neighborhood effects, we removed the frequency component from the GNM and compared the performance of this reduced model (combined with sequence probabilities) with the full GNM. Removing the frequency component significantly reduced the fraction of variance explained from 29 to 25%, $F(2,46) = 12$, $p < .001$. This indicates that wordlikeness depends on the frequency of neighboring words as well as the degree of similarity to them. The relative weight assigned to each neighbor as a function of its token frequency, as determined by regression parameters in the full GNM, is graphed in Fig. 1. This figure also shows the frequency effects found in Experiment 2 using the same edit distance metric and also using an alternative distance metric discussed later.

[6] It is conceivable that multisyllabic neighbors might also contribute to neighborhood effects for monosyllabic stimuli. For reasons of computational complexity (finding all items' n nearest neighbors requires a step through the entire corpus for each stimulus item), we follow previous work in restricting our consideration to monosyllabic neighbors (cf. Luce, 1986).

[7] Regression models were fit to mean item ratings to determine sums of squares for the predictor variables. The same models were fit to the ratings for each subject to determine corresponding sums of squares for subject by predictor interactions. The complete set of tests for Experiment 1 involved 75 regressions with the GNM.

TABLE 2

Bivariate Correlations (× 100) among Predictors of Wordlikeness

	Metric		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	R^2 Exp1	R^2 Exp2
												Phonotactics													
1.	1P	Mean		98	19	52	48				28				23		17		17	8	14	48	47	5	6
2.		Lg	98		15	50		45				27				33		39	21	5	13	49	46	7	8
3.		*	19	15		32			14				13				7		–28	44	28	2	19	0	0
4.		Lg*	52	50	32					55				25				52	–74	72	55	47	72	11	8
5.	2P	Mean	48					95	45	77	35				35		25		22	9	26	53	11	5	15
6.		Lg		45			95		35	77		29				34		46	26	4	23	51	9	5	15
7.		*			14		45	35		49			39				32		–29	49	47	38	24	6	10
8.		Lg*				55	77	77	49					25				62	–40	54	56	57	38	10	19
9.	3P	Mean	28				35					95	44	87	42		26		25	0	22	14	8	8	4
10.		Lg		27				29			95		35	91		45		38	27	–5	18	12	3	10	4
11.		*			13				39		44	35		48			63		–34	36	35	17	30	4	4
12.		Lg*				25				25	87	91	48					50	–14	27	40	18	24	14	7
13.	SP	Mean	23				35				42					77	87		–30	44	43	29	43	4	5
14.		Lg		33				34				45			77			86	–33	53	54	39	50	9	7
15.	ORP	*	17		7		25		32		26		63		87			54	–21	29	29	18	35	2	4
16.		Lg*		39		52		46		62		38		50		86	54		–28	52	56	44	46	9	13
17.	Length		17	21	–28	–74	22	26	–29	–40	25	27	–34	–14	–30	–33	–21	–28		–78	–53	–14	–47	3	1
												Neighborhoods													
18.	NNB		8	5	44	72	9	4	49	54	0	–5	36	27	44	53	29	52	–78		75	30	46	8	7
19.	GNM	ED	14	13	28	55	26	23	47	56	22	18	35	40	43	54	29	56	–53	75		37	30	18	22
20.		SPM	48	49	2	47	53	51	38	57	14	12	17	18	29	39	18	44	–14	30	37		45	21	29
21.		PSL	47	46	19	72	11	9	24	38	8	3	30	24	43	50	35	46	–47	46	30	45		3	4

Note. Includes phonotactic probabilities for phonemes (1P), bigram transitions (2P), trigram transitions (3P), syllable part transitions (SP), and onset–rime combinations (ORP), which are computed as mean probabilities, logged means (Lg), product probabilities (*), or logged products (Lg*). *Length* counts the number of phonemes. Measures of lexical neighborhoods include number of neighbors (NNB), and the GNM fit to data from Experiment 2 using estimates of phonological distance based on edit distance (ED), syllable part mismatches (SPM), and phoneme subset lattices (PSL). The last two columns show the amount of item variance each metric accounts for in Experiments 1 and 2. Some correlations are omitted for clarity.

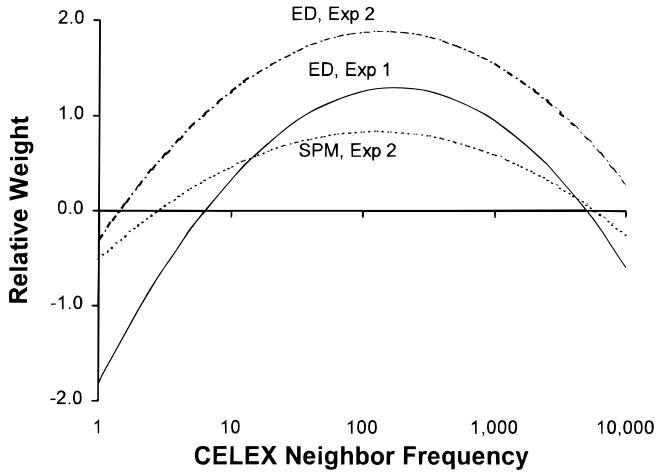

FIG. 1. Best fit GNM quadratic frequency weighting for Experiments 1 and 2, using estimates of phonological distance based on edit distance (ED) and syllable part mismatches (SPM). Graph shows the relative weight assigned to the similarity between a probe and each neighboring word as a function of the word's token frequency. A frequency of 100 represents about 5.6 occurrences per million words.

As is clear from the graph, the effect of frequency was nonmonotonic. A word of very low *or* very high frequency has little effect on wordlikeness, even if it is similar to a probe item. Words with token frequencies between about 10 and 1000 (equivalent to about 0.6 and 60 occurrences per million words, respectively) had the greatest influence on wordlikeness (among the 105,672 neighbors included in the analysis the average log token frequency was 1.97, corresponding to a CELEX frequency of 93.3, or about 6 occurrences per million words). This suggests that, in general, nonwords which are similar to medium-frequency words will tend toward higher wordlikeness.

Discussion

Experiment 1 examined three issues not addressed by previous research on sequence typicality: the importance of phonotactic influences on sequence typicality in nonwords, the importance of lexical influences, and the relationship between phonotactics and the lexicon. The central finding of Experiment 1 is that both sequence probabilities and lexical similarity produced significant unique effects. This result confirms the role of sequence probabilities in wordlikeness. This has been widely assumed, but previous studies had not controlled for lexical effects, so that lexical influences could not be ruled out as an alternative explanation for seemingly phonotactic effects. At the same time, this result establishes lexical neighborhoods, which had been given little attention by previous research on wordlikeness, as an important factor. Finally, the finding that sequence probabilities and lexical neighborhoods exhibit unique effects suggests that these represent cognitively distinct sources of knowledge. Lexical influences do not subsume phonotactic effects, which suggests that phonotactic knowledge is not merely implicit in our knowledge of words. We return to this issue in the General Discussion.

In measuring neighborhood effects, the GNM provided a substantial improvement over the simple number of neighbors density metric. The GNM also revealed that the effect of a neighboring word was modulated by its frequency as well as its degree of similarity to a probe. The frequency effects observed in our study cannot be due to unusual phonotactic properties of very high frequency words, because separate terms were included for phonotactic probabilities. To some extent the frequency effects we observed are consistent with effects reported elsewhere in lexical processing (e.g., Goldiamond & Hawkins,

1958; Luce & Pisoni, 1998; Luce et al., 1990; Whaley, 1978). What is striking about our results is that the frequency effect was nonmonotonic. Again, we will return to this finding in the General Discussion.

Finally, Experiment 1 assessed the overall ability of sequence probabilities and lexical effects to explain wordlikeness. Any complete theory of wordlikeness must explain the rating differences we observed among the items in our study. Our results confirmed our intuition that phonotactic probabilities were of limited explanatory power. Lexical neighborhoods were more influential in determining wordlikeness. Even so, sequence probability and lexical measures together accounted for some 30% of variance in item ratings, leaving 70% unaccounted for. This unaccounted variance is shared knowledge about sequence typicality which is unexplained. The extent of unaccounted variance suggests severe limitations to the current understanding of this aspect of linguistic knowledge, but how general is this rather unexpected result? The only previous studies which report measures of the magnitude of effect their predictors had on wordlikeness report much larger values, including R^2 values as high as 96 (Greenberg & Jenkins, 1964), and 76% (Frisch et al., 2000).[8] However, this is where the differences in design between previous research and our own study become crucial. Because the isolates of Experiment 1 were chosen at random, the R^2 values we obtain estimate the general relationship between those predictors and wordlikeness for monosyllabic nonwords in the neighborhood of arbitrary isolates. This is a specialized population of nonwords, to be sure, but the point is that our estimates of explained variance extend beyond the specific items actually tested in our study. In contrast, inferences cannot be drawn about populations of nonwords from the sample R^2 values reported in these previous studies due to the methods of stimulus selection they employed. In each case, these studies chose stimuli to yield either a dichotomous high–low phonotactic probability distribution (Frisch et al.)[9] or a flat distribution which covered a range of probability values (Greenberg & Jenkins). Though this method of stimulus selection will have maximized the power of these studies for significance tests, the gains in power are achieved by magnifying the population correlations under test, producing larger correlations in the stimulus sample. To illustrate the problem this poses for assessing the importance of various predictors, we selected a subset of our stimulus items covering the 25% most extreme phonotactic trigram values. In this dichotomous subset, the fraction of wordlikeness variance accounted for is $R^2 = 24$ and 20% for phonotactic and orthotactic trigrams, respectively. These values are substantially inflated relative to those obtained across the entire set of stimulus items ($R^2 = 8$ and 15%, as shown in Table 1). Perhaps even worse, the R^2 values are differentially inflated by the selection process: Phonotactic trigrams have greater predictive value in the subset than do orthotactic trigrams, but the opposite is true in the entire set of items and, by inference, in the population from which the original set was sampled. Because the R^2 values reported in previous studies are inflated by the stimulus selection process, comparisons are not possible across studies and, more important, the values reported overestimate to an unknown degree the ability of the predictors under test to account for wordlikeness in nonwords more generally.

Because there are no other studies available for comparison, a replication can help clarify whether our predictors are genuinely as limited in explanatory value as it seems. Such a replication is also desirable for the other key findings of Experiment 1.

EXPERIMENT 2

Experiment 2 is an auditory version of the same task. The difference in modality and concomitant differences in the duration of the study

[8] The correlation coefficients reported in Greenberg and Jenkins (1964) and in Frisch et al. (2000) have been squared here for comparison purposes.

[9] Collapsing across different syllable lengths, the stimuli of Frisch et al. (2000) cover a broad range of phonotactic probability with a fairly flat distribution, but the distribution of phonotactic probabilities within each length group is dichotomous.

represent a considerable change from Experiment 1, which allows us to examine the robustness of its key findings. A purely auditory presentation also allows us to clarify whether the orthotactic contribution arose solely as a consequence of written presentation or whether knowledge of orthotactic patterns plays a role more generally.

Method

Participants. Twelve University of Warwick undergraduates and graduates, 6 Cardiff University graduates, and 6 Oxford University graduates were paid for participating in this experiment.

Materials. The stimulus set was the same as in Experiment 1, minus misspelled items. The items were recorded in a professional radio studio by a male speaker unaware of the purpose of the experiment. There were 328 items, words and nonwords, in total. The items were recorded in four different orders (two forward and two reverse orders from Experiment 1).

Procedure. Participants took part in small groups of two and three. The experiment lasted about an hour. Participants listened to a tape recording in which all items were spoken in the same syntactic frame, for example:

> One. 'Slontch.' How typical sounding is 'slontch'?

Results

Several items on some of the hour-long lists were mispronounced by the speaker; the ratings of those items were dropped from the analysis. A total of 12 data points were affected in this way, but no single item was lost entirely. As in Experiment 1, the ratings were subjected to an arcsine transformation.

Across individual responses, the differences between items were large relative to noise in the data, $\omega^2 = .21$. The similarity between Experiment 2 and Experiment 1 ($\omega^2 = .22$) indicates that the level of agreement between participants in relative item ratings was the same for both tasks.

On average, the transformed ratings for near misses ($M = .41$, $SE = .006$, $n = 237$) were somewhat higher than those for isolates ($M = .33$, $SE = .013$, $n = 22$). These aggregate ratings are virtually identical to those obtained in Experiment 1. The correlation between mean item ratings in Experiments 1 and 2 was $r = .60$.

Sequence probabilities. The simple effect of each of the phonotactic and orthotactic measures on its own was significant, Fs(1,23) > 16, ps < .001. Together, phonotactic and orthotactic probabilities accounted for about 17% of the variance between items (see Table 1). The combination of phonotactic and orthotactic probabilities was only a little better than phonotactics alone ($R^2 = 16\%$), and the combination of phonotactic bigrams, trigrams, and syllable parts was little better than bigrams alone ($R^2 = 15\%$). Orthotactic bigrams and trigrams accounted for very little variance in this oral task ($R^2 = 3\%$).

Lexical neighborhoods. By itself, NNB ($R^2 = 7.2\%$) accounted for much less item variance than did the GNM ($R^2 = 22\%$). Neighborhood effects increased the total item variance explained from 17% (sequence probabilities alone) to 22% (with NNB) and 31% (with the GNM). The unique contribution of each neighborhood model was significant: $R^2 = 5.4$ and 14% for NNB and GNM, respectively, $F(1,23) = 22$ and $F(4,92) = 28$, ps < .001.

After partialling out neighborhood effects, the total remaining effect of phonotactic and orthotactic probabilities was significant, $R^2 = 15$ (partialling out NNB) and 9% (partialling out the GNM), Fs(5,115) = 31 and 25, ps < .001. Phonotactic probabilities had a significant unique effect after partialling out orthotactics in addition to NNB, $R^2 = 14\%$, $F(3,69) = 47$, $p < .001$. In contrast, orthotactics had no significant unique effect after partialling out phonotactics and the NNB, $R^2 = 0.4\%$, $F(2,46) = 1.6$, $p = .22$.

Removing the frequency component from the GNM significantly reduced the fraction of variance explained from 31 to 30%, $F(2,46) = 8.2$, $p < .001$. This replicates the finding of Experiment 1 that wordlikeness depends on the frequency of neighboring words as well as the degree of similarity to them. Again, frequency effects were nonmonotonic, with influence rising with token frequency up to a maximum for medium-frequency neighbors and then falling off again at higher token frequencies (see Fig. 1).

Discussion

While some differences to Experiment 1 emerged, the key findings were replicated. Again, neighborhood density and similarity to individual words had significant effects, over and above the effects of sequence probabilities. Also, the frequency of neighboring words had a significant, nonmonotonic effect on wordlikeness. Lexical effects were stronger than effects due to sequence probabilities, but sequence probabilities were not subsumed by lexical ones.

In Experiment 2, phonotactic bigram probabilities were much better than trigram or syllable part probabilities at predicting wordlikeness ratings. This contrasts with Experiment 1, in which trigram probabilities (orthotactic as well as phonotactic) were better than bigram probabilities at predicting wordlikeness. It may be that the transitory nature of the auditory stimuli in Experiment 2 made it relatively difficult for participants to analyze stimuli into larger trigram chunks. Also, the oral mode of presentation in Experiment 2 evidently eliminated the effects of orthotactic probabilities. Nevertheless, the orthotactic influence in Experiment 1 seems to have had little effect on the relationship between phonotactic probabilities and lexical neighborhoods. Indeed, it is remarkable that all key findings replicated across two studies conducted in different modalities, with significant concurrent differences in the duration of the experiment (20 min vs. slightly over 60 min). This is of methodological consequence, since administering a questionnaire is far easier for both experimenters and participants where so much material is involved. Our results suggest that, in this context at least, written presentation may be effective if it is ensured that stimulus items receive unambiguous spellings as did our materials and if orthotactic effects are factored out in the analysis.

As in Experiment 1, a large fraction of item variance in Experiment 2 remains unaccounted for. This suggests that our current understanding of wordlikeness is still rather limited, even when lexical influences are considered in addition to phonotactics. Such a limitation would have readily been masked in previous research by the emphasis on significance testing. It would also not become apparent in other tasks, particularly online processing tasks, in which factors beyond sequence typicality would contribute significantly to the systematic variability between items. Because there are no obvious alternatives to lexical knowledge and phonotactic knowledge as the determinants of our wordlikeness ratings, this result suggests that it is lexical neighborhoods and/or phonotactics themselves which are only partially understood. The implications of this finding extend beyond sequence typicality and affect any domain in psycholinguistics that has drawn on phonotactic probabilities or lexical neighborhoods. Consequently, we followed these results with an extensive search for better measures of phonotactics and lexical neighborhoods.

ALTERNATIVE MODELS OF PHONOTACTICS AND LEXICAL NEIGHBORS

Clearly no work could entirely exhaust the range of possible measures of phonotactics or lexical influence. We thus take a two-step approach. We first try to delimit conceptually the key dimensions of variation governing alternate measures and then sample different measures varying along each of these dimensions. Finally, we combine the most successful measures and evaluate them against the data from Experiment 2 to determine whether they account for the missing variance.

Measuring Phonotactic Probabilities

Where might better phonotactic measures be found? Conceptually, the space of possibilities is delimited by the following dimensions of variation: (1) one might consider different units of analysis; (2) one might calculate the statistics over a different, more appropriate or psychologically relevant corpus; (3) one might use different psychological functions (e.g., logs) to relate phonotactic probabilities to behavior; or (4) the individual component probabilities might be combined in a different way.

Bigrams and trigrams. We first explored two variations in the corpus over which these statistics are compiled. Since our experimental materials were monosyllabic, participants might eval-

uate them with respect to transition probabilities of English monosyllables, rather than those of the entire lexicon. However, bigram and trigram probabilities computed over the monosyllabic entries in CELEX accounted for much less of the variance between items in Experiment 2 (R^2 = 5.4 and 0.9% for bigrams and trigrams, respectively) than did probabilities computed over the entire lexicon (R^2 = 15 and 4.0%).

The second corpus variation involved word frequency. The simplest way to compute bi- and trigram probabilities is to base these on a large dictionary of English. This method differs from an analysis of real speech in that the dictionary lists the sounds of each word just once, so that the bigrams of common words are underrepresented compared to their occurrence in real speech. The statistics of real speech are better approximated by weighting the bigrams of each word in the dictionary by its frequency of use (cf. Jusczyk et al., 1994; Vitevitch & Luce, 1998, 1999; Vitevitch et al., 1997). However, the probabilities for our stimulus set computed with and without weighting by words' log token frequencies (as listed in CELEX) were almost perfectly correlated: For all five of our phonotactic and orthotactic measures, whether computed across the entire CELEX database or across monosyllables alone, the lowest correlation for corresponding measures with and without frequency weighting was r = .97 (inspection revealed no clear nonlinear relationships between weighted and unweighted measures). This high degree of correspondence suggests that little is to be gained by taking word frequency into account in these sequence probabilities.

Along the other dimensions of variation, we examined other ways of combining the component bigram (or trigram) probabilities into a composite measure for the entire word and we also considered a different psychological function. Our original measures above averaged over the component probabilities for each item, whereas some studies have used unaveraged product probabilities and have also taken logs of probabilities rather than using them directly (e.g., Coleman & Pierrehumbert, 1997). These computational differences produce some measures of item probability which are only moderately correlated with each other across our stimulus set, as shown in Table 2. The combination which accounts for the most variance is log product probabilities (R^2 = 19 and 7% for bigrams and trigrams, resp., in Experiment 2). The superiority of log product probabilities over our original unlogged, averaged measures is probably due to word length effects (see Coleman & Pierrehumbert, 1997). Whereas averaging component probabilities normalizes for word length, the use of product probabilities penalizes longer words and assigns them lower probabilities. This allows product probabilities to capture the fact that wordlikeness ratings generally decrease with word length (r = −.12 in Experiment 2). However, word length effects also arise naturally in models of lexical neighborhoods (longer words have fewer neighbors), so the gain obtained through the use of products in phonotactic measures might ultimately be superfluous.

Single phoneme probabilities. In principle, single phoneme probabilities are independent of bigram and trigram transition probabilities (e.g., a high probability bigram transition could involve low frequency phonemes which occur in just a few contexts). Across our stimulus set, bivariate correlations between single phoneme probabilities and bigram probabilities ranged from r = .14 to .55, depending on how the probabilities were computed (see Table 2). These correlations leave ample room for independent effects on wordlikeness to emerge. Accordingly, we tested the ability of logged and unlogged mean and product probabilities to account for the experimental wordlikeness ratings. The best phoneme probability metric used log product probabilities (R^2 = 8% for Experiment 2).

Subsyllabic constituents. In addition to sequence probabilities based on phoneme transitions, the analyses of Experiments 1 and 2 considered a syllable part metric based on the probability of transitions between a syllable's onset, nucleus, and coda. Again, computing these probabilities relative to monosyllables (R^2 = 2.7% in Experiment 2) instead of the whole lexicon (R^2 = 5.2%) substantially reduces the amount of item variance explained by syllable parts. By contrast, applying a log transformation to probabilities computed across the whole lexicon increases the amount of variance explained (R^2 = 7.5%).

We also examined an alternative measure of subsyllabic probabilities. The onset–rime probability metric (ORP; Coleman & Pierrehumbert, 1997) takes the syllable onset and rime (the vowel and any following consonants) as the basic units of analysis. It emphasizes the importance of nucleus–coda co-occurrence information and assumes that onset and rime are statistically independent (cf. Treiman, 1988, 1995). ORP distinguishes between initial, medial, and final syllables and also between stressed and stressless syllables. We assumed our (monosyllabic) stimuli were stressed and evaluated them on the basis of probabilities for word-initial onsets and word-final rimes in stressed syllables. Log ORP values ($R^2 = 13\%$ for Experiment 2) were better than log syllable part probabilities ($R^2 = 7\%$). However, the relatively simple log product bigram probabilities ($R^2 = 19\%$) were better than ORP.

Summary of phonotactic alternatives. We examined four key dimensions of the space of possible phonotactic measures, testing additional basic units of analysis (single phonemes, onsets-and-rimes), different corpora (monosyllables, token frequency), combination rules (products), and psychological functions (logs). We examined 18 additional measures, but none was radically better than our original measures. Log product probabilities provided a modest but consistent gain. In addition, both single phoneme probabilities and onset–rime probabilities look like promising additions to the set of phonotactic measures. Although additional combinations of unit of analysis, corpus, combination rule, and psychological function remain untested, the fact that none of the many variations we tried led to striking improvements makes it seem unlikely that any vastly superior combination actually exists. This points toward lexical neighborhoods as the route toward a better explanation of sequence typicality.

Lexical Neighborhoods

The same basic dimensions of variation govern the space of possible models of lexical neighborhood. However, the complexities involved mean that the space of alternative models is vast and greatly exceeds that sketched for phonotactics. Furthermore, there has been virtually no exploration of these issues in previous research. The most problematic issues concern basic units of analysis and combination rules. These form the central question for any model of lexical neighborhood: What makes two words similar? Phonological theory decomposes words into feet, syllables, syllable parts, phonemes, and phonological features. Any of these might be the basic unit of comparison from which the similarity between two words is composed. What makes these candidates so complex to evaluate is that each candidate is associated with a range of possibilities as to how the comparisons are to be conducted. Assuming, for example, phonemes as the basic unit of comparison, which phoneme comparisons between the two words are relevant? Are *stick* and *ticks* more or less similar than *stick* and *tick*? That is, do matching phonemes in different positions (e.g., the /s/ in *stick* and *ticks*) contribute to the overall similarity between words, or are the psychologically relevant comparisons restricted to phonemes in corresponding positions? Furthermore, corresponding positions in words of different length might be defined in many different ways. A serious test of phonemes as the basic unit of analysis would have to explore all of these possibilities.

To some extent, one could bypass the issue of how to compute similarity by collecting similarity judgments or confusability data for pairs of words (e.g., Luce, 1986). However, it would be an enormous task to collect such data for, say, the 408 nearest neighbors of each of our stimulus items. Potentially, whole-word similarity might be derived from the confusability of subword parts (as in Luce), but the same questions arise: What are the appropriate subword parts, and how do they combine to determine whole-word similarity?

Even the conceptually straightforward dimensions of variation such as choice of corpus harbor considerable difficulties. It would be desirable to examine the same variation between measures based on the entire CELEX lexicon and measures based on monosyllables. However, there are significant computational costs involved, given that finding a word's *n* nearest neighbors requires a step through the entire cor-

pus for each stimulus item; this has restricted previous work (e.g., Luce, 1986) and ours to monosyllables only.

Given the range of possibilities and the sizeable effort involved in testing a model such as the GNM, nothing like an exhaustive search for better models of lexical neighborhood could be conducted here. We tested two variants of the GNM which differ from our original version in the basic unit of analysis for computing similarity to neighbors and also in the way these units are combined. These variants maintain the same corpus (monosyllables), psychological function (exponential), and way of aggregating whole-word similarities across all neighbors into a single value for each probe item.

Phoneme subset lattice distance metric. The first new variant of the GNM employs a phonological distance metric which gives credit for shared phonemes, even in different positions. The words *stick* and *ticks* are composed of the same phonemes arranged in different sequences. The edit distance metric used in the GNM above gives no credit for shared phonemes in different word positions, so *stick* and *ticks* are no more similar than *trick* and *ticks*. The phoneme subset lattice metric extends the natural class lattice metric used for phoneme dissimilarity in Frisch (1996). A comparison of two words proceeds by listing all subsets of phonemes in each word, whether the phonemes are adjacent or not, but preserving the order of the phonemes (e.g., /stu/ *stew* includes the subsets /stu/, /st/, /su/, /tu/, /s/, /t/, and /u/). The phoneme subsets are augmented by the subsets of phonological features for each phoneme, to obtain a combined list of phoneme and feature subsets for each word. A single metric of phonological distance between two words is computed by dividing the number of different subsets by the total number of subsets in the two words.

Syllable part mismatch distance metric. A second new variant of the GNM adopts syllable parts (onset, nucleus, and coda) as the basic subword unit of analysis in order to explore the possibility that syllable parts are weighted differentially in determining similarity. Sendlmeier (1987) suggests that onsets are more important than codas in word similarity; Nelson and Nelson (1970) make the opposite claim. Stemberger (1994) argues that, at least for speech production, vowels are of primary importance in phonological similarity, followed by word-initial (onset) consonants, and then word-final consonants. It seems reasonable to consider a distance metric which allows differential weighting for comparisons between different syllable parts in two words. The syllable part feature mismatch metric computes three values for each pair of syllables to be compared, one for corresponding onsets, one for nucleuses, and one for codas. For simplicity, this metric dispenses with the detailed phonological feature representations employed in our other metrics of phonological distance. For each syllable part, we compile a set of major class features (place of articulation, manner of articulation, and voicing) represented within the syllable part (e.g., the onset of *skit* includes the features [alveolar], [velar], [voiceless], [fricative], and [stop]). To the set of features we add the phoneme subsequences which are present in the syllable part (e.g., /s/, /k/, and /sk/ for the onset /sk/). This preserves positional information so that syllables such as /æps/ and /æsp/ are distinguished from each other. Finally, corresponding syllable parts are compared by dividing the number of features and subsequences they share by the total number of features and subsequences in both syllable parts and subtracting this ratio from 1. A weighted sum of syllable part scores yields a single syllable part feature mismatch metric for the pair of words. The relative weights assigned to the syllable parts are determined by regression.

Evaluation of GNM variants. For each of the two new metrics of phonological distance, we identified the nearest 408 real-word neighbors to each probe (i.e., 105,672 total neighbors for each metric). The neighbors and their token frequencies were used in the GNM to model the ratings from Experiment 2. The phoneme subset lattice metric (R^2 = 4%) explained much less variance between items than did the original edit distance metric (R^2 = 22%). The syllable part mismatch metric (R^2 = 30%) explained more variance than did the edit distance metric. Inspection of the weighting parameters for different syllable parts tentatively suggests that onsets

were more important than codas, though the difference was small. Vowels carried no weight whatsoever.[10]

In summary, our examination of the space of possible models of lexical neighborhood made clear both that there is much more to explore and that much less ground has been covered by previous research. Tests of two alternative models found that giving credit for shared phonemes in different syllable parts did not help, whereas differential weighting by syllable position seems a promising road to pursue.

Combined and Unique Effects of the Best Predictors

Our explorations of alternative measures identified a number of measures which individually provide somewhat better accounts of wordlikeness than the factors considered in the original analyses. To see how these measures perform in combination, we tested them together against the data from Experiment 2. This analysis included five measures of phonotactic probability, including log product probabilities for single phonemes, bigram transitions, and trigram transitions, plus log syllable part and onset–rime probabilities. The GNM was used to model neighborhood effects with the syllable part mismatch metric of phonological distance. Orthotactic probabilities, which had no significant effect in Experiment 2, were left out.

Without the GNM, the new set of phonotactic metrics accounted for $R^2 = 23\%$ of variance between items (compared to 16% for the original metrics). Neighborhood density (NNB) had no significant unique effect after the new phonotactic metrics were partialled out, $F < 1$.

When we combined the new phonotactic probabilities with the GNM, the full model accounted for $R^2 = 38\%$ of variance between items (compared to 31% reported above with the original set of phonotactic probabilities and the edit distance metric of phonological similarity). The unique contributions were $R^2 = 15$ and 9% for the GNM and phonotactics, respectively, $F(6,138) = 21$ and $F(5,115) = 17$, ps $< .001$. Removing the frequency component from the GNM significantly reduced the fraction of variance explained to $R^2 = 36\%$, $F(2,46) = 21$, $p < .001$. The frequency effects were nonmonotonic, as shown in Fig. 1.

Neither the consideration of alternative phonotactic measures nor of alternative measures of lexical influence substantially alters the basic findings reported above. Our main findings of independent lexical and phonotactic effects, superior lexical influence, and nonmonotonic frequency effects are confirmed with the new set of measures. Moreover, despite an extensive search for the best measures of phonotactic probability and neighborhood effects, the best set of predictors accounts for only 38% of wordlikeness variance in the population of nonwords we sampled. Although this is an improvement over the 31% explained by the original predictors, it still leaves a considerable proportion of the variance unexplained. To eliminate the possibility that the remaining unexplained variance is just noise, we tested it against the residual error variance and found it to be statistically significant, $R^2 = 62\%$, $F(247,5608) = 5.1$, $p < .001$. This result confirms the conclusion that an entirely adequate account of wordlikeness has not yet been found.

GENERAL DISCUSSION

The research reported here examined the nature of sequence typicality, an important factor in many areas of speech processing and verbal memory. We sought to determine: (1) the extent to which sequence typicality is based on phonotactic probabilities (as has often been assumed); (2) whether lexical influences have their own, possibly even more important role in determining sequence typicality; and (3) whether knowledge of sequence typicality is implicit in the lexicon and completely subsumed by neighborhood effects or whether it is based on an independent store of abstract statistical knowledge.

Sequence typicality has been linked to phonotactic probabilities in a number of previous stud-

[10] The surprising absence of any contribution by vowels to phonological (dis)similarity in the syllable part mismatch version of the GNM may be due in part to the limited nature of the feature comparisons employed—two vowels are either the same or else they differ in place of articulation. It is possible that effects of vowel comparisons will emerge from more sophisticated models that incorporate degrees of similarity among vowels.

ies (Coleman & Pierrehumbert, 1997; Dankovicova et al., 1998; Frisch et al., 2000; Gathercole & Martin, 1996; Vitevitch et al., 1997). However, the results of these studies are confounded, since phonotactic probabilities are generally correlated with neighborhood density (Landauer & Streeter, 1973; Frauenfelder & Schreuder, 1992). This confound raises the possibility that seemingly phonotactic effects in these studies were in fact lexical. Our finding that phonotactic probabilities had significant unique effects above and beyond the effects of lexical neighborhoods is the strongest evidence to date for the role of phonotactics in sequence typicality. This finding replicated across modalities of stimulus presentation and held across a variety of measures of both lexical neighborhoods and sequence probabilities. However, estimates of the variance explained by sequence probabilities alone confirmed our suspicion that phonotactic probabilities explain less about sequence typicality than is commonly assumed. A combination of the five best measures of phonotactic probabilities accounted for about 23% of the variance between items in wordlikeness. Furthermore, extensive examination of alternative phonotactic measures makes it seem unlikely that this figure can be much improved upon.

Previous evidence for lexical influence on sequence typicality has been limited (Greenberg & Jenkins, 1964; Martin & Gathercole, reported in Gathercole & Martin, 1996; Frisch et al., 2000), and the lexical measures used in these studies are confounded with phonotactic probabilities. Accordingly, our finding that lexical neighborhoods had significant unique effects above and beyond the effects of phonotactic probabilities is also the strongest evidence to date for neighborhood effects in sequence typicality. Our results confirm our earlier suspicion that lexical influences are not only important, but play a bigger role in sequence typicality than do phonotactic probabilities, at least in the kind of monosyllabic nonwords we examined. This conclusion contrasts with Frisch et al. (2000), who found that phonotactic probability was a marginally better predictor of wordlikeness than was neighborhood density for the multisyllabic nonwords they examined. Several factors may contribute to this difference. First, Frisch et al.'s method of stimulus selection potentially inflates the effect of phonotactic probabilities relative to neighborhood effects. This kind of relative inflation is exemplified above by the dichotomous subset of our stimuli chosen to obtain extreme phonotactic trigram probabilities. In this subset, the relative importance of phonotactic and orthotactic trigrams was reversed relative to the relation those factors had in the original sample and, by inference, in the wider population of nonwords from which they were chosen. Frisch et al.'s choice of high and low phonotactic probability items makes it impossible to say whether the relationship they observed between phonotactic probabilities and neighborhood effects is indicative of their relative effects in the wider population. Second, Frisch et al. used a neighborhood density metric (a variant of NNB) to model neighborhood effects. Our own results confirmed that NNB was a poorer predictor of wordlikeness than were the best measures of phonotactic probabilities. However, we found NNB itself to be a relatively poor measure of neighborhood effects. The GNM, in conjunction with either the edit distance or the syllable part mismatch metrics of phonological distance, provided a better explanation of wordlikeness than did any individual phonotactic probability metric or, indeed, any combination of phonotactic probability metrics. This result underscores the shortcomings of NNB as a measure of lexical influence and emphasizes the gradient nature of neighbor status as represented in the GNM. Third, Frisch et al. examined wordlikeness in multisyllabic nonwords, whereas our study focused on monosyllables. It is conceivable that the relative importance of sequence probabilities and neighborhood effects changes as a function of word length. However, given the other differences between these studies, this final possibility is merely speculative.

The GNM identified nonmonotonic frequency effects. While the influence of a lexical neighbor increased as a function of its frequency at the lower end of the frequency spectrum, its influence actually decreased again for very common words. Nonlinear frequency effects have been indicated by previous findings that log fre-

quency is a better predictor than raw frequency counts in a variety of linguistic tasks (e.g., Goldiamond & Hawkins, 1958). Nonmonotonic frequency effects, however, have previously been reported only in the context of inflectional morphology (Bybee, 1995; Hahn et al., 1998; Moder, 1992). Within a schema model of productive phonological patterns, Bybee suggests that very high frequency bestows greater lexical autonomy or distinctiveness, attenuating links between the affected words and other words with similar phonological patterns. Very high frequency words therefore contribute little to the productivity of phonological patterns. The GNM provides a slightly different explanation in terms of psychological distance. The nonmonotonic frequency effects suggest that phonological space is warped around very high frequency words so that the effective similarity between them and their neighbors is less than would be expected on the basis of phonological characteristics alone. As a result, a high frequency word interacts less with its neighbor than does a medium frequency word.

Our findings also address the issue of the cognitive status of phonotactic knowledge in its relation to the lexicon. The question of whether phonotactic knowledge is a genuine, separate body of linguistic knowledge, or is merely implicit in the lexicon, has been raised in a variety of psycholinguistic contexts (cf. Gathercole & Martin, 1996; McClelland & Elman, 1986; Vitevitch & Luce, 1998, 1999). Our finding of unique effects of both provides the most direct evidence so far for the two as distinct components.

In the context of sequence typicality, Vitevitch and Luce (1998, 1999) argued for distinct, dissociable sources of phonotactic and lexical knowledge. In these studies, the existence of two separate sources was inferred from divergent patterns obtained in the speeded processing of words and nonwords. However, these experiments did not manipulate the two factors independently: Stimuli were either high in both phonotactic probability and neighborhood density or they were low in both. These studies thus necessarily provide only indirect support for the claim that phonotactic and lexical influences are distinct. Vitevitch and Luce argued that the differences they observe between words and nonwords can be explained by assuming that the processing of real words is dominated by lexical influences whereas nonword processing is dominated by phonotactics. Moreover, to explain different patterns across different tasks, the relative influence of both sources must be modulated by specific task demands. The results of our direct comparison of lexical neighborhoods and phonotactic probabilities on the same task lend support to Vitevitch and Luce's arguments for two separable knowledge sources.

Beyond the realm of sequence typicality, the two prime areas of language processing that have seen debate about whether or not processing draws on probabilistic phonotactic knowledge above and beyond lexical knowledge are the recognition of ambiguous phonemes and error patterns found in speech production. The debate about phonotactic influences on phoneme recognition was initiated by McClellend and Elman's (1986) TRACE model of spoken word recognition which attributed seemingly phonotactic effects to top-down feedback from the lexical level. However, these effects might also arise from recurrent connections on the phonological level in purely bottom-up models of word recognition (see, e.g., Norris, 1994). In principle, recurrent connections which enable a network to retain information about preceding network states could pick up the kind of statistical contingencies embodied in our phonotactic transition probabilities. Recurrent network connections have also been incorporated into a model of speech production (Dell, Juliano, and Ghovindjee, 1993) to account for phonotactic patterns in speech errors by extracting salient statistical properties of sound sequences. Most models of speech production, however, have incorporated a more limited view of phonotactic knowledge in the form of *frames* which distinguish legal from illegal sequences (e.g., Dell, 1986; MacKay, 1987; Meyer, 1990; Stemberger, 1985; Wheeler & Touretzky, 1997).

In both contexts, our results are broadly consistent with those models that can incorporate phonotactic probability through some form of recurrence. Conversely, they are at odds with

models which lack such information, whether those models assume that all phonotactic knowledge is implicit in the lexicon or whether the phonotactic knowledge is limited to inviolable rules or constraints in the form of phonotactic frames (since our stimuli are well formed with respect to English syllable structure, categorical rules are powerless to explain differences in typicality among our items; also see Coleman & Pierrehumbert, 1997). However, two qualifications must accompany these general comments. On the one hand, models of specific tasks (e.g., phoneme identification) might legitimately exclude phonotactic information as irrelevant to that particular task—different tasks might well draw on different sources of information. Even so, our results have implications for what might be deemed a parsimonious explanation of other tasks. For example, McClelland and Elman (1986) argued that their TRACE model of speech perception can account for apparent effects of phoneme transition probabilities in a phoneme identification task, without any explicit representation of subword transition probabilities. If the existence of explicit subword transition probabilities is independently motivated by results such as ours, then the appeal of McClelland and Elman's account of phoneme identification is greatly reduced (also see Pitt & McQueen, 1998).

On the other hand, network models that include recurrent phonological layers are consistent with our results only in a general sense. Recurrence gives such models the power, in principle, to extract at least some phonotactic probabilities. However, a practical difficulty which remains to be overcome is to get networks with recurrent connections to operate on a lexicon of a realistic size (cf. Norris, 1994). This is critical because we found that phonotactic probabilities computed over monosyllables alone were not nearly as good at accounting for sequence typicality as those computed over the entire lexicon. Even if the practical difficulty can be overcome, it remains to be seen whether recurrent connections can extract the right kind of phonotactic probabilities needed to explain the relevant phonotactic effects.

Our final result was unexpected. Even the best combination of phonotactic probabilities selected from the many measures we examined, combined with the best measure of neighborhood effects, accounted for only about half of the variance between items. Sequence typicality still seems to be only partially understood, which in turn limits our understanding of tasks in which sequence typicality has been acknowledged as important (e.g., naming, recall from verbal short-term memory, phoneme identification). Since there are no obvious alternatives to lexical knowledge and phonotactic knowledge for explaining wordlikeness, we conclude that it is lexical neighborhoods and/or phonotactics themselves which are only partially understood. In particular, the vast space of possible models of lexical neighborhoods is still virtually unexplored, and it would seem to be here that the greatest room for improvement lies. Given the importance of lexical influences throughout psycholinguistics (including speech production, word recognition, inflectional morphology, and phonological development), the search for more comprehensive models of lexical neighborhood must be a key area for future research. Whether more comprehensive models might even one day subsume the current contribution of phonotactics to sequence typicality remains to be seen.

APPENDIX

Isolates

drolf	greltch	prunth	shrupt	smisp	stolf	threlth	zinth
drusp	gwesht	shandge	slesk	snulp	swesk	throndge	
glemp	krenth	shresp	slontch	spulsh	swust	trinth	

Near misses

binth	drilf	glump	misp	shadge	slintch	spulp	thrilth
blemp	drisp	greld	nulp	shan	slisk	spulse	thrindge
blesk	droff	grell	pinth	shendge	slisp	spust	throng
breltch	drosp	grelm	plemp	shesp	slon	stelf	thronn
brelth	drump	grentch	plunth	shindge	slotch	stilf	thrupt
brenth	drup	gresht	presp	shinth	slulp	stoff	thrusp
bresp	drupt	gresp	printh	shondge	sluntch	stulf	tolf
brondge	drusk	gretch	prolf	shrapt	slusk	stulp	treltch
brunth	druss	griltch	prundge	shremp	slust	stust	trenth
clemp	drust	grinth	prunt	shrep	smimp	sulp	tresp
clenth	dusp	grolf	pruntch	shrept	smip	sulsh	trilth
clontch	dwesht	grondge	prupt	shress	smiss	sweck	trin
cren	dwesk	grunth	punth	shrest	smist	swelk	trindge
crend	finth	grupt	quenth	shript	smust	swesht	trintch
crendge	flemp	grusp	reltch	shrisp	snalp	swess	trith
crent	flesk	gweft	relth	shruct	snisp	swest	trolf
crentch	flontch	gwelt	renth	shruft	snulf	swisk	trupt
cresp	freltch	gwept	resp	shrunt	snulk	swisp	trusp
creth	frenth	gwesh	rinth	shrup	snull	swist	twinth
crinth	frinth	gwet	rolf	shrust	snult	swontch	wesht
crondge	frondge	hinth	rondge	shrut	snump	swuft	wesk
crunth	frunth	inth	rupt	shundge	snup	swunt	winth
crupt	frupt	jinth	rusp	sinth	snust	swupt	wust
crusp	frusp	kenth	sandge	sisp	solf	swutt	yinth
dinth	geltch	kinth	scolf	skisp	sontch	threll	zilth
dolf	gemp	kwesk	scontch	sleck	spelsh	threlm	zin
drelf	gesht	lemp	sculp	slemp	spesk	threlsh	zindge
dreltch	glem	lesk	sculsh	slentch	spuldge	threlt	zint
drenth	glep	linth	scust	sless	spulk	threth	zintch
dresp	glimp	minth	sesk	slest			

REFERENCES

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

Bailey, T. M., & Hahn, U. (1998). Determinants of wordlikeness. *Proceedings of the Cognitive Science Society*, **20,** 90–95.

Bailey, T. M., & Pothos, E. M. (1998). Unconfounding rules and similarity in artificial grammar learning. *Proceedings of the Cognitive Science Society*, **20,** 96–101.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, **10,** 425–455.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, **33,** 111–153.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language*, **17,** 205–215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, **22,** 727–735.

Coleman, J. S., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *Computational Phonology*, **III,** 49–56.

Dankovicova, J., West, P., Coleman, J., & Slater, A. (1998). *Phonotactic grammaticality is gradient*. Poster presented at the 6th International Conference on Laboratory Phonology, University of York, 2–4 July 1998.

Dell, G. S. (1986). A spreading activation theory of retrieval in language production. *Psychological Review*, **93,** 283–321.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, **17,** 149–195.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. Booij, & J. van Marle (Eds.), *Yearbook of Morphology 1991*. The Netherlands: Kluwer.

Frisch, S. (1996). *Similarity and frequency in phonology*. Ph.D. thesis, Northwestern University, Evanston, Illinois.

Frisch, S., Broe, M., & Pierrehumbert, J. (1997). Similarity and phonotactics in Arabic. *Rutgers Optimality Archive* [Online], ROA-223-1097. Available at http://www.web-slingerz.com/cgi-bin/oa_list.cgi.

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

Gathercole, S. E., Hitch, G. J., Service, E. S., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, **33,** 966–979.

Gathercole, S. E., & Martin, A. J. (1996). Interactive processes in phonological memory. In M.A. Conway (Ed.), *Cognitive models of memory*. Hove, UK: Psychology Press/MIT Press.

Gaygen, D. E. (1997). *The effects of probabilistic phonotactics on the segmentation of continuous speech*. Doctoral dissertation, University at Buffalo, Buffalo, NY.

Goldiamond, I., & Hawkins, W. F. (1958). Vexierversuch: The logarithmic relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology*, **56,** 457–463.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, **20,** 157–177.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single or dual route? *Cognitive Psychology*, **41,** 313–360.

Hahn, U., Nakisa, R. C., Bailey, T. M., Holmes, M., Kemp, D., & Palmer, L. (1998). Experimental evidence against the dual route account of inflectional morphology, *Proceedings of the Cognitive Science Society*, **20,** 472–477.

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

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24,** 695–711.

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 and Verbal Behavior*, **12,** 119–131.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary. *Cognition*, **50,** 239–269.

Lorch, R. F., Jr, & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16,** 149–157.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Doctoral dissertation, Indian University, Bloomington, IN.

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

Luce, P. A., Pisoni, D. B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. T. M Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 105–121), Cambridge, MA: MIT Press.

MacKay, D. G. (1987). *The organization of perception and action: A theory of language and other cognitive skills*. New York: Springer-Verlag.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10,** 29–63.

McClelland, J. L., & Elman, J. L. (1986). The Trace model of speech perception. *Cognitive Psychology*, **18,** 1–86.

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

Meyer, A. S. (1990). The phonological encoding of successive syllables. *Journal of Memory and Language*, **30,** 69–89.

Moder, C. L. (1992). *Productivity and categorization in morphological classes*. Ph.D. thesis, State University of New York at Buffalo, NY.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, **66,** 911–936.

Morton, J. (1979). Word recognition. In J. Morton, & J. C. Marshall (Eds.), *Psycholinguistics: Vol. 2. Structures and processes* (pp. 107–156), London: Paul Elek.

Nakisa, R. C., & Hahn, U. (1996). Where defaults don't help: the case of the German plural system. *Proceedings of the Cognitive Science Society*, **18,** 177–182.

Nelson, D. L., & Nelson, L. D. (1970). Rated acoustic (articulatory) similarity for word pairs varying in number and ordinal position of common letters. *Psychonomic Science*, **19,** 81–82.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1996). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, **23,** 873–889.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, **52,** 189–234.

Nosofsky, R. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115,** 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of the relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14,** 700–708.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, **34,** 393–418.

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

Pothos, E. M., & Bailey, T. M. (1997). Rules vs. similarity in artificial grammar learning. In *Proceedings of SimCat97, an Interdisciplinary Workshop on Similarity and Categorization*. Dept. of Artificial Intelligence, University of Edinburgh.

Pothos, E. M., & Bailey, T. M. (2000). The role of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26,** 847–886.

Sendlmeier, W. F. (1987). Auditive judgments of word similarity. *Zeitschrift fur Phonetik Sprachwissenshaft und Kommunikationsforschung*, **40,** 538–546.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science, *Science*, **237,** 1317–1323.

Stemberger, J. P. (1985). An interactive activation model of language production. In A. Ellis (Ed.), *Progress in the psychology of language* (Vol. 1). London: Erlbaum.

Stemberger, J. P. (1994). Rule-less morphology at the phonology-lexicon interface. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 147–169), Amsterdam: Benjamins.

Treiman, R. (1988). Distributional constraints and syllable structure in English. *Journal of Phonetics*, **16,** 221–229.

Treiman, R. (1995). Errors in short-term memory for speech: A developmental study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21,** 1197–1208.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, **40,** 211–228.

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

Vitevitch, M. S. & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, **40,** 374–408.

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

Whaley, C. P. (1978). Word-non-word classification time. *Journal of Verbal Learning and Verbal Behavior*, **17,** 143–154.

Wheeler, D. W., & Touretzkey, D. S. (1997). A parallel licensing model of normal slips and phonemic paraphasias. *Brain and Language*, **59,** 147–201.

(Received February 5, 2000)

(Revision received July 25, 2000; published online March 15, 2001)