Phonotactic and Morphological Effects in the Acceptability of Pseudowords

Jeremy M. Needle, Northwestern University
Janet B. Pierrehumbert, University of Oxford
Jennifer B. Hay, University of Canterbury

Author Note

Jeremy M. Needle, Department of Linguistics, Northwestern University.

Janet B. Pierrehumbert, Department of Engineering Science, University of Oxford;
Department of Linguistics, Northwestern University;
New Zealand Institute of Language Brain and Behaviour, University of Canterbury.

Jennifer B. Hay, Department of Linguistics, University of Canterbury;
New Zealand Institute of Language Brain and Behaviour, University of Canterbury.

This paper was presented at the 2017 Workshop on Morphological Typology and Linguistic Cognition, held at the Linguistic Institute in Lexington, KY. It has been accepted for the volume *Morphological Typology and Linguistic Cognition*, edited by Andrea Sims & Adam Ussishkin.

This project was made possible through a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Correspondence concerning this article should be addressed to Jeremy M. Needle, Department of Linguistics, Northwestern University, Evanston, IL 60208.

Contact: jneedle@u.northwestern.edu

Abstract

We develop a large set of pseudowords that systematically varies length and phonotactic probability, and obtain acceptability ratings using an online interface. We find that phonotactic likelihood and the presence of an apparent morphological parse both significantly predict acceptability; pseudowords containing known morphemes are more acceptable than otherwise comparable pseudowords that do not. We find support for the conjecture that novel words with apparent morphology are advantaged as additions to the lexicon. The resulting lexicon, as observed, is one in which long words are not a random sampling of phonotactically acceptable wordforms, but instead tend to be completely or partially decomposable into morphemes.

Keywords: morphological decomposition, phonotactics, pseudowords, wordlikeness

Phonotactic and Morphological Effects in the Acceptability of Pseudowords

1. Introduction

A central goal of phonology is to characterize the possible words of individual languages. In any language, the lexicon contains only a fraction of the phonologically possible wordforms. All other forms that are possible (but have no meaning) are *pseudowords*. The term *nonwords* is reserved here for strings that are phonologically impossible. Wordlikeness judgments reveal that the distinction between pseudowords and nonwords is a gradient one. Some pseudowords are judged to be extremely typical for the target language; should a conventional meaning become associated with them, they would be strong contenders to be added to the vocabulary. Others are moderately or barely acceptable. The statistical prediction of the full range of such gradient wordlikeness judgments is a major research issue, which this paper addresses. In this study, we explore the roles of phonotactics and of (partial or complete) morphological parsing in determining the acceptability of novel possible words in English.

The gradient acceptability of wordforms for different individuals depends on their own linguistic experiences. The experience-based perspective is well-established for phonology and morphology, especially as a way to understand why uniform rule-based models can fail to predict match real language patterns. To briefly give examples, Bybee (1988) posits that morphological patterns are described by schemas which vary in strength depending on factors such as word frequency; Pierrehumbert (2003) describes morphophonology as a probabilistic system learned by generalizing over word types in the lexicon. Daland, Sims, & Pierrehumbert (2007) develop a multi-agent diachronic model of gaps in Russian verbal paradigms, in which the production model for each generation of speakers samples from learned distributions of forms. A review of the evidence for experienced-based models of morphology may be found in Racz, Pierrehumbert, Hay, & Papp (2015). The effects of experience naturally differ across languages: Havas, Waris, Vaquero, Rodríguez-Fornells, & Laine (2015) use an artificial-language-learning experiment involving a gender-marking affix to demonstrate that L1 Finnish speakers (whose language "calls for continuous morphological decomposition" though it lacks gender) perform better than L1 speakers of Spanish (a more fusional language that does mark gender). These examples provide a variety of models in which the lexicon is shaped by experience with language. Accordingly, the effects considered here derive from linguistic experience: e.g., probabilistic phonotactics, lexical neighborhood density, or individual vocabulary size.

We build on substantial previous research regarding the gradient effects of phonotactics and lexical neighborhoods in wordlikeness judgments; the effects of phonotactics and lexical neighborhoods in processing; and the effects of morphology in processing. While there has been significant study of these effects in different experiment tasks, the relationships among these factors remain unclear. For example, phonotactics and lexical neighborhood density have both been shown to correlate with higher acceptability judgments for pseudowords (Bailey & Hahn 2001). These factors have contrasting effects in production: Edwards, Beckman, & Munson (2004) found that children produced nonwords faster and more accurately when the nonwords contained frequent (i.e., better) phonotactics, and Kapatsinski & Johnston (2010) found that pseudowords with better phonotactics are preferentially selected in a picture naming task. In contrast, Luce & Pisoni (1998) showed that items with dense neighborhoods were slower in naming. Phonotactics and lexical neighborhoods also have contrasting effects in perception; see review in Vitevich, Luce, Pisoni, & Auer (1999).

Results are somewhat sparse with regard to morphological effects, and have been mainly documented in perception (rather than production or acceptability judgments), leaving a gap in the current understanding. Caramazza, Laudanna, & Romani (1988) found evidence that the presence of real morphemes in nonwords made lexical decision responses slower and less accurate; that is, the presence of one or two real morphemes made it harder to correctly decide that a nonword was not a real word. This may imply that the presence of real morphemes makes nonwords more acceptable, though we note the difficulty of generalizing observations between perception, production, and acceptability tasks. A number of visual priming lexical decision studies support the idea that shallow morphological decomposition affects lexical perception: Beyersmann, Casalis, Ziegler, & Grainger (2015) found evidence from a visual priming lexical decision task that lexical decision was facilitated for stem targets when complex primes containing those stems were shown; priming obtained when the complex primes were fully or only partially decomposable into real morphemes. A brief summary of the principal findings from the literature is given in Table 1, and a more thorough review is found in Section 2.

While the lexical decision studies mentioned above describe online effects in perception, the present study considers effects on wordlikeness ratings of pseudowords. Wordlikeness ratings are a slower offline measure and may be more equally dependent on systems of perception and production, including morpho-semantic pressures. The current study uses an internet-mediated task to collect human wordlikeness judgments of pseudowords (described in Section 4). The large stimulus set evenly samples the phonotactic space, from highly improbable to highly probable pseudowords, and includes a range of word lengths from 4 to 7 phonemes (described in Section 3). This design allows us to find patterns in pseudoword acceptability across the possible space. First, we replicate and extend the positive gradient effect of phonotactic probability on wordlikeness judgments, across the full ranges of phonotactic probability and item length (see Section 5). We also replicate the positive effect of lexical neighborhood density on wordlikeness, when applicable (see Section 5). Examination of the residuals for the replication analyses suggested a fruitful post hoc analysis of shallow morphological complexity (*pseudomorphology*). We investigate the role of pseudomorphology in the wordlikeness judgments using an automatic analysis of the highly varied pseudo-compounding and pseudo-suffixation that are exhibited in our phonotactically-balanced set (see Section 6). We show that morpho-orthographic decomposition provides a positive effect on wordlikeness judgments for pseudowords.

	Acceptability	**Perception**	**Production**
Good Phonotactics	+ Positive correlation (Bailey & Hahn 2001)	+ Facilitatory (Hay, Pierrehumbert, & Beckman 2004)	+ + Faster and more accurate (Edwards, Beckman, & Munson 2004)
Dense Neighborhood	+ Positive correlation (Bailey & Hahn 2001)	– Inhibitory (Vitevitch & Luce 2016)	– + Slower, but more accurate (Luce & Pisoni 1998)
Morphology	(+) Focus of this study	+ Facilitatory (Caramazza, Laudanna, & Romani 1988)	?

Table 1. Summary of principal reported results for word perception, production, and acceptability in English. Example citations are given.

This study advances the body of research on pseudoword acceptability by demonstrating the importance of shallow morphology in acceptability judgments. These findings offer insight into the processing of novel words, which have no established meanings, and often are not completely decomposable into morpho-orthographic strings. Lexical innovation and encoding are key components in the process by which the lexicon grows and changes. Existing phonotactic and morphological patterns influence (and are influenced by) the encoding and adoption of new words in a feedback loop. Based on evidence from the current study, we suggest that the acceptability of novel words is enhanced by the recognition of established morphemes. With the further assumption that highly acceptable complex new words are more likely to enter the lexicon and be reused in the future, this result implies a dynamic in which existing morphemes are reinforced, and morphologically complex new words are preferred over simplex words of comparable length.

2. Background: Wordlikeness and Word Processing

The over-arching goal of the reported experiment is to provide a large dataset enabling replication and exploration of the factors influencing acceptability, including measures of phonotactics and neighbourhood density. The analysis presented in this paper establishes the contribution of these factors, and then pursues a post-hoc analysis relating to morphological effects on wordlikeness. In Sections 2.1 and 2.2, we review two general factors that are known to influence the wordlikeness of a pseudoword. One is the overall constraints on combinations of phonological elements in the language (*phonotactics*). The other is the extent to which the word is similar to, or reminds people of, specific words already known. A major approach to this lexical similarity is *lexical neighborhood density*. The influences of phonotactic probability and lexical neighborhood density are correlated, because a phonological combination has high probability if it is found in many words. However, the correlation is not perfect; the lexicon is composed of a haphazard subset of the allowable forms, and words that are similar to a pseudoword may or may not match the same parts of the pseudoword. Studies of speech processing have revealed that the two factors are dissociable (Vitevitch et al. 1999; Storkel, Armbrüster, & Hogan 2006), and so we consider them separately.

Morphological decomposability of the pseudowords may be understood as a further form of similarity between a complex pseudoword and the lexicon. There has been little work on the effects of morphological decomposability on wordlikeness. However, extensive work on its effects in processing, which we review in Section 2.3, points to the strong possibility of a positive effect.

2.1. Phonotactics

Speaker knowledge of phonotactics is gradient and probabilistic, so that there is a full spectrum of acceptability for possible words. This range of phonotactic acceptability is derived from lexical statistics: items with common sound sequences are judged better than those with rarer ones (Coleman & Pierrehumbert 1997; Vitevitch & Luce 1999; Frisch, Large, & Pisoni 2000; Bailey & Hahn 2001; Hay, Pierrehumbert, & Beckman 2004; Vitevitch & Luce 2004). Note that phonotactic knowledge draws on lexical statistics over word types, not tokens (Frisch, Large, Zawaydeh, & Pisoni 2001; Hay et al. 2004; Richtsmeier 2011). Modeling phonotactic likelihood probabilistically is the most common type of generative-grammar approach to wordlikeness, operating at the level of phones. A probabilistic model describes the observed phone sequence patterns in the language, building on frequency statistics over the set of all words in the speaker's experience. The resulting model describes the total space of possible phone sequences for that experience, so that it can parse or generate not only the input words, but a large set of unseen sequences.

The size of sequences considered in phonotactic models varies. Biphone statistics are widely used, offering a major improvement over uniphone statistics by capturing the tendency of consonants and vowels to alternate. But biphone statistics do not fully capture the syllable structure of languages such as English. Systematic effects at larger time scales include constraints on syllable contacts, distinctive patterns at word edges, and effects of word stress (see review in Pierrehumbert 2003). In order to capture these effects, other approaches use larger units of analysis: triphones, onsets/rimes, syllables, etc. (Coleman & Pierrehumbert 1997; Hay et al. 2004). Positional phonotactics are sometimes used to capture additional information from the lexicon (e.g., patterns relevant to word stress, syllable boundaries, and word boundaries), but such an approach is not feasible for our stimulus set. The stimuli for the present study include of a large number of short and long random pseudowords, which range from very pronounceable to unpronounceable. This means that it is not feasible to estimate syllable structure and stress for all of the forms in the stimulus set. It is also unlikely that even the 8400 stimuli would provide adequate balance and statistical power to consider positional phonotactics within the variety of possible stress and syllable contexts. The non-positional phonotactic model includes triphone units, which are large enough to encode patterns of syllable structure in English, as well as word-initial and word-final patterns. In addition, generalized triphone statistics provide coverage of syllable contact patterns (i.e., segmental probabilities across syllable boundaries)

Biphone and triphone models have a privileged status in phonotactics because they provide an efficient means for both word parsing (i.e., deciding if any given input string is licit, and calculating its probability) and word generation (Manning & Schütze 1999). They perform well in comparison to a lexicon that merely lists encountered words, because of their capacity to accept new, out-of-vocabulary words, while also being able to reject very unlikely words (i.e., words with very low or zero phonotactic probability scores). Because these sequential models are the simplest learnable system, it is important to explore the limits of their performance; more elaborate methods must be justified by surpassing that performance. The tractability of probabilistic n-gram approaches also mean that they are pervasive in computational applications like phoneme to grapheme (P2G) conversion and automatic speech recognition (Hahn, Vozila, & Bisani 2012; Jurafsky & Martin 2000).

Segmental n-gram approaches such as the biphone and triphone models used here have important limitations. Evidence for n-grams becomes increasingly sparse as the n-gram size increases, and some attested word patterns are well-explained by more abstract phonological elements (e.g., features, syllable structures) (Pierrehumbert 2003; Kager & Pater 2012). For humans, sparseness may mean that triphone statistics are not generally learnable; but they are potentially learnable for frequent triphones. Speakers may be able to make use of larger n-gram knowledge (e.g., triphones) when it is available, and 'back off' to their broader knowledge of biphone statistics otherwise. In natural language processing, 'smoothing' of higher order n-gram statistics by backing off or interpolating to lower-order statistics is used to mitigate sparse sampling issues (Jurafsky & Martin 2000); it is interesting to ask whether the cognitive system effectively uses the same strategy. In the analyses presented here, biphone and triphone phonotactics are treated as distinct factors, and their correlations with each other and with the wordlikeness ratings are assessed; simple weighted combination is represented by the independent inclusion of the biphone and triphone factors within the linear mixed-effects regression (LMER) models.

Word length must also be considered in wordlikeness models. Our stimulus set systematically covers the space of possible forms with 4, 5, 6, and 7 phones. This provides items long enough for pseudomorphology to appear, and allows the statistical models to control for length. An important constraint on possible wordforms is that long words are dispreferred. Simply recombining phonological elements in valid strings of arbitrary length would produce an exponentially increasing distribution of overall word lengths. In fact, the distribution is close to log-normal (Limpert, Stahel, & Abbt 2001). This result can be derived by imposing a cost for each additional unit (a mechanism stipulated in Daland 2015). To approximate the cost of additional units, we provide unnormalized phonotactic scores. Because the log of a likelihood is negative, each additional unit invariably lowers the score; but this approach is compatible with the finding that long words comprised of more probable parts are judged to have similar wordlikeness to short words made of less probable parts (Frisch et al. 2000). The effect of length in our dataset is illustrated by Figure 3 (Section 5), but we were not able to statistically assess item length effects due to technical limitations (described in Section 5).

2.2. Lexical Neighborhood Density

The lexical neighborhood is a major approach to word similarity; while phonotactic probability describes similarity to the overall lexicon, lexical neighborhood density relates to specific words in the lexicon. This method assumes that a form that differs from an existing word by exactly one phoneme counts as extremely similar to it. The set of such words—the lexical neighborhood of the target form—is the set of real words that can be formed by adding, deleting, or substituting a single phoneme (i.e., a phoneme edit distance of 1) (Coltheart, Davelaar, Jonasson, & Besner 1977; Grainger 1990; Luce, Pisoni, & Goldinger 1990; Marian, Bartolotti, Chabal, & Shook 2012). For real words, the effects of lexical neighborhood size on processing are dissociable from the phonotactics, and can vary depending on the task, either enhancing or degrading performance (e.g., accuracy or response time) (Vitevitch, Stamer, & Sereno 2008; Heller 2014). Short pseudowords, such as monosyllables and disyllables, are judged to be more wordlike if the lexical neighborhood is large than if it is small (Bailey & Hahn 2001). This result is easily understood as indicating that similarity to many existing words makes a pseudoword seem more like a real word.

The applicability of lexical neighborhoods for a general theory of wordlikeness is limited, however, by two properties of the way it is normally computed. First, nonwords such as *spt* may be completely unpronounceable and yet have many neighbors (*apt, opt, set, sat, spa, spit,* etc.). Second, long wordforms often have no neighbors, even if they are highly acceptable. Because the chance that a phonologically legal sequence is an actual word decreases with word length, the chance that a pseudoword has a minimal pair also decreases. The standard lexical neighborhood calculation depends solely on the number of minimal pairs, ignoring long word pairs that may be highly similar to each other because of the many respects in which they match. The forms *see* and *sue* are lexical neighbors, although they differ in 50% of their length; *mediation* and *radiation* are not, even though they match in a greater percentage of their length and might be easily confused in noisy conditions. Luce & Pisoni (1998), Bailey & Hahn (2001), Hahn & Bailey (2005), and Kapatsinski (2006) advance proposals to mitigate these problems by various elaborations of the basic approach, including expanded neighborhood definitions, more nuanced edit distance calculations, and length-normalization. The cognitive status of these more complex models is unclear, and the simple form is still commonly used (Storkel 2004; Marian et al. 2012; Heller 2014).

Further issues surrounding the lexical neighborhoods of long words are evident for languages in which words are normally longer than in English because of highly productive morphology. While neighborhood density interferes with the speed and accuracy of lexical processing in English, presumably due to the effects of lexical competition, it facilitates lexical processing in Spanish (Vitevitch & Rodríguez 2005). This result may be due to the fact that lexical neighbors are much more likely to be morphological relatives in Spanish than in English. Words sharing a morpheme are similar in meaning as well as form, so psycholinguistic research on morphological processing offers additional insight into how lexical similiarities may influence the wordlikeness of pseudowords.

2.3. Morphology in Processing

One study shows that phonotactic cues to morphology influence acceptability ratings of pseudowords: in their experiment on phonotactic effects on wordlikeness ratings, Hay et al. (2004) find that ratings are best predicted by the likelihood of the single best morphology-based parse. If the pseudoword contains a medial cluster that is more likely than not to span a morphological boundary, the word is evaluated as if it were morphologically complex. Since the stimuli in their experiment include no real English stems, this means that a bottom-up decomposition of the forms based on the phonotactics was involved.

At the same time, a large body of experimental research with visual priming for lexical decision has found evidence for shallow decomposition of words into morphemes (morpho-orthographic segmentation). Taft (2004) argues that morphological decomposition is obligatory when possible. Contra earlier findings (Marslen-Wilson, Tyler, Waksler, & Older 1994), this process is not constrained by the actual morpho-semantic derivation of the word. This means that *cleaner* primes *clean* (a transparent morpho-semantic relationship), but *corner* also primes *corn* (a morpho-orthographic relationship) (Rastle, Davis, & New 2004). Rastle et al. also considered the distinction between complete morpho-orthographic parses (e.g., *corner* is *corn* and the suffix *-er*) and partial parses (*cornea* is *corn* and the non-morpheme *ea*); they found that complete parses produced significantly more priming. Further research has yielded evidence that complete morpho-orthographical parsing is not required for priming of embedded stems (*cornea* priming *corn*): building on an ERP study by Morris, Porter, Grainger, & Holcomb (2011), Beyersmann et al. (2015) used a masked priming lexical decision experiment in French to compare priming across a variety of conditions. Primes included suffixed and unsuffixed pseudowords along with real complex words. English analogues for the French stimuli would be *teacher*, *teachness*, and *teachald* as primes for *teach*. For participants with high French proficiency (vocabulary and spelling performance), Beyersmann et al. found equivalent priming for all three related primes (in comparison to unrelated items). For low-proficiency participants, the non-suffixed items produced less priming. The results for high-proficiency participants imply that complete morpho-orthographic segmentation is not required for the activation of embedded stems; embedded stems are perceived even when the word includes additional non-morphemic letters. This pattern suggests that people are using a stem-finding approach (in which they identify embedded stems when possible), instead of relying on an affix-stripping mechanism (as in Taft & Forster, 1975). The reduced priming for non-suffixed items for low-proficiency participants may suggest that individuals with smaller vocabularies are more reliant on affix-finding than on stem-finding.

The results of Beyersmann et al. (2015) are extended to prefixed items in Beyersmann, Cavalli, Casalis, & Colé (2016); and a similar experiment in German (Hasenäcker, Beyersmann, & Schroeder 2016) replicates the pattern with adults and children. Hasenäcker et al. show priming for suffixed words, suffixed pseudowords, and non-suffixed pseudowords in comparison to unrelated controls; they also find that German adults showed relatively less facilitation for non-suffixed pseudowords versus suffixed pseudowords, while German children showed no difference between the two. They suggest that the adults are making use of affix-finding, while the children are relying on stem-finding; these strategies may have different cost–benefit ratios in German versus French, so that the German children in Hasenäcker et al. (2016) resemble the high-proficiency participants in Beyersmann et al. (2015).

Morpho-orthographic segmentation has also been shown for compound words, in addition to suffixed and prefixed stimuli: in a study of the processing of ambiguous novel compounds, Libben, Derwing, & Almeida (1999) find evidence for a prelexical parser that makes all possible analyses available to the lexicon. In a related study using semantic ratings with familiarity-decision and priming tasks, Libben, Gibson, Yoon, & Sandra (2003) show evidence for decomposition of semantically opaque compounds. Kuperman, Bertram, & Baayen (2008) report that information about both stems and suffixes within complex Finnish compounds is used immediately, before the full word has been accessed.

Insofar as facilitatory effects in processing are correlated with positive influences on wordlikeness judgments, these studies lead to the hypothesis that a shallow morphological parse should enhance wordlikeness of pseudowords, whether or not the parse provides a coherent semantic interpretation. A further and more anecdotal piece of evidence suggesting this hypothesis is the behavior of familiarity ratings for rare words. It is well-known that word frequency and word familiarity ratings are poorly correlated (Nusbaum, Pisoni, & Davis 1984; Connine, Mullenix, Shernoff, & Yelen 1990). The Hoosier Mental Lexicon (19,320 words) described by Nusbaum et al. includes many compounds like *manhunt* that have low frequencies and high familiarity ratings. Using substring matching to identify words with a (shallow) compound parse, we considered the forms with the lowest frequency (F=1). For this subset (N=10,355), we found that forms had a median familiarity of 4.8 on a scale of 1 to 7 if there was no parse (N=3,012), 5.1 if there was a partial parse (N=3,821), and 5.8 if the form was fully decomposable (N=1,522). This observation, which has not been previously reported to our knowledge, suggests that familiar subparts can boost the apparent familiarity of forms that are not in fact very familiar.

3. Materials

We developed a new, flexible pseudoword generator and used it to generate a set of 8400 pseudowords: *PseudoLex*. PseudoLex was normed and validated through an experiment collecting wordlikeness judgments. Our generator was designed to enable replication and extension of the relationship between wordlikeness and phonotactics. For the post hoc analysis of morphological decomposition effects, shallow morphology was estimated using automatic processes (see Section 3.4).

3.1. How PseudoLex Items are Generated

For PseudoLex, statistical phonotactic models are used to generate the 8400 items. The models are trained on a CELEX-based corpus of 11382 monomorphemic words in phonemic representation (Baayen, Piepenbrock, & Gulikers 1995). The training corpus was hand-edited to ensure that words were monomorphemic (Hay et al. 2004). Complex words were excluded from the training corpus because many contain phone sequences that are not found in monomorphemic words (e.g., *hotdog*, *kindness*), and we wanted to avoid forms with a strong phonotactic cue for a morphological boundary in the outputs of the generator. Three models are trained: triphone, biphone, and uniphone. Word boundaries are encoded as null phones; no other positional information is included. The triphone and biphone models are stored in the form of n-gram transitional probabilities. The uniphone model is stored as a table of overall phone probabilities. The trained models are used to generate random pseudowords of 4, 5, 6, and 7 phones. The uniphone model is also used to generate nonwords, which serve as corpus-matched filler items. The trained models are used to assign biphone and triphone scores to items generated with either grammar, and to ensure the nonwords are indeed illegal strings. Because the illegal filler items contain sequences with transition probabilities of zero, these items do not have well-defined scores when scores are calculated in the standard manner using log probabilities.

3.2. Phoneme to Grapheme Conversion Using Phonetisaurus

For experiments presenting pseudowords visually, stimuli need to be represented orthographically. Phoneme-to-grapheme (P2G) conversion is required for our phonemically-generated pseudowords. Phoneme-to-grapheme conversion is an issue for opaque spelling systems (e.g., that of English), in which the mappings between phonemes and graphemes (*graphones*) are frequently irregular or ambiguous. There is frequently more than one single correct orthographic rendering for a given phonemic pseudoword, and vice versa. Figure 1 gives an example of possible graphone mappings in the word *phoenix*. Proficient speakers of a language are skilled at this process, but hand-coding items is both laborious and subject to bias. To address this problem, we used *Phonetisaurus*, a state-of-the-art computational tool for G2P conversion (Novak, Yang, Minematsu, & Hirose 2011). In testing with other top G2P tools, Phonetisaurus has excellent accuracy (Hahn et al. 2012). Using a computer-based tool avoids the biases of hand-coding, and quickly handles the thousands of items required for this study. Detailed discussion of Phonetisaurus and our use of it is found in Appendix C.

```
PH  OE  N  I  X
|   |   |  |  |
f   i   n  I  ks
```

Figure 1. Example graphone mappings.

3.3. Wordlist

We generated 8400 pseudowords based on a monomorphemic subset of CELEX. Because some CELEX pronunciations come from a non-rhotic variety of English, pseudowords containing /r/-colored vowels or linking-/r/ segments were excluded to make this stimulus set useful across a wider range of populations. The phonemically generated pseudowords were converted to orthographic representation for the visual wordlikeness task by the Phonetisaurus tool (also trained on a CELEX lexicon). We excluded items that: 1) failed the G2P mapping stability filter (see Appendix C); 2) contained orthographic substring matches to a compiled knockout list of 1042 vulgar or obscene terms; 3) were homophones of existing words in CELEX; or 4) were homographs of existing words. Homographs were detected using the Corpus of Contemporary American English (COCA) (Davies 2008); COCA was used for this purpose because it is slightly larger than CELEX (100,803 vs. 89,871 wordforms), and it has better coverage of American vernacular. Homophones and homographs were not common: for example, of length 6 pseudoword candidates, 0.14% were excluded as homophones, and 0.28% of items were excluded as homographs.

Biphone and triphone probability scores are calculated for each generated pseudoword. To ensure coverage of the full range of phonotactic likelihood, the stimulus set consisted of 1200 items in each of 7 categories: items from the first, second, and third tertiles of triphone scores; the same distribution for biphone scores; and uniphone items illegal in the biphone and triphone grammars. For each category, 300 items were generated with 4, 5, 6, and 7 phones, to create 28 cells (see Table 2). These lengths were chosen to include and extend on pseudoword stimuli used in previous studies. Random sampling of stimuli in each cell ensures that the full ranges of scores are evenly covered, and that neither the biphone nor triphone model is privileged. Biphone and triphone scores for each item are strongly correlated (r = 0.81), but this design means they have equal footing in our models.

Item Length	Uniphone Generated	Biphone Generated			Triphone Generated		
		Low Score	Med Score	High Score	Low Score	Med Score	High Score
4 phones	ngiac kjkd	liku orphab	roiet emboy	hanch swong	jolsh ertav	focar theoroi	morbi lont
5 phones	ccusfc ootplp	elvial thyroil	lemurch caread	pardos digot	ofluth jaystow	daporp biahaw	peleos sordna
6 phones	tnjayout udgvtnm	auxald arthralm	eyprithy axallia	allownser phispath	odeckyo poutiki	eptuo whenmaph	pulview egugong
7 phones	nftcngick dfpkeps	uccoirstoi thworbizar	totisual esierrian	fequoisa drublod	urialerau ogunkeb	loiterpum afrannoys	obversing doyenvom

Table 2. Example stimuli for the 28 cells of the current study. Each cell represents 300 stimuli.

3.4. Morphological Decomposition

To explore the role of morphological similarities to existing words, the stimuli were analyzed post hoc for suffixation and compounding patterns. These are both heavily used in English words, whereas prefixation is less common. Items were coded as having a full suffixation parse, a partial suffixation parse, or no suffixation parse ('suffix_full', 'suffix_partial', 'suffix_none'); and as having a full compound parse, a partial compound parse, or no compound parse ('compound_full', 'compound_partial', 'compound_none').

Suffixation was determined using the standard Lancaster stemmer, as implemented in NLTK (Paice 1990; Bird, Loper, & Klein 2009). Compound parses were found by substring matches to the CELEX English lexicon. Neither method uses syntactic or semantic analysis, and both are based on orthography. We define a full suffixation parse as occurring when the stemmer output is a real English word (CELEX English) (e.g., *puck* + *-ing*). Note that no constraints on part-of-speech have been imposed. A partial (*pseudosuffixation*) analysis occurs when the output stem is a pseudoword (e.g. *thraf* + *-ium*, *surpit* + *-ual*). The minimum suffix length is 1 letter, and the minimum length of the residue after suffix parsing is 2 letters. Similarly, a full compound analysis means that the pseudoword is a concatenation of two existing English words (e.g., *hypodeck, aftertook, sellfilth*); the minimum length of subword is 3 letters, and the minimum length of residue after compound parsing is 2 letters. A partial (*pseudocompound*) analysis contains one English word, with the residue being a pseudoword. Examples of forms with a partial compound analysis include *churcharou* and *affreap*. These compounds have no established meanings. Because no syntactic analysis is performed, they do not necessarily conform to productive compounding strategies for English. However, meanings for many of them can be imagined. For example, if a pickpocket picks valuables from pockets, a 'sellfilth' might sell unsanitary products, or filthy gossip for tabloids. Additional examples of decomposable stimuli are shown on Table 4 in Section 6.1.

This morphological analysis implicitly assumes that participants are recognizing apparent morphemes whenever they are present (according to the compounding and suffixation estimators). The statistical analyses presented do not consider that there is variation in the parseability of the items; e.g., that some morphemes are more frequent than others, or that the phonotactic probability of the estimated morpheme boundary may be higher or lower. Needle & Pierrehumbert (in press) have shown that even strong boundary cues had a modest effect on the parsing of partially-decomposable pseudowords, while the apparent presence of morphemes was a stronger effect. Because the stimuli in the current study were generated from a monomorphemic wordset specifically to avoid strong boundary cues, this simplification is appropriate (see Section 3.1). Overall, our morphological analysis is conservative: we did not look for analyses involving prefixes, or words embedded in the middle of a pseudoword leaving unanalyzed material on both sides. We selected our method because it is highly replicable and minimizes the need for additional assumptions. Example parses are shown in Table 4; note that some real stems may be unfamiliar to the participants, so some 'full' parses could function as 'partial' parses.

Item Length	No Parse	Suffixation Parse		Compound Parse	
	None	Partial	Full	Partial	Full
4 phones	peld shreath	lurp+ed onf+er	hep+s	ay+leach re+bay	boo+goo ark+off
5 phones	snumph dovio	murph+al bluck+ed	kilo+th	push+el yo+down	bow+gush wool+pay
6 phones	phanuct obstoon	roid+als phasan+ia	burthen+th	ang+stalk oro+fowl	dig+wick drown+joy
7 phones	phalamang wodazook	cinct+ual sug+anian	–	cook+ivert fank+foil	face+dummy hypo+deck

Table 4. Example parsesfor stimuli by length and parse type. Morpheme boundaries are marked with '+', and '–' indicates no examples exist.

Of items with any parse, 29% have both suffixation and compound parses. The analyses generated by the Lancaster stemmer and the compounding analysis may include both spurious and missed parses. For example, the forms *snuffy* and *crassy* are not decomposed, because the stemmer recognizes the suffix *-y* only after specific consonants. The form *erfletul* is analyzed as containing the suffix *-ul*, which would not be familiar to most English speakers. These errors occur because the irregularities in English lead to a tradeoff between accuracy and precision in the rules. Note also that the Lancaster stemmer matches multiple suffixes in succession; e.g., in *dumpouser*, both *-er* and *-ous* are matched (*dump* + *-ous* + *-er*). In our analyses, such cases are treated as if the parse yielded a single combined suffix (*ouser*). When affix combinations occur in the lexicon, it can be semantically and statistically justified to treat them as morphemes in their own right (Stump 2018). Thus, it is not clear if participants would obligatorily decompose all apparent suffixes in pseudowords. We did not wish to compromise the objectivity of our analysis by readjusting the rule set post hoc. We will return below to the consequences of this situation for the data analysis.

4. Data Collection

The norming study used the PseudoLex stimuli in a visual wordlikeness task. In an online Amazon Mechanical Turk experiment, 1440 native US English speakers provided Likert-scale wordlikeness judgments of 140 pseudowords each, as well as completing a vocabulary assessment, and a rhyming task.

4.1. Methods

4.1.1. Participants. The study collected data from 1440 participants via Amazon Mechanical Turk (825 female, 608 male; 7 participants declined to provide gender). All participants were English speakers (5 participants reported other "main" languages, but their performance passed the quality control standards of the experiment), and 1438 participants currently reside in the United States. Reported birth years range from 1945 to 1996 (26 participants declined to answer). All participants completed the experiment between 2014-06-02 and 2014-06-13. Participants were paid $3 for completing the task.

Recruiting participants through AMT and other online sources is increasingly popular in psycholinguistics because it can efficiently provide large datasets of high quality (Snow, O'Connor, Jurafsky, & Ng 2008; Warriner, Kuperman, & Brysbaert 2013). Wurm, Cano, & Barenboym (2011) report higher response variability for an online versus an in-lab task. Some of this variability may arise from uncontrolled variability in the experimental conditions. However, some is likely to reflect individual variation and capturing it may be a useful step towards understanding the natural range in human cognition. Current lab studies are unduly reliant on Western college undergraduates as participants (Henrich, Heine, & Norenzayan 2010), and online data collection makes it possible to recruit a more diverse participant pool (Gosling, Sandy, John, & Potter 2010).

4.1.2. Materials and Presentation. The 8400 pseudowords, as described above, were block-randomly distributed into 1440 experiment scripts of 140 stimuli each (5 stimuli for each of the 28 cells). The scripts are semi-overlapping, so the design gathered 24 ratings from different participants for each pseudoword. The experiment included 2 supplemental tasks: a word familiarity task to assess vocabulary level (based on Frisch & Brea-Spahn 2010), and a rhyming task. The vocabulary task includes 70 items: 10 nonce words (e.g., *impiroxin*), 10 very common words (e.g., *statue*), and 50 test words of varying familiarity (e.g., *tabby*). The 50-item rhyming task was developed to assess dialect differences. No significant effects of rhyming task performance were found, so the results are not presented here.

4.1.3. Procedure. Participants chose the experimental "human intelligence task" from the AMT interface and were directed to the web-based experiment. The experiment consisted of three tasks. The pseudoword rating task was first. Each participant was instructed to give each item a rating of 'English-like-ness' on a 5-point Likert scale. Participants were told to pronounce each word aloud, and to base ratings on the sound, not the spelling, of the pseudowords. The experiment enforced a 600ms delay between the presentation of each pseudoword and the acceptance of a response. After completing the 140 pseudoword ratings, the participant performed the second task: 50 pairs of rhyming judgments. The third task was the vocabulary assessment. Participants were instructed to rate each word by how familiar it seemed on a 5-point Likert scale. The nonce words and highly familiar words in the test are used as catch items to exclude participants who do not follow the instructions. The ratings of the 50 test words are used to calculate the vocabulary score, with all words weighted equally. The three tasks together took a maximum of 30 minutes. Instructions for each task are found in the appendices.

4.2. List of Effects to be Replicated

Prior to investigating the role of morphological decomposition in judgments of pseudowords, we first verify that the experiment replicates some important effects previously reported. In addition to pure replication, we are interested to see how these effects are shown for the PseudoLex stimuli, which are designed to be more varied than the pseudoword stimuli in many previous studies.

4.2.1. Phonotactic likelihood. Phonotactic likelihood has been shown to correlate with wordlikeness judgments in previous research, but previous studies have largely focused on shorter words. Here, we seek to replicate the correlation for the shorter items, and determine the extent to which it holds for longer items and for less-probable items. We evaluate both triphone and biphone models. Traditional biphone-only versions fail to capture some phonotactic constraints that are known to be psycholinguistically relevant, as discussed above. Triphone models can capture some of these effects, such as word-edge and syllable contact effects, as well as short morphemes. However, because there are many more possible triphones than biphones, triphone statistics cannot be estimated as reliably from a lexicon of realistic size; see further discussion in Pierrehumbert (2003). Here we ask whether triphone statistics can improve model predictions, in comparison to biphone statistics alone. Biphone and triphone phonotactic probability scores for each item are cumulative log transitional probabilities, centered in the LMER models. Nonword items do not have a well-defined log probability score and were excluded from LMER analysis. These illegal items should be rated less wordlike than the pseudoword items.

4.2.2. Orthotactics. PseudoLex was designed to minimize the effects of irregularities in the English spelling system. We verify this effort by asking whether orthotactic scores provide any additional predictive power beyond phonotactic scores.

4.2.3. Vocabulary level. Frisch & Brea-Spahn (2010) found that participants with larger vocabularies judge items more favorably, suggesting that high-vocabulary participants are more familiar with rare phonotactic sequences. We seek to replicate this effect with the more varied set of pseudowords found in PseudoLex. Vocabulary level for each participant is a continuous integer measure from 50 to 250, the sum of the Likert ratings for the 50 test items (M = 168, s.d. = 32.32). This measure was centered in LMER models.

4.2.4. Word length. Controlling for local phonotactic likelihood, longer items should have lower wordlikeness judgments (Frisch et al. 2000). A phonotactic likelihood score that is not normalized for item length predicts this effect qualitatively, because the overall score tends to decrease with each additional phone. PseudoLex includes a phonotactically balanced sample of words of four different lengths (4, 5, 6, and 7 phones). We ask whether there is a systematic decrease of rating for these pseudowords, which represent a more diverse set than those used in Frisch et al. (2000).

4.2.5. Lexical neighborhood size. In previous studies of wordlikeness, an important predictor is lexical neighborhood size, defined as the number of words with a string-edit distance of 1 from the target word. This measure was developed for studies of monosyllabic pseudowords. We ask whether it is also relevant for the more comprehensive sampling of the phonological space in PseudoLex. The orthographic neighborhood size for each pseudoword was calculated using CLEARPOND (Marian et al. 2012). The CLEARPOND lexicon is built from the SUBTLEX movie subtitle database, a more natural and current lexical inventory than the Hoosier Mental Lexicon used in earlier work. The measure ranges from 0 to 19 (M = 0.51, s.d. = 1.46). The distribution is highly skewed (for 79% of items, the neighborhood size was 0), so neighborhood density was included in models as a Boolean factor ('Does the item have neighbors?': 'True' or 'False').

5. Replication Results

Before presenting our complete statistical analysis, we illustrate graphically how three of the most important influences on wordlikeness appear in our data: biphone score, triphone score, and word length. Figure 2 plots the relationship of biphone score and triphone score to wordlikeness judgments. The scores shown on the x-axis are cumulative log transitional probabilities for biphones and triphones. In both plots, the mean ratings for nonword items are much worse than the least likely pseudowords. Biphone and triphone scores appear strongly positively related to wordlikeness.

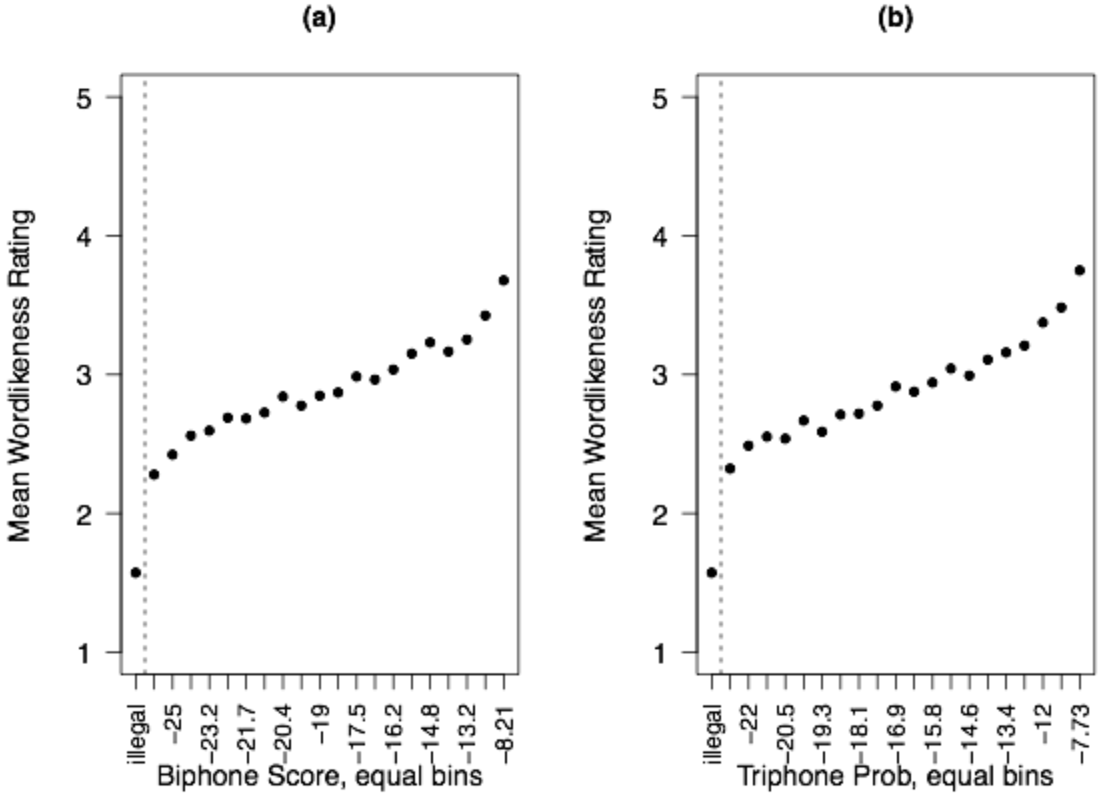

Figure 2. Mean wordlikeness rating by log phonotactic probability scores. Bins contain equal observation counts, pooled over all lengths: a) biphone score, b) triphone score. The filler items (labeled as “illegal” on the x-axis) are rated lower than the lowest-scored legal items. On the average, biphone and triphone scores both correlate positively with wordlikeness ratings.

Figure 3 illustrates the decline in word score and wordlikeness ratings with word length for the biphone and triphone scores.

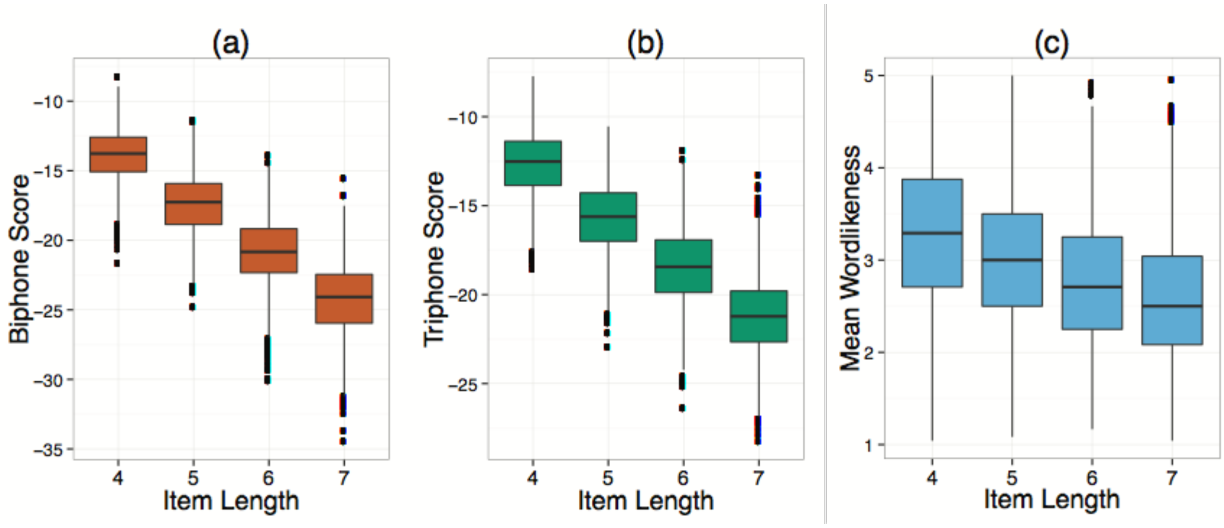

Figure 3. Boxplots of phonotactic scores and wordlikeness ratings, separated by word length: (a) biphone scores, (b) triphone scores, (c) wordlikeness ratings.

The relationship of wordlikeness ratings to all five of the effects to be replicated was evaluated using linear mixed-effects regression (LMER) implemented in R package lme4 (Bates, Maechler, Bolker, & Walker 2015) in R (R Core Team 2014). We first consider models of the replication effects, excluding orthotactics (Sections 5.1 to 5.4). Then, we test the effects of substituting orthotactics for phonotactics in equivalent models (Section 5.5). Later, we investigate the consequences of adding morphological effects (Section 6). All models include random intercepts for subjects and items. All continuous measures were centered (i.e., biphone and triphone probability scores, vocabulary level). Because of issues of stability and convergence with the LMER models, it was necessary to divide the analysis into 4 models by item length; these technical issues may be related to the high correlation between word scores and length. This means that the relationship of length to wordlikeness judgments cannot be directly statistically evaluated here, either as a main effect, or in interaction with other effects. The length factor appears to be related to multiple other factors affecting wordlikeness (e.g., morphological decomposition, discussed later), both positively and negatively, so isolating a possible effect of length per se will require further research to control for these length-related factors.

For each length, an initial model was defined to include all main effects (biphone and triphone score, vocabulary level, neighborhood density) and all 2-way interactions of the main effects. These initial models were pruned to yield the final models; during pruning, factors were removed if their inclusion could not be supported (i.e., caused failures of model convergence), or for insignificance (i.e., $t < 2$). To prevent unreasonable collinearity in the model, a criterion of kappa < 10 was imposed; all kappa values in the models presented are less than 7. The significance of all reported factors and interactions was confirmed using model comparison ($p < 0.05$, X^2 method); these values are reported in Appendix E. The four resulting models (Replication models) are summarized in Table 3; information for factors excluded from a model is marked by ‘–’. Models in the Replication set are suffixed with ‘A’.

Replication Model Factors	Length 4A		Length 5A		Length 6A		Length 7A	
	β	*t*	*β*	*t*	*β*	*t*	*β*	*t*
biphone	0.046	4.69	0.058	7.40	0.061	9.22	0.067	11.39
triphone	0.109	10.40	0.073	8.95	0.081	11.07	0.070	10.19
vocabulary	0.003	7.35	0.003	7.75	0.004	8.27	0.004	9.07
neighbors	0.564	16.23	0.559	13.97	0.811	9.57	0.693	2.43
neighbors:vocabulary	-0.001	-3.24	–	–	–	–	–	–
biphone:vocabulary	–	–	–	–	–	–	0.0001	2.49

Table 3. Summary of factors for the four Replication LMER models. Factor estimates and t-values are given for each model. Only significant factors are shown.

5.1. Phonotactic score. In the models for each length category, both biphone and triphone phonotactic scores were significant positive predictors of wordlikeness rating; increased phonotactic score was associated with increased wordlikeness ratings. Model comparison showed that both biphone and triphone factors significantly improved the model fits, and that removing the triphone score generally reduced model fit more than removing the biphone score: for Length 4A, the difference in $X^2(1)$ for dropping biphone vs. triphone is 21.91 vs. 105.05; for Length 5A, 53.99 vs. 78.28; for Length 6A, 83.04 vs. 118.52; and for Length 7A, 125.20 vs. 100.89. The increased model fit from including the triphone score in the models may indicate that biphone-only scores fail to capture many aspects of English syllable structure and syllable contact constraints that are captured by triphone scores. Triphones may also capture some highly productive morphemes.

5.2. Vocabulary level. The participants' vocabulary level is a significant positive predictor of wordlikeness ratings across item lengths; high-vocabulary participants show a general tendency to rate items higher. Factor estimates and t-values are reported in Table 3 as 'vocabulary'. Vocabulary level is also involved in significant interactions, reported below.

5.3. Lexical neighborhood. The presence of one or more orthographic neighbors provides a significant positive influence on an item's wordlikeness rating (see Table 3, 'neighbors'). This effect is present across all lengths, though the number of items with one or more neighbors falls sharply as item length increases: at length 4, there are 1067 such items (of the 1800 total items), while length 5, 6, and 7 have 352, 60, and 5, respectively.

5.4. Factor interactions. As shown on Table 3, there are two significant interactions in this set of models. The Length 4A model contains an interaction of the neighborhood density factor with vocabulary level ('neighbors:vocabulary'): the wordlikeness boost for having neighbors is larger for participants with lower vocabulary levels. The Length 7A model contains an interaction of biphone score with vocabulary level ('biphone:vocabulary'): the positive effect of biphone score on rating is larger for participants with higher vocabulary levels. This effect is small and does not survive in any of the Decomposition models (in the follow-up analysis, below).

5.5. Orthotactics. The 8400 pseudoword items in PseudoLex were designed to have a close correlation between phonotactic score and orthotactic score. This allows the items to be used in visual experiments with confidence that orthotactic effects are not being confused with phonotactic effects in participants' ratings. In the subset of items with legal bigraph and trigraph scores (5572 items), the correlation of orthotactic and phonotactic score is high: r = 0.84 for digrams, r = 0.82 for trigrams. When combining orthotactic and phonotactic scores into the same LMER models, there were issues with convergence and stability. We instead compared the wordlikeness effects of orthotactics and phonotactics by running an additional set of LMER models using the 5572-item subset; these correspond to the Replication models described on Table 3, in which bigraph and trigraph orthotactic score factors were substituted for the biphone and triphone phonotactic score factors. These equivalent models are similar overall; to the extent that they differ, the fit of the phonotactic versions is slightly superior. The correlation of residuals between the phonotactic and orthographic models is > 0.999, indicating that the deviations of specific items from the overall trends in each model are similar.

6. Effects of Pseudomorphology

The Replication models presented on Table 3 enable the Decomposition analyses for wordlikeness rating by controlling for the fixed effects in the models: biphone and triphone phonotactic scores, lexical neighborhood effects, and participant vocabulary levels. The Replication models include random effects in the form of intercepts for each item and participant, which function as idiosyncratic adjustments to the predicted ratings; e.g., participant intercepts adjust for a specific participant's tendency to rate items higher when controlling for other factors, and item intercepts adjust for a specific item's tendency to be rated higher when controlling for other factors. Patterns in the item intercepts can provide a clue that the model is missing important factors affecting wordlikeness. We examined the items with high and low intercepts (i.e., items consistently rated more or less wordlike than predicted), and we noticed that high-intercept items often contained recognizable morphemes, whereas low-intercept items never did. In the following analysis, we demonstrate that items which may be parsed as containing at least one morpheme are rated significantly more wordlike than items lacking a morphological parse.

6.1. Morphological decomposition. Two morphological processes were explored: suffixation and compounding. Recall from Section 3.4 that items were coded as having a full suffixation parse, a partial suffixation parse, or no suffixation parse ('suffix_full', 'suffix_partial', 'suffix_none'); and as having a full compound parse, a partial compound parse, or no compound parse ('compound_full', 'compound_partial', 'compound_none'); these categories are described in more detail in Section 3.4. Of items with any parse, 29% have both suffixation and compound parses. The effect of compound parses on the distribution of the intercepts is shown in Figure 4; items with a suffixation parse are excluded. Longer items are more likely to have a compound parse than shorter items; after length 4, presence of a parse is significantly more likely than no parse. We also see the positive effect of a compound parse on the intercept: in each case, the mass of the distribution for pseudocompound items is further towards the right (the items are more wordlike) than for noncompound items. The same relationship also holds for complete compound parses versus partial compound parses; however, such items are rare (less than 10% of all compound parses).

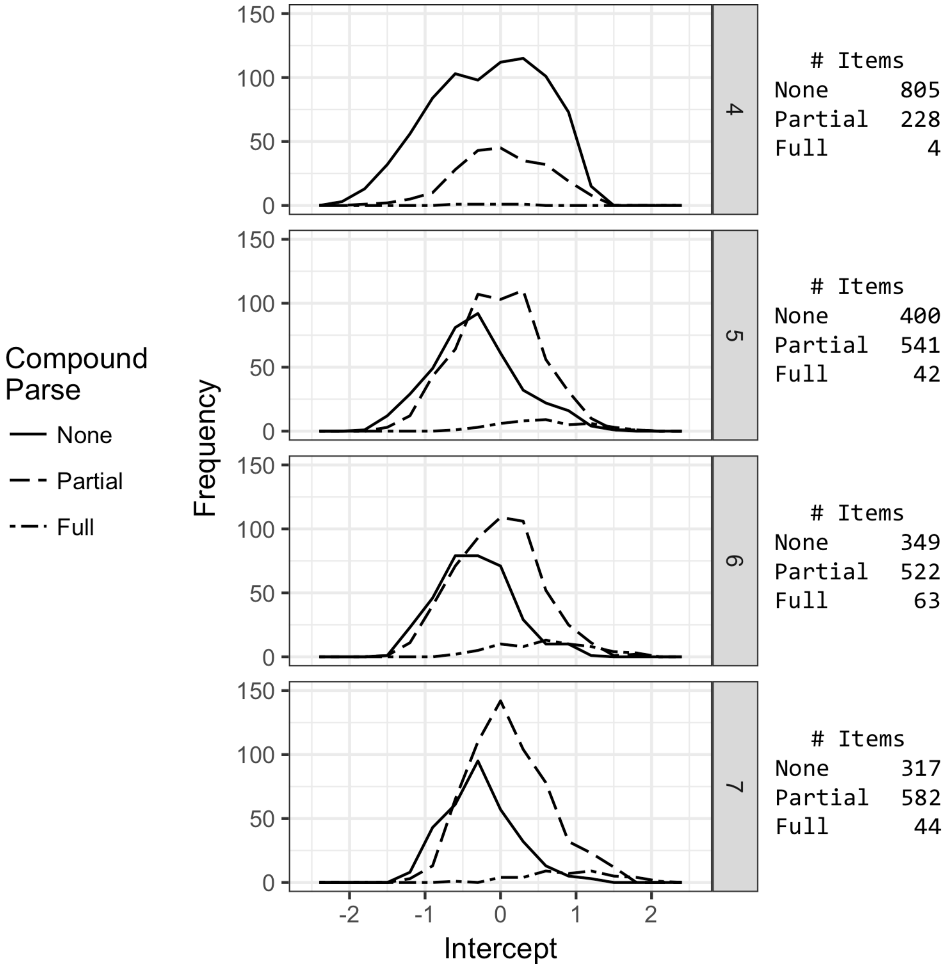

Figure 4. Effect of a compound parse on the distribution of wordlikeness intercepts, for pseudowords of length 4 to 7. Each panel shows superimposed histograms of the number of pseudowords having the indicated intercept value. The count of items in each category is given to the right.

The pattern is similar for suffixation, though there are key differences. Figure 5 displays the results of the suffixation analysis in the same format; items with a compound parse are excluded. Longer words are more likely to have a suffixation analysis than shorter words. Items with a partial suffixation analysis are generally rated higher than items with no suffixation analysis. The most notable difference between Figure 4 and Figure 5 is that suffixation analyses are less common than compound analyses for all pseudowords except those of length 4, where they are much more common. Similarly, full suffixation analyses are much less common than full compound analyses as length increases.

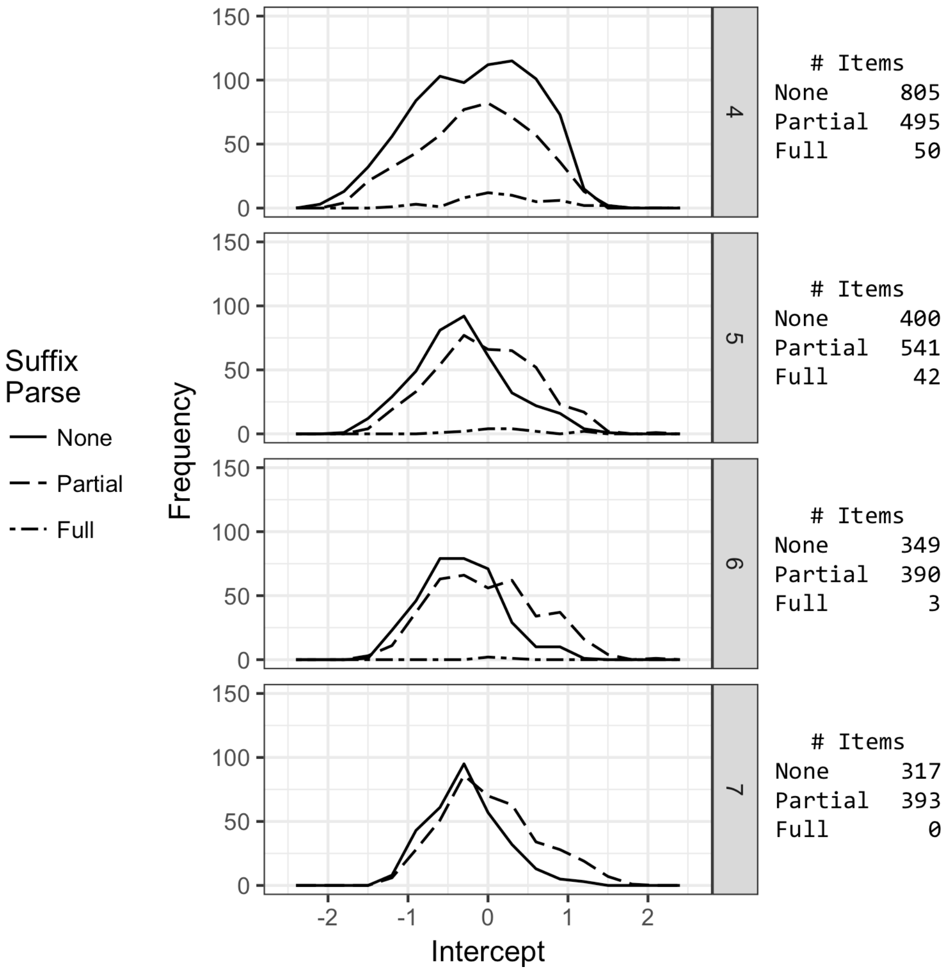

Figure 5. Effect of a suffix parse on the distribution of wordlikeness intercepts, for pseudowords of length 4 to 7. Each panel shows superimposed histograms of the number of pseudowords having the indicated intercept value. The count of items in each category is given to the right.

6.2. Modeling decomposition effects. In order to evaluate the significance of the patterns described in Section 6.1, a new set of mixed-effects models ('Decomposition') was generated by including both the suffixation and compounding factors as fixed effects; models in the Decomposition set are suffixed with 'B'. As before, an initial model was defined to include all main effects (biphone and triphone score, vocabulary level, neighborhood density, suffixation, and compounding) and all 2-way interactions of the main effects. These initial models were pruned by removing insignificant factors to yield the final models; see Table 5. Suffixation and compounding factors are combined in these models, meaning that a single item may simultaneously benefit from both parses; it is even possible that both methods result in the same parse. For example, *dumpouser* is parsed as being the suffixation of *dump* + *-ouser*, but also as a partial compound of *dump* with the pseudoword remainder *ouser*.

In this augmented model set, the influence or significance of the previously-reported main effects (biphone and triphone score, vocabulary level, and neighborhood density) are similar to the Replication models. The interaction of 'neighbors:vocabulary' in the Length 4A model is also nearly identical in 4B, but the interaction of biphone:vocabulary does not carry over. This stability indicates that the morphological factors are explaining additional variation in wordlikeness. However, the effects of morphological factors are complex, with effect directions and significance levels differing for items of different lengths. Some of the statistical interactions are interpretable, while others appear to arise as artifacts from the automated analysis.

Decomposition Model Factors	Length 4B		Length 5B		Length 6B		Length 7B	
	β	*t*	*β*	*t*	*β*	*t*	*β*	*t*
biphone	0.049	5.13	0.058	7.93	0.059	9.60	0.057	7.37
triphone	0.106	10.49	0.086	7.86	0.071	10.32	0.055	8.71
vocabulary	0.003	7.35	0.003	7.75	0.003	6.29	0.003	5.74
neighbors	0.617	18.20	0.688	12.96	0.745	9.32	0.748	2.87
compound_partial	0.404	11.12	0.312	9.62	0.279	9.51	0.345	12.07
compound_full	0.366	1.67	0.975	11.72	0.813	12.61	1.147	15.53
suffix_partial	–	–	0.212	7.11	0.243	8.50	0.245	8.75
suffix_full	–	–	0.521	7.87	0.320	2.81	0.354	1.60
neighbors:vocabulary	-0.001	-3.25	–	–	–	–	–	–
neighbors:compound_partial	–	–	-0.272	-3.80	–	–	–	–
neighbors:compound_full	–	–	-0.555	-2.27	–	–	–	–
biphone:suffix_partial	–	–	–	–	–	–	0.022	2.09
biphone:suffix_full	–	–	–	–	–	–	0.021	0.29
triphone:suffix_partial	–	–	-0.032	-2.19	–	–	–	–
triphone:suffix_full	–	–	-0.092	-3.11	–	–	–	–
vocabulary:compound_partial	–	–	–	–	0.001	3.31	0.001	3.90
vocabulary:compound_full	–	–	–	–	0.002	3.34	0.002	3.01
vocabulary:suffix_partial	–	–	–	–	–	–	0.001	2.26
vocabulary:suffix_full	–	–	–	–	–	–	0.003	1.63

Table 5. Summary of factors for the four Decomposition LMER models. Factor estimates and t-values are given for each model. Only significant factors are shown.

For all lengths, the presence of a partial compound parse (‘compound_partial’) has a significant and positive effect on wordlikeness rating (see factor estimates and t-values on Table

5). For all lengths except length 4, the presence of a complete compound parse (‘compound_full’) yields a significant and larger positive effect on wordlikeness; because there are only 9 compound_full items at length 4, this gap may be due to insufficient power. In general, the wordlikeness increase from a compounding parse is larger than the increase from a suffixation parse. The compound effect may increase with item length, perhaps because of greater salience for embedded words that are longer.

The compound parse factor also has 3 significant interactions across the model set. In the Length 6B and Length 7B models, compound parse interacts with vocabulary level (‘vocabulary:compound_partial’, ‘vocabulary:compound_full’): the positive effect of vocabulary level is significantly increased when a partial compound parse is present, and further increased when a complete compound parse is present. This interaction is illustrated for Length 7 in Figure 6.

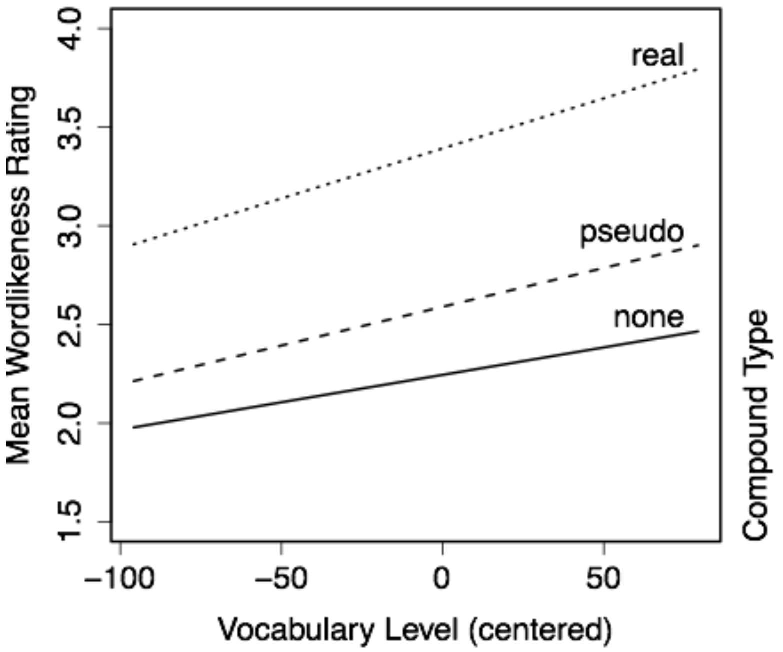

Figure 6. Interaction of vocabulary level with compound type as captured in Decomposition models, for pseudowords of length 7. Matches to existing words of English have a greater positive effect on ratings by people who know more words.

Compound parsing also interacts with neighborhood density in the Length 5B model (‘neighbors:compound_partial’, ‘neighbors:compound_full’). The positive effect of lexical neighbors on wordlikeness rating is significantly reduced when a partial or full compound parse is present. Examination of the specific items that are responsible for this interaction suggests, however, that it is an artifact of unreliable morphological analysis for items of length 5. The difference between full compounds with and without lexical neighbors rests on only three items that are analyzed as full compounds and have lexical neighbors: *chippert*, *yonnet*, and *modgem*. It is far from clear that the embedded words in these items are as psychologically salient as the corresponding full compounds without lexical neighbors, such as *arcterm* and *bowgush*. Amongst words with lexical neighbors, the distinction between pseudocompounds and non-compounds also appears to be unreliable for items of length 5. Some items with salient embedded words, such as *loyalk* and *moisto*, are analyzed as non-compounds, whereas highly similar forms such as *mortark* are analyzed as pseudocompounds. The presence of *mortar* in *mortark* is probably more salient than the word *ark* found by the algorithm. Such examples raise the possibility that the benefit of having a lexical neighbor might really be uniform across words of different morphological status. However, more detailed psycholinguistic studies of morphological decomposition would be a prerequisite to developing a more sophisticated parsing algorithm that could avoid idiosyncratic analyses like those just mentioned.

The suffixation parse factor (labeled 'suffix_full' and 'suffix_partial') could not be included in the Length 4B model, because its inclusion made the model unstable. For the other lengths, the presence of a partial suffix parse ('suffix_partial') yields a significant positive effect on wordlikeness (see factor estimates and t-values on Table 5). The presence of a complete suffix parse ('suffix_full') has a significant and larger positive effect in the models for Length 5B and Length 6B; note that the t-value of 'suffix_full' in Length 7B falls below our significance criterion (*t* = 1.60), though the suffix parse factor as a whole is significant.

The suffixation factor has 3 significant interactions across the model set. In the Length 7B model, suffixation interacts with vocabulary level ('vocabulary:suffix_partial') and with biphone probability ('biphone:suffix_partial'); as with the suffix parse main effect, the specific factor level 'biphone:suffix_full' in this model falls below our significance criterion (*t* = 0.29). The positive effect of higher vocabulary is significantly higher when a partial suffix parse is present. This effect is analogous to the effect of a compound analysis, but is smaller, as shown in Figure 7. The cross of the 'full' and 'partial' lines is unlikely to be consequential, because only a small minority of participants have such low vocabulary scores.

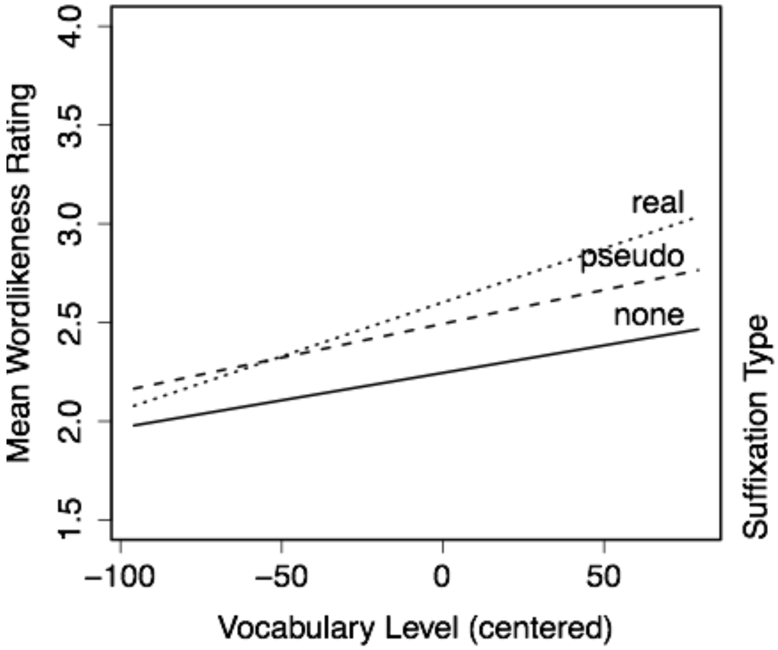

Figure 7. Interaction of vocabulary level with suffixation as captured in Decomposition models for length 7. Matches to real suffixes and words of English have a greater positive effect on ratings by people who know more words.

The positive effect of biphone score on wordlikeness is significantly higher when a partial suffix parse is present for length 7. This effect presents an interesting contrast to the interaction between triphone score in the Length 5B model; here the positive effect of triphone score is significantly reduced when a partial suffix parse is present. These effects are contrasted in Figure 8. The category of 'full' suffixations is omitted from Figure 8a because there are only 4 such forms, of which two were probably mis-parsed. The category of 'full' suffixations is omitted from Figure 8b because all such items also had analyses as compounds, and as a result the effect size distinguishing the 'partial' items from the 'full' items is extremely small. Note that the lowest biphone scores for length 7 are lower than the lowest triphone scores for length 5, and the ratings reflect this fact.

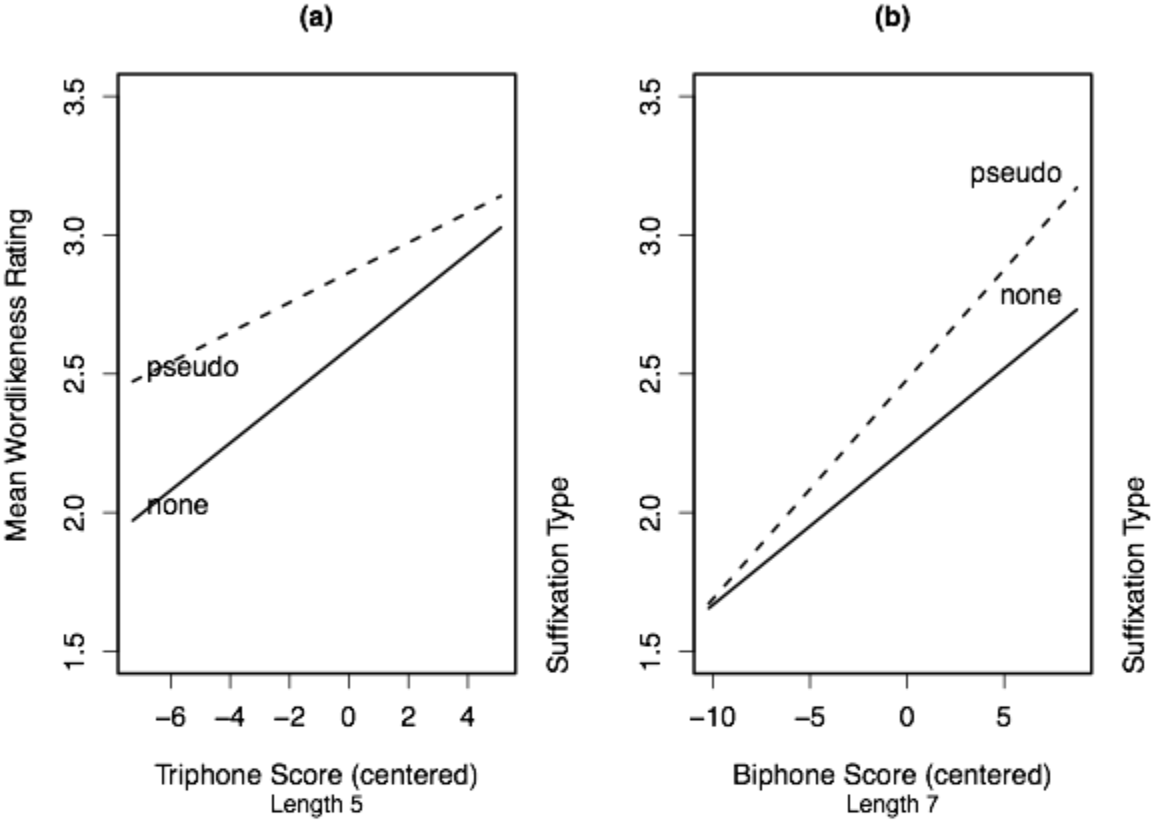

Figure 8. Interactions of phonotactic score with suffixation. (a) triphone scores for pseudo-suffixed and non-suffixed items of length 5. (b) biphone scores for pseudo-suffixed and non-suffixed items of length 7.

7. Discussion

The wordlikeness results in the Replication models replicate effects for biphone and triphone likelihood, vocabulary level, word length, and lexical neighborhood. In particular, biphone and triphone scores both make significant contributions to predicting wordlikeness judgments of English-based pseudowords over the varying lengths and wide continuum of phonotactic probabilities provided in PseudoLex. These results suggest that speakers can make judgments using more detailed phonotactic knowledge (triphone statistics), while also using more abstract biphone knowledge. In addition, when scores based on orthotactic probabilities were substituted for those based on phonotactic probabilities in the Replication model, the orthotactic and phonotactic scores were shown to provide effectively the same information for the PseudoLex inventory.

Statistical analysis of lexical similarity in the form of potential morphological parse reveals highly significant effects on wordlikeness. Because pseudowords have no established meanings, and because many of the parses do not conform to the syntactic and semantic constraints of English morphology, the benefit from such parses supports the morpho-orthographic account of lexical processing with partial parsing, as suggested by priming research such as Beyersmann et al. (2015). We have shown increased wordlikeness ratings related to partial morpho-orthographic segmentation, even if phonotactic likelihood and other factors are controlled. Our experiment demonstrates that this effect obtains not only in the on-line processing tasks explored by previous researchers, but in wordlikeness judgments. Deeper analysis of course becomes possible as words become well-learned and acquire fully elaborated representations. More generally, subword sequences that correspond to real morphemes improve wordlikeness because they suggest associations with real words that go beyond mere phonological resemblance. The result that the wordlikeness benefit is greater for pseudocompounds than for pseudosuffixed forms follows from the fact that full word matches generally represent a more substantial degree of similarity than subword matches. This possibility is argued by Grainger & Beyersmann (2017), who present a model in which people prioritize finding full word matches aligned to either the left or right edge of a stimulus. The pseudosuffixed forms in our dataset can be items consisting of a non-stem and a suffix (*thrafium*), meaning that morpho-orthographic segmentation of these items must proceed beyond stem-finding to a secondary phase.

While the contributions of phonotactic probabilities appear largely independent in our analysis, there were two significant interactions involving the suffixation factor. In the Length 7B model, higher biphone scores increased the positive effect of having a suffix parse (see 'biphone:suffix_partial'). This may indicate that it is difficult for a suffix alone to redeem the poor phonotactics of a poor stem, particularly as the stem would be notably longer than the suffix for items of length 7. However, the Length 5B model shows the reverse pattern: higher triphone score reduced the positive effect of a suffix parse (see 'triphone: suffix_partial'). It is possible that these opposite patterns come about because triphones are capturing many of the suffixes coded in this analysis, creating a redundancy.

A traditional lexical neighborhood metric has limited efficacy in predicting wordlikeness in PseudoLex, due to the fact that most of the 8400 words have no lexical neighbors. While the presence of lexical neighbors was a significant factor across all lengths, it was necessary to rely on a Boolean version of the factor because only 21% of items had any neighbors at all. Although lexical neighborhood density is a strong predictor of wordlikeness for short words, pseudomorphology has also emerged from our study as a more powerful way of looking at resemblances to pre-existing words, when a more natural range of word lengths is considered. This means that these two approaches to lexical similarity, as implemented in the current analysis, are complementary. Both lexical neighborhood and morphological parse are significant, but the significance and effect sizes are different at different lengths. Lexical neighborhood effects are most useful for shorter words (those most likely to have neighbors), while parsing is relatively more useful for longer words (those more likely to contain recognizable morphemes).

In addition to properties intrinsic to the stimuli, individual participant differences are important in predicting wordlikeness ratings. The main effect of vocabulary level is a replication of Frisch & Brea-Spahn (2010). For items with the same phonotactic likelihood, participants with higher vocabulary levels gave higher wordlikeness ratings. Vocabulary level was also shown to modulate other factors in the final set of models: lexical neighborhood, compound parsing, and suffix parsing. In both the Length 4A and 4B models, where the effect of lexical neighborhood is most important, the positive effect of lexical neighbors is relatively small for participants with larger vocabularies. This pattern may result from high-vocabulary individuals having access to a greater variety of wordlikeness factors (richer phonotactics, larger inventory of morphemes for decomposition), which could de-emphasize lexical neighborhood effects. In contrast, a larger vocabulary seems to enhance the ability to decompose potentially complex forms. For the longer words (in the Length 6B and Length 7B models), the positive effect of both compound parse (Length 6B and Length 7B) and suffix parse (Length 7B only) increase as vocabulary increases. We may see this pattern because having a larger vocabulary means more known morphemes for decomposition. The pattern could also occur because high-vocabulary individuals are skilled both at decomposing new words and at generalizing word formation patterns, creating a system of positive feedback. This possibility matches the findings of Beyersmann et al. (2015) that high-proficiency speakers showed more priming from stimuli containing embedded stems without complete morphological parses.

8. Conclusion

This study provides evidence that people process novel words using their morphological knowledge, in addition to lexical and phonological statistics. The Replication models show the expected wordlikeness effects of phonotactic likelihood, word length, lexical neighborhood, and subject vocabulary size, over broad ranges of these factors; and the Decomposition models demonstrate significant positive wordlikeness effects of suffixation and compounding parses beyond the effects in the Replication models. Such effects resemble the morpho-orthographic segmentation effects shown in lexical decision priming literature, providing convergent evidence of morphological decomposition of novel wordforms. Specifically, priming experiments show that lexical decision reaction times are shorter when a morpho-orthographic relationship obtains between a prime and a target; we suggest that this pattern is related to the current finding that wordlikeness ratings are higher for pseudowords that are morpho-orthographically decomposable.

The existence of both shallow (morpho-orthographic) and deep (morpho-semantic) processes has previously been shown for real words and pseudowords in perception experiments. It can be interpreted as an efficient, flexible strategy for perception in noisy and variable contexts (Sanford & Graesser 2006). We can consider our results in the context of two competing accounts of lexical perception: the multiple-route-type approach of Grainger & Beyersmann (2017), and the Naive Discriminative Learning-based approach (NDL) described in Milin, Feldman, Ramscar, Hendrix, & Baayen (2017). Grainger and Beyersmann argue for a version of morpho-orthographic segmentation in which the first step is not affix-stripping, but instead stem-finding. Specifically, edge-aligned embedded word activation precedes the activation of affixes in reading. This stipulation provides a mechanism for the priming evidence shown for stimuli like *cornea* as well as the compound-type pseudowords in the present study (*churcharou*, *affreap*). After this phase, the activation of affix-like string takes place; this encompasses our results for suffixation-type pseudowords (*thrafium*, *surpitual*). If a stem-first process is similarly underlying wordlikeness judgments for our pseudoword stimuli, items with compound-type parses should receive a larger wordlikeness boost than those with suffixation-type parses (i.e., items lacking real word stems); such a pattern is suggested in the current data.

In contrast, the Milin et al. (2017) model does not make use of morphemes at all. Boundary-sensitive trigram units (i.e., trigraphs with a symbol for word boundaries) are statistically associated with semantic units (*lexomes*). In this model, an NDL method is used to learn the associations between spellings and meanings, and within meanings (i.e., semantically-related words, synonyms, etc.), with the result that speakers are able to connect similar words (e.g., *work* and *worker*) without representing a traditional morphological parse (e.g., *worker* = *work* + *-er*). Milin et al. show that the resulting model is able to accurately predict the size of priming effects for *corner*-type stimuli (fully decomposable) and *cornea*-type stimuli (partially decomposable). Under the further assumption that these predictions of lexical decision latency are related to wordlikeness judgments of pseudowords, it seems that this approach lends itself to explaining the wordlikeness effects in the current study of apparent morphology in fully decomposable and partially decomposable pseudowords. However, it is not clear how it would predict improved wordlikeness for fully decomposable pseudowords over partially decomposable pseudowords. The contribution of affixes within the NDL-based model is less important than that of embedded words, insofar as affixes are shorter on the average than embedded words. As with the Grainger and Beyersmann model, this could go towards explaining why compound-type pseudomorphology yields a greater wordlikeness improvement than suffixation-type pseudomorphology in the current study.

While these two accounts differ widely in their underlying mechanisms, both are compatible with our findings that pseudowords are rated more wordlike when they appear to contain morphemes, even when they are not exhaustively decomposable. The decomposition of pseudowords is an important component of our understanding of lexical innovation and morphological productivity. All new words were once pseudowords, and it appears that more wordlike pseudowords are more likely to become new words. The enhanced acceptability of partially decomposable pseudowords should give them an advantage in being added to the lexicon over phonotactically legal words of comparable length. While the evidence in the current data is limited, the greater advantage for fully decomposable pseudowords suggests that pseudowords containing ‘cranberry morphemes’ (words with a partial parse), though viable, should be less readily assimilated. Together, these patterns mean that amongst new word candidates, there is a competitive advantage for forms that include existing morphemes.

The morphology advantage may also be demonstrated in the pseudomorphological transformations that take place during borrowings and in folk etymologies. For example, the morpheme *fish* was included in the word *crayfish* when borrowed from the French word *crevis*, which makes sense for a water-dwelling seafood animal. Morpho-semantic factors can also influence existing words in the lexicon, as illustrated by a form of reanalysis called ‘eggcorns’. This term was coined in 2003 to describe the reanalysis of less familiar words into similar-sounding novel words with appropriate semantics (Liberman 2003); the morphologically-complex word *eggcorn* derives from a reanalysis of *acorn.* The morphology advantage demonstrated in our results points to a step in the process of language innovation where morphological complexity can be favored, even without a clear semantic basis.

Word frequency and word familiarity are generally correlated, but it is known that this correlation breaks down for longer, morphologically complex words. As discussed in Section 2.3. existing words with a shallow compound parse seem more familiar than their token frequency would otherwise imply. While a more thorough experimental evaluation is clearly needed, it seems plausible that shallow parsing might in general enhance the apparent familiarity of rare or novel word forms. In our randomly generated pseudoword set, the prevalence of morphological parses increases as item length increases. The influence of morphology in turn increases with length, and the relative wordlikeness contributions of phonotactics and especially lexical neighborhood are reduced. This pattern means that longer words can be supported in the lexicon by participation in morphological families of related words. The familiarity effect for real complex words may be so strong that people rate completely novel words that have an apparent morphological analysis as familiar. This dynamic predicts a lexicon containing more long words than phonotactic scores alone would predict; but the longer words should be clustered in morphologically related constellations rather than evenly spread through the phonotactic space.

References

Baayen, R. H., R. Piepenbrock, & L. Gulikers. 1995. The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Bailey, Todd M., & Ulricke Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods?. *Journal of Memory and Language* 44(4), 568-591.

Bates, Douglas, Martin Mächler, Benjamin M. Bolker, & Steven C. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1-48.

Beyersmann, Elisabeth, Séverine Casalis, Johannes C. Ziegler, & Jonathan Grainger. 2015. Language proficiency and morpho-orthographic segmentation. *Psychonomic Bulletin & Review* 22(4), 1054-1061.

Beyersmann, Elisabeth, Eddy Cavalli, Séverine Casalis & Pascale Colé. 2016. Embedded stem priming effects in prefixed and suffixed pseudowords. *Scientific Studies of Reading* 20(3), 220-230.

Bird, Steven, Ewan Klein, & Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Bybee, Joan L. 1988. Morphology as lexical organization. *Theoretical Morphology.* 119-141.

Caramazza, Alfonso, Alessandro Laudanna, & Cristina Romani. 1988. Lexical access and inflectional morphology. *Cognition*, 28(3), 297-332.

Coleman, John, & Janet Pierrehumbert. 1997. Stochastic Phonological Grammars and Acceptability. *3rd Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, 12 July 1997. Association for Computational Linguistics, Somerset, NJ. 49-56.

Coltheart, Max, Eileen Davelaar, Jon Torfi Jonasson, & Derek Besner. 1977. Access to the internal lexicon. In S. Dornic (Ed), *Attention and Performance VI* (535-555). Hillsdale, NJ: Erlbaum.

Connine, Cynthia M., John Mullennix, Eve Shernoff, & Jennifer Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(6), 1084-1096.

Daland, Robert. 2015. Long words in maximum entropy phonotactic grammars. *Phonology* 32(3), 353-383.

Daland, Robert, Andrea D. Sims, & Janet Pierrehumbert. (2007). Much ado about nothing: A social network model of Russian paradigmatic gaps. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics* (pp. 936-943).

Davies, Mark. 2008-. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Edwards, Jan, Mary E. Beckman, & Benjamin Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research* 47(2), 421.

Frisch, Stefan A., & Maria R. Brea-Spahn. 2010. Metalinguistic judgments of phonotactics by monolinguals and bilinguals. *Laboratory Phonology* 1(2), 345-360.

Frisch, Stefan A., Nathan R. Large, & David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4), 481-496.

Frisch, Stefan A., Nathan R. Large, Bushra Zawaydeh, & David B. Pisoni. 2001. Emergent phonotactic generalizations in English and Arabic. *Typological Studies in Language* 45, 159-180.

Gosling, Samuel D., Carson J. Sandy, Oliver P. John, & Jeff Potter. 2010. Wired but not weird: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences* 33(2-3), 94-95.

Grainger, Jonathan. 1990. Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language* 29(2), 228-244.

Grainger, Jonathan, & Elisabeth Beyersmann. 2017. Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In *Psychology of Learning and Motivation* (Vol. 67, pp. 285-317). Academic Press.

Hahn, Stefan, Paul Vozila, & Maximilian Bisani. 2012, September. Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks. In *Thirteenth Annual Conference of the International Speech Communication Association* (pp. 2537-2540). Portland, OR: International Speech Communication Association.

Hahn, Ulrike, & Todd M. Bailey. 2005. What makes words sound similar?. *Cognition* 97(3), 227-267.

Hasenäcker, Jana, Elisabeth Beyersmann, & Sascha Schroeder. 2016. Masked morphological priming in German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology* 7, 929.

Havas, Viktória, Otto Waris, Lucía Vaquero, Antoni Rodríguez-Fornells, & Matti Laine. Morphological learning in a novel language: A cross-language comparison. *The Quarterly Journal of Experimental Psychology* 68(7), 1426-1441.

Hay, Jennifer, Janet Pierrehumbert, & Mary Beckman. 2004. Speech Perception, Well-Formedness, and the Statistics of the Lexicon. *Papers in Laboratory Phonology VI*, Cambridge University Press, Cambridge, UK, 58-74.

Henrich, Joseph, Steven J. Heine, & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 61- 83.

Heller, Jordana. 2014. *Contextual Constraints on Phonological Activation During Sentence Production*. (Doctoral dissertation, Northwestern University).

Jurafsky, Daniel, & James H. Martin. 2000. *Speech and language processing*. Prentice-Hall.

Kager, René, & Joe Pater. 2012. Phonotactics as phonology: knowledge of a complex restriction in Dutch. *Phonology* 29(01), 81-111.

Kapatsinski, Vsevolod. 2006. Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Speech Research Lab Progress Report* 27, 133-152.

Kapatsinski, Vsvolod, & Lamia Johnston. 2010. Investigating phonotactics using xenolinguistics: A novel word-picture matching paradigm. *Proceedings of the Annual Meeting of the Cognitive Science Society* 32.

Kuperman, Victor, Raymond Bertram, & R. Harald Baayen. 2008. Morphological Dynamics in Compound Processing. *Language and Cognitive Processes* 23(7-8), 1089-1132.

Libben, Gary, Bruce L. Derwing, & Roberto G. De Almeida. 1999. Ambiguous Novel Compounds and Models of Morphological Parsing. *Brain and Language* 68, 378-386.

Libben, Gary, Martha Gibson, Yeo Bom Yoon, & Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84, 50-64.

Liberman, Mark. 2003. Egg corns: folk etymology, malapropism, mondegreen. *Language Log*. Available online at http://itre.cis.upenn.edu/~myl/languagelog/archives/000018.html.

Limpert, Eckhard, Werner A. Stahel, & Markus Abbt. 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51, 341-352.

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

Luce, Paul A., David B. Pisoni, & Steven D. Goldinger. 1990. Similarity neighborhoods of spoken words. In G. T. M. Altmann (ed.) *Cognitive models of speech processing: psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press. 105–121.

Manning, Christopher D., & Hinrich Schütze. 1999. *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT Press.

Marian, Viorica, James Bartolotti, Sarah Chabal, & Anthony Shook. 2012. CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS One* 7(8), e43230.

Marslen-Wilson, William, Lorraine K. Tyler, Rachelle Waksler, & Lianne Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101(1), 3.

Milin, Petar, Laurie Beth Feldman, Michael Ramscar, Peter Hendrix, & R. Harald Baayen. (2017). Discrimination in lexical decision. *PloS One* 12(2), e0171935.

Morris, Joanna, James H. Porter, Jonathan Grainger, & Phillip J. Holcomb. 2011. Effects of lexical status and morphological complexity in masked priming: An ERP study. *Language and Cognitive Processes* 26(4-6), 558-599.

Needle, Jeremy M., & Janet B. Pierrehumbert. In press. Gendered Associations of English Morphology. *Journal of Laboratory Phonology*.

Novak, Josef R., Nobuaki Minematsu, & Keikichi Hirose. 2012, July. WFST-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding. In *10th International Workshop on Finite State Methods and Natural Language Processing* (p. 45).

Novak, Josef, Dong Yang, Nobuaki Minematsu, & Keikichi Hirose. 2011. *Initial and Evaluations of an Open Source WFST-based Phoneticizer*. The University of Tokyo, Tokyo Institute of Technology.

Nusbaum, Howard C., David B. Pisoni, & Christopher K. Davis. 1984. Sizing up the Hoosier mental lexicon. *Research on spoken language processing report no* 10, University of Indiana, 357-76.

Paice, Chris D. 1990. Another Stemmer. *ACM SIGIR Forum* 24.3, 56-61.

Pierrehumbert, Janet B. 2003. Probabilistic Phonology: Discrimination and Robustness. In R. Bod, J. Hay and S. Jannedy (eds.), *Probability Theory in Linguistics*. The MIT Press, Cambridge, MA, 177-228.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

Rácz, Péter, Janet B. Pierrehumbert, Jennifer B. Hay, & Viktória Papp. 2015. Morphological Emergence. In Brian MacWhinney and William O'Grady (eds.), *The Handbook of Language Emergence*. Wiley Blackwell, 123-146.

Rastle, Kathleen, Matthew H. Davis, & Boris New. 2004. The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review* 11(6), 1090-1098.

Richtsmeier, Peter T. 2011. Word-types not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* 2, 157-183.

Sanford, Anthony J., & Arthur C. Graesser. 2006. Shallow processing and underspecification. *Discourse Processes* 42(2), 99-108.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, & Andrew Y. Ng. 2008, October. Cheap and fast--but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Association for Computational Linguistics.

Storkel, Holly L. 2004. Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics* 25(02), 201-221.

Storkel, Holly L., Jonna Armbrüster, & Tiffany P. Hogan. 2006. Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research* 49(6), 1175-1192.

Stump, Gregory T. 2018. Some sources of apparent gaps in derivational paradigms. *Morphology*.

Taft, Marcus. 2004. Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology* 57A(4), 745-765.

Taft, Marcus, & Kenneth I. Forster. 1975. Lexical Storage and Retrieval of Prefixed Words. *Journal of Verbal Learning and Verbal Behavior* 14, 638-647.

Vaden, Kenneth I., H. R. Halpin, & Gregory S. Hickok. 2009. *Irvine Phonotactic Online Dictionary, Version 2.0*. [Data file]. Available from http://www.iphod.com.

Vitevitch, Michael S., & Paul A. Luce. 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language* 40, 374–408.

Vitevitch, Michael S., & Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3), 481-487.

Vitevitch, Michael S., Paul A. Luce, David B. Pisoni, & Edward T. Auer. 1999. Phonotactics, Neighborhood Activation, and Lexical Access for Spoken Words. *Brain and Language* 68(1-2), 306-311.

Vitevitch, Michael S., & Eva Rodríguez. 2005. Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders* 3(1), 64-73.

Vitevitch, Michael S., Melissa K. Stamer, & Joan A. Sereno. 2008. Word length and lexical competition: Longer is the same as shorter. *Language and Speech* 51(4), 361-383.

Warriner, Amy Beth, Victor Kuperman, & Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4), 1191-1207.

Wurm, Lee H., Annmarie Cano, & Diana A. Barenboym. 2011. Ratings gathered on-line versus in person. *The Mental Lexicon* 6(2), 325-350.

Appendix A
Nonword Instructions

You will be shown a series of made-up words, one at a time. Pronounce each word you see out loud, as best you can.

Your task is to rate each word for how 'English-like' it is: how much it sounds like a normal word of English that you simply never learned before.

Focus on how the word sounds, not on the spelling.

Using the five labeled buttons below the word, you will give each word a rating from 1 to 5:

(5) means the word is a perfectly good, normal-sounding English word;

(1) means the word is awful or impossible-sounding as a word of English.

Here is an example:

[Insert Appendix A Image 1 here.]

After you've rated the word, press the 'Next' button to continue. You will be told when you've finished. When you finish rating all the made-up words, there will be new instructions for the next part of the HIT.

To take this HIT, you have to be a native speaker of English, 18 years or older.

Please be aware that some of our tasks are incompatible with earlier ones. If you have completed a previous task that this one is incompatible with, you will not be able to take this HIT.

We monitor our results to make sure that participants are attentive. If you do not give the task enough attention, you risk being excluded from taking any of our future HITs.

You may review these instructions. When you are ready, please press the 'Next' button to begin.

Appendix B
Vocabulary Instructions

In the last part of this task, you will see seventy ‘words’, one after the other. Some are real words of English, while some are made-up nonwords.

Your task is to indicate your familiarity with each word on a 1-5 scale.

The scale is from least familiar (1) to most familiar (5), and should be applied as follows:

1 = totally unknown; I have never seen or heard this word.

2 = unfamiliar; I may have seen or heard this word, but I don’t know what it means, and I would not use this word.

3 = somewhat familiar; I have seen or heard this word, I have some idea of what it means but I am not completely sure, and I would probably not use this word.

4 = familiar; I have definitely seen or heard this word, I think I know what it means, and I would use this word.

5 = very familiar; I have definitely seen or heard this word, I am sure that I know what it means, and I would be very comfortable using the word myself.

Please be as honest as you can in your responses.

Work as quickly as you can without sacrificing accuracy.

Here is the first Example Question:

[Insert Appendix B Image 1 here.]

This is a very familiar real word of English, so the correct choice is (5).

When you are ready, click the (Next) button to continue.”

Here is the second Example Question:

[Insert Appendix B Image 2 here.]

This is a made-up nonword, so the correct choice is (1).

When you are ready, click the (Next) button to begin.

Appendix C
Phonetisaurus Description

Phonetisaurus uses WFSTs (weighted finite state transducers) in a modified EM-driven alignment algorithm (Novak, Minematsu, & Hirose 2012). This approach finds the optimal set of correspondences between strings of letters and phonemes. Phonetisaurus takes a user-supplied pronunciation dictionary input (i.e., a word list in paired phonemic and orthographic forms). This list is used to fit an optimal model of G2P mappings. Like other high-performance G2P tools, Phonetisaurus makes use of joint n-gram 'graphone' units, which define mappings between orthographic n-grams and phonemic n-grams (e.g., grapheme *ee* maps to phoneme /i/). Note that the graphone approach is inherently bidirectional, so it can be used for both G2P and P2G. Graphone mappings may be simple (e.g., in /kæt/ <–> *cat*, an orthographic 1-gram *t* is mapped to a phonemic 1-gram /t/), or complex, with the graphones consisting of different size n-grams (in *tax*, *x* –> /ks/, or in *fish*, *sh* –> /ʃ/).

Mappings are frequently not unique, so that either a grapheme or a phoneme may correspond to multiple different counterparts, depending on context and variation. For example, given *cat* (/kæt/) and *kit* (/kɪt/), /k/ –> *c* or *k*; given *cats* and *dogs*, *s* –> /s/ or /z/. In the the probabilistic Phonetisaurus graphone model, these multiple mappings are weighted based on the input corpus. Graphone mappings may be ambiguous in a given word; e.g., a null mapping so that a grapheme can be silent (in *knight*, *k* –> null), could also be learned as the mapping *kn* –> /n/. Phonetisaurus considers these alternatives to build an optimal model for the input corpus overall. The n-gram representational structure used is context-sensitive, in the same way that phonotactic triphones capture additional structure over biphones. Phonetisaurus uses a multiple n-gram model (up to 8-grams), to capture idiosyncratic spellings for longer strings and (pseudo)morphemes (e.g., *-tion* in *nation*, or *-tuous* in *fatuous*).

The Phonetisaurus output is a ranked set of candidates: e.g., /tæks/ might yield in descending order *tacks*, *tax*, *taks*, *tacs*. Note the two worst examples are not matches to real words, but they are pronounceable and encode the intended phonemes. It is possible for the orthographic representations to be ambiguous in pronunciation. This ambiguity, which is unavoidable for a natural language like English, is particularly dangerous for pseudowords; by definition, subjects will have no previous experience with these exact words to guide their pronunciation. This presents a problem for linguistic experiments depending on the control of specific phonological characteristics in the stimuli. To address this issue, the orthographic representations of pseudoword items were converted back to phonemic representation using the same trained Phonetisaurus model. Any items for which the resulting phonemic output did not match the original phonemic input were excluded from use. This can occur both due to spelling system ambiguity, and P2G system errors; instability is particularly expected for the unpronounceable nonwords. This mapping stability filter gives confidence that the intended pronunciation for stimuli is the most likely one for the orthographic form presented. Table C-1 provides examples of errors that the stability filter removes. In an initial batch of 120000 candidate items, 38073 items passed the filter (32%).

Phonemic Input	Graphemic Form	Phonemic Output
/ɛvrəʤuɚ/	evrdu	/vdu/
/dɛljuɚs/	deuous	/fjuɚs/
/hɔdɛld/	hordeld	/hɔd/
/ɔɪŋkɔps/	oincorps	/ɔɪnkɔ/
/jjnkɔR/	jaruk	/ʤɑruk/

Table C-1. Examples of P2G/G2P instability results.

Appendix D
Shared Corpus Format

The PseudoLex corpus of pseudowords and nonwords will be made available as a resource for further research: ‘pseudoLex_share1.csv’. The corpus contains 201,600 wordlikeness ratings of 8400 stimuli by 1440 subjects (24 ratings per stimulus). This data is shared in CSV tidy format with the following fields: “subjID” (arbitrary subject ID number), “gender” (subject-reported ‘Male’, ‘Female’, or ‘Decline to answer’), “birthYear” (subject-reported year of birth, where 0 indicates ‘declined to answer’), “vocabLevel” (the score of the subject on the vocabulary test), “cmu” (the stimulus in CMU phonemic representation), “disc” (stimulus in DISC/CELEX phonemic representation), “ortho” (stimulus in orthographic representation), “length” (stimulus length in phones), “rating” (Likert 1-5 rating of the item by the subject), “uniScore” (stimulus uniphone log phonotactic probability), “biScore” (stimulus cumulative log transitional biphone phonotactic probability), and “triScore” (stimulus cumulative log transitional triphone phonotactic probability). When an item is illegal under the biphone or triphone phonotactic grammars, ‘biScore’ or ‘triScore’ is listed as ‘0’ in this data file. For detailed descriptions of these fields and the data-gathering methods, see Sections 4 and 5.

Appendix E

Extended Model Comparison Statistics

Replication Model Factors	Length 4A		Length 5A		Length 6A		Length 7A	
	X^2	p	X^2	p	X^2	p	X^2	p
biphone	21.9	<0.001	54.0	<0.001	83.0	<0.001	125.2	<0.001
triphone	105.1	<0.001	78.3	<0.001	118.5	<0.001	100.9	<0.001
vocabulary	53.6	<0.001	58.8	<0.001	66.9	<0.001	80.0	<0.001
neighbors	245.9	<0.001	185.4	<0.001	89.2	<0.001	5.9	0.015
neighbors:vocabulary	10.5	0.001	-	-	-	-	-	-
biphone:vocabulary	-	-	-	-	-	-	6.2	0.013

Table E-1. Drop-1 model comparison statistics for Replication models. Unavailable comparisons indicated by ‘-’.

Decomposition Model Factors	Length 4B		Length 5B		Length 6B		Length 7B	
	X^2	*p*	X^2	*p*	X^2	*p*	X^2	*p*
biphone	26.2	<0.001	61.8	<0.001	89.8	<0.001	-	-
triphone	106.8	<0.001	-	-	103.4	<0.001	74.3	<0.001
vocabulary	53.7	<0.001	58.8	<0.001	-	-	-	-
neighbors	304.1	<0.001	-	-	84.8	<0.001	8.2	0.004
compound	120.9	<0.001	-	-	189.7	<0.001	290.6	<0.001
suffix	-	-	93.8	<0.001	73.8	<0.001	76.2	<0.001
neighbors:vocabulary	10.6	0.001	-	-	-	-	-	-
neighbors:compound	-	-	17.3	<0.001	-	-	-	-
biphone:suffix	-	-	-	-	-	-	-	-
triphone:suffix	-	-	11.7	0.003	-	-	-	-
vocabulary:compound	-	-	-	-	17.7	0.000	19.8	<0.001
vocabulary:suffix	-	-	-	-	-	-	7.3	0.026
biphone	26.2	<0.001	61.8	<0.001	89.8	<0.001	-	-
triphone	106.8	<0.001	-	-	103.4	<0.001	74.3	<0.001
vocabulary	53.7	<0.001	58.8	<0.001	-	-	-	-
neighbors	304.1	<0.001	-	-	84.8	<0.001	8.2	0.004
compound	120.9	<0.001	-	-	189.7	<0.001	290.6	<0.001
suffix	-	-	93.8	<0.001	73.8	<0.001	76.2	<0.001
neighbors:vocabulary	10.6	0.001	-	-	-	-	-	-

Table E-2. Drop-1 model comparison statistics for Decomposition models. Unavailable comparisons indicated by ‘-’.