NORTHWESTERN UNIVERSITY

Gradient Typicality and Indexical Associations in Morphology

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Jeremy Michael Needle

EVANSTON, ILLINOIS

September 2018

Abstract

This dissertation focuses on the topic of pseudowords and how speakers pseudoword processing relates to that of real words. Three main lines of inquiry are pursued with respect to pseudowords and real words: mechanisms of gradient well-formedness, theories of morphological decomposition, and indexical associations for morphemes in complex words. It argues for an integrated model which considers real words and novel words using common mechanisms, and which takes into account both morphological structure and indexical information.

Chapter 3 expands on studies of pseudoword wordlikeness by collecting wordlikeness judgments for a large corpus of pseudowords which comprehensively sample the space of phonotactic probability for pseudowords that are both short and long. The positive effect of phonotactic likelihood is replicated over the whole domain of likely and unlikely forms, and the realistic limitations of simple neighborhood density measures are shown. Post-hoc analysis also suggests that participants perceived apparent morphology in the pseudowords, and gave such items higher wordlikeness ratings.

Chapter 4 demonstrates that participants also give gradient typicality judgments for real words, in contrast to the predictions of some traditional and current theories of well-formedness. In this new analysis of data from Bailey and Hahn (2001), significant variation is observed in the judgments of both real words and pseudowords, and this variation correlates with similar factors for both categories.

Chapter 5 follows upon the morphological findings in Chapter 2 by eliciting explicit morphological decompositions from participants for complex real words and apparently-complex

pseudowords. While it is unclear in Chapter 2 whether participants are actually aware of morpheme strings per se in the pseudoword stimuli, it is shown in Chapter 4 that participant accuracy in decomposing real and pseudowords exceeds the baseline levels derived from chance or from morphologically-unaware phonotactic statistics.

With previous chapters establishing that wordlikeness judgments are influenced by different aspects of similarity within the lexicon (i.e., phonotactic probability, neighborhood density, morphology), Chapter 5 investigates indexical effects of morphology on complex pseudowords. Much of lexical innovation in English involves morphology, and Chapter 5 finds that the gendered experience people have with morphemes influences their associations for novel words containing those morphemes.

Chapter 6 summarizes the findings for real words and pseudowords, considers their relation to theories of lexical processing and morphological decomposition, discusses consequences for mechanisms of language change, and proposes steps for future research.

Acknowledgments

First, I would like to thank my advisor, Janet Pierrehumbert, who has invested an extraordinary amount of time and support into my academic training. Her enthusiastic scientific curiosity has inspired every step of this project, and her patience and guidance have made completing it possible. I am also indebted to my committee members, Matt Goldrick, Jen Hay, and Klinton Bicknell, and to my teachers, for sharing their knowledge and time. I would like to thank the many people who have contributed to this project over these several years in the form of listening, feedback, and critique. Finally, this entire undertaking would not have been possible without the support of my family, friends, graduate student cohort, and labmates: thank you for all being in this together with me.

Preface

Information is included here about the publication status and authorship for each dissertation chapter:

Chapter 2, “Shallow Morphological Processing in Pseudowords”, has been accepted for the edited volume of proceedings of the 2017 workshop on “Morphological Typology and Linguistic Cognition” (Cambridge University Press). The authors for Chapter 2 are Jeremy M. Needle (Northwestern University), Janet B. Pierrehumbert (University of Oxford), and Jennifer B. Hay (University of Canterbury).

The authors for Chapter 3, “Gradient Typicality Judgments of English Words”, are Jeremy M. Needle (Northwestern University), Janet B. Pierrehumbert (University of Oxford), Todd M. Bailey (Cardiff University), and Ulricke Hahn (Birkbeck, University of London).

The authors for Chapter 4, “Explicit Morphological Decomposition of Complex English Pseudowords”, are Jeremy M. Needle (Northwestern University) and Janet B. Pierrehumbert (University of Oxford).

Chapter 5, “Gendered Associations of English Morphology”, has been accepted for the special issue of the Journal of Laboratory Phonology on “Abstraction, Diversity and Speech Dynamics”. The authors for Chapter 5 are Jeremy M. Needle (Northwestern University) and Janet B. Pierrehumbert (University of Oxford).

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.

Table of Contents

List of Tables, Illustrations, Figures, or Graphs

1. Introduction

This dissertation focuses on the topic of pseudowords and how speakers pseudoword processing relates to that of real words. Three main lines of inquiry are pursued with respect to pseudowords and real words: mechanisms of gradient well-formedness, theories of morphological decomposition, and indexical associations for morphemes in complex words.

Gradient well-formedness has been previously demonstrated for pseudowords, though many experiment sets of pseudoword items are limited to short and even monosyllabic forms (e.g., Bailey & Hahn, 2001). In addition, these sets typically represent a small portion of the space of possible forms, with a bias toward the most wordlike pseudowords. This dissertation begins with the collection of wordlikeness judgments for a large set of pseudowords which systematically samples the full space of phonotactic probability for short and longer forms (Chapter 2). It then presents evidence for gradient typicality in morphologically simple real words, and reviews how models of wordlikeness can adequately capture the gradience seen for both real words and pseudowords (Chapter 3).

Different models of morphological decomposition may require semantic transparency, exhaustive parsing, or only partial orthographic similarity (cf. Marslen-Wilson, Tyler, Waksler, & Older, 1994; Taft & Forster, 1975; Rastle, Davis, & New, 2004). Post-hoc analysis in Chapter 2 suggests that participants recognized apparent morphemes in a large subset of the random-generated pseudowords, and that their wordlikeness judgments were increased for those items; these findings raise the question of whether decomposition is involved in pseudowords processing. Chapter 4 uses a new experimental paradigm to demonstrate that participants are highly successful in giving explicit decomposition responses for complex pseudowords partially constructed of real morphemes.

If morphological decomposition influences pseudoword processing, there could be important consequences for models of lexical innovation. Chapter 5 builds on the experimental paradigm for decomposition to look for effects of indexical associations (here, gender associations) of known morphemes on pseudowords. Indexical associations have been widely documented for whole words (e.g., R. Lakoff, 1973) and for phonetic patterns (e.g., Labov, 2001; Eckert, 2008), but Chapter 5 provides evidence suggesting that similar effects are possible for derivational morphemes per se.

Together, the results for gradient well-formedness, morphological decomposition, and indexical associations make important connections to models of linguistic change, and of language diversity within connected populations. First, they bridge studies of well-formedness and phonotactic learning with lexical processing and psycholinguistics. Second, they enhance our understanding morphological processing and the cognitive architecture underlying theories of morphological productivity. Third, they suggest new avenues for cognitive processing models of socio-indexical features. Finally, they suggest a unified idea of variation and change, in the perspective of lexical processing and psycholinguistics. Lexical processing and perception of words is a critical step in the cycle of language change: new words are created out of the space of possible words, and then these new words must be spread to other speakers; speakers perceive and adopt new words, according to their biases and experiences. In this way, models of lexical processing are the lynchpin to language change. For the improvement of those models, this dissertation identifies gaps between experimental results and existing theories, while also suggesting aspects of these models which can be unified in ways that mutually benefit the understanding of those factors.

1.1. Conventions of terminology

This dissertation focuses on the question of phonotactic well-formedness, though the concept of well-formedness can be applied to many aspects of a word, and indeed to other levels of linguistic inquiry (e.g., to syntax). The term *word* is used to describe any isolated string of letters or phonemes, regardless of lexical semantics. When words are presented as examples here, they are styled in all capital letters (e.g., WORD). To make connections between approaches to well-formedness and lexical processing models, a clear discussion of the terminology for these concepts is required at the outset. The term *well-formedness* itself pertains to the characterization of possible words in a language: well-formed words are those which are acceptable to speakers, while any other words are ill-formed. Descriptions of well-formedness are grounded in empirical observation of acceptability. For example, analysis of corpora yields positive evidence about well-formedness because all attested words must be well-formed. However, the set of well-formed words is by no means limited to attested words. It is clear that this cannot be the case, because well-formedness knowledge is critical for language learning and language change. If speakers thought that only attested words were well-formed, no new words could arise, because they would be ill-formed and unacceptable; nor could a language learner, whose personal lexicon is limited, accept and learn any unknown words.

In fact, experimental evidence shows that speakers can find both known and unknown words to be acceptable. Because the set of well-formed words includes both attested words and possible words that are not part of the lexicon, a three-way distinction in terms is used: *real* *words*, *pseudowords*, and *nonwords*. Real words (e.g., PELT) are attested and acceptable. Pseudowords (e.g., PELP) are unattested, but possible and acceptable. Nonwords (e.g., BNICK) are impossible and unacceptable. More nuanced evidence of acceptability is shown by experimental judgments of wordlikeness for words; participants are asked to rate how much a word is like words of English, understanding that a pseudoword may be wordlike even though it is not a word they recognize. The task is to compare a word to all their knowledge of real words and judge its similarity. Previous studies have demonstrated that human judgments of wordlikeness for pseudowords are not limited to the binary distinction between well-formed and ill-formed, but are gradient on this continuum: the pseudoword GWAG might be less acceptable than PELP, even though neither is totally impossible.

In order to characterize these gradient judgments of wordlikeness, studies have demonstrated a number of factors which correlate with observed wordlikeness. Two widespread approaches are phonotactic probability, and lexical neighborhood density. Though they use different perspectives, both of these can be considered as measures of overall lexical similarity based primarily on word type frequencies. Phonotactic probability methods induce a generative probabilistic grammar from the lexicon, which can generate all real words in the lexicon. The grammar can also generate pseudowords that are similar to the real words, and wordlikeness is a correlate of how likely a pseudoword is to be generated. For example, there are several words which contain PEL (as in the pseudoword PELP), but very few that contain GWA (as in GWAG), and none that contain BNI (as in the nonword BNICK). The grammar is more likely to generate PELP, which therefore is more wordlike than GWAG. In contrast, a simple lexical neighborhood density method predicts wordlikeness based on the number of real words that are very similar to a given word (e.g., words which differ by a single change of a letter or phoneme). This means that the more wordlike PELP has several neighbors (HELP, KELP, PULP, PELT, PEEP, PEP), while the less wordlike GWAG has fewer (SWAG, WAG, GAG).

Both phonotactic probability and lexical neighborhood density approaches to wordlikeness make gradient predictions for pseudowords, as exemplified above. In the same way, these models make gradient wordlikeness predictions for real words: real words may be more or less likely to be generated by a phonotactic grammar, and real words may have many or few neighbors. This reveals a terminological problem: if a participant is asked to judge how much a particular real word of English is like real words of English, they would probably respond that it is maximally wordlike—by definition. The concept more frequently used for real words is typicality: how typical of an English word is the stimulus? For the purposes of this dissertation, typicality is effectively equivalent to wordlikeness and the concept of overall lexical similarity, so these terms are used as appropriate for the experimental context. The studies in Chapters 2, 4, and 5 are concerned with pseudowords and use wordlikeness, while Chapter 3 focuses on real words and uses typicality.

The two models of word acceptability described above are quite parsimonious, primarily making use of type frequencies for real words. However, there are many other aspects of lexical information available to speakers. Some of this additional information is commonly used to elaborate both phonotactic probability and lexical neighborhood density models. For example, speakers are known to be sensitive to token frequency in addition to type frequency; people know that the word HELLO is much more common than DASTARDLY, and this knowledge could influence wordlikeness judgments. These models can incorporate this speaker knowledge to improve their accuracy. Other aspects of speaker knowledge are less commonly used; a speaker is not limited to simply being aware of how often a given word is encountered, but experiences rich information about the linguistic context. Speakers know what kind of people often use the word, in what situations, in what locations, and so on. This contextual knowledge that relates to the social use of words is indexical information, because it can provide linguistic pointers to social identity; for example, it was once the case that using profanity like DAMN was stereotypically associated with men, while women were instead associated with euphemized variants like DARN.

1.2. Outline of the dissertation

The characterization of possible words in a language is central to the study of lexical processing, as well as of lexical innovation: new words must be former pseudowords, and speakers must be able to distinguish known words from merely plausible ones. Chapter 2 expands on these studies by collecting wordlikeness judgments for a large corpus of pseudowords which comprehensively sample the space of phonotactic probability for pseudowords that are both short and long. The positive effect of phonotactic likelihood is replicated over the whole domain of likely and unlikely words, and the realistic limitations of simple neighborhood density measures are shown. Post-hoc analysis also suggests that participants perceived apparent morphology in the pseudowords, and gave such words higher wordlikeness ratings.

Chapter 3 demonstrates that participants also give gradient typicality judgments for real words, in contrast to the predictions of some traditional and current theories of well-formedness. In this new analysis of data from Bailey and Hahn (2001), significant variation is observed in the judgments of both real words and pseudowords, and this variation correlates with similar factors for both categories.

Chapter 4 follows upon the morphological findings in Chapter 2 by eliciting explicit morphological decompositions from participants for complex real words and apparently-complex pseudowords. While it is unclear in Chapter 2 whether participants are actually aware of morpheme strings per se in the pseudoword stimuli, it is shown in Chapter 4 that participant accuracy in decomposing real and pseudowords exceeds the baseline levels derived from chance or from morphologically-unaware phonotactic statistics.

With previous chapters establishing that wordlikeness judgments are influenced by different aspects of similarity within the lexicon (i.e., phonotactic probability, neighborhood density, morphology), Chapter 5 investigates indexical effects of morphology on complex pseudowords. Much of lexical innovation in English involves morphology, and Chapter 5 finds that the gendered experience people have with morphemes influences their associations for novel words containing those morphemes.

Chapter 6 summarizes the findings for real words and pseudowords, considers their relation to theories of lexical processing and morphological decomposition, discusses consequences for mechanisms of language change, and proposes steps for future research.

2. Shallow Morphological Processing in Pseudowords

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. The distinction between pseudowords and nonwords is a gradient one, as revealed in wordlikeness judgments. 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.

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. In this study, we explore the role of partial and complete morphological decomposition in determining the acceptability of novel possible words (pseudowords). The influence of morphological decomposition in processing known words has been previously investigated via semantic priming. In a seminal priming study, Marslen-Wilson et al. (1994) show that transparently-derived words prime embedded words, but words without transparent derivation do not: e.g., CLEANER primes CLEAN but TINSEL does not prime TIN. They also argue that semantically opaque complex words (such as DEPARTMENT) and pseudoaffixed words (such as CORNER and PIGMENT) are not decomposed. Their priming results point to a processing model in which morphological decomposition involves deep semantic processing. Under such a model, no decomposition is predicted for pseudowords, which lack semantic information.

However, further work has altered this perspective. Taft (2004) argues instead that morphological decomposition is obligatory when possible, with differences in processing deriving from relative frequencies of the base and affixes. Rastle et al. (2004) show no difference in priming between transparent words (CLEANER) and words with opaque or false derivation (DEPARTMENT or CORNER); and both of these produce more priming than words with only partial formal similarity (BROTHEL may slightly prime BROTH). In a study of the processing of ambiguous novel compounds, Libben, Derwing, and Almeida (1999) find evidence for a prelexical parser that makes all possible analyses available to the lexicon. In a related study using semantic ratings plus familiarity-decision and priming tasks, Libben, Gibson, Yoon, and Sandra (2003) show evidence for decomposition of semantically opaque compounds. Lehtonen, Monahan and Poeppel (2011) find MEG evidence for early decomposition of pseudoaffixed words, and Kuperman, Bertram, and Baayen (2008) report that information about both stems and suffixes within complex Finnish compounds is used immediately, before the full word has been accessed. While these results suggest that processing is only shallow (not reliant on semantics), recent work has argued that the deep and shallow processes operate in parallel, with evidence of morphosemantic effects from the very beginning of processing (Feldman, Milin, Cho, del Prado Martín, & O'Connor, 2015). Feldman et al. (2015) show graded priming effects across a range of lag times: relative to unrelated pairs, the most priming was shown for transparent pairs (TEACHER–TEACH), and moderate priming was shown for opaque, pseudoaffixed, or partially-decomposable pairs (CORNER–CORN, RATIFY–RAT, or CORNEA–CORN). Such results suggest two processes: deep morphosemantic analysis of known words, as well as shallow, form-based decomposition.

These findings offer only a little insight into the processing of novel words, which have no established meanings. It is known that phonotactic cues to morphology influence acceptability ratings of pseudowords: in their study of phonotactic effects on wordlikeness ratings, Hay, Pierrehumbert, and Beckman (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. Analyzing rejection times in a lexical decision task, Taft and Forster (1975) find that prefixed pseudowords are analyzed into constituent morphemes before lexical access occurs. Deacon and Kirby (2004) report that general morphological awareness is a predictor of children's success in reading novel or made-up words. In the current study, we investigate the potential role of such pseudomorphology in wordlikeness judgments using an analysis of the highly varied pseudo-compounding and pseudo-suffixation that are exhibited in our phonotactically balanced set of short and long pseudowords.

2.1. Predictors of Wordlikeness

In order to test the effects of pseudomorphological parsing in our dataset, we establish a baseline to control for other known factors of wordlikeness. The wordlikeness of a pseudoword is influenced by two general factors. 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 they already know. These influences 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 (Storkel, Armbrüster, & Hogan, 2006), and so we consider them separately.

2.1.1. 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 et al., 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).

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. In addition, the probabilistic n-gram approach is pervasive in computational applications like phoneme to grapheme (P2G) conversion and automatic speech recognition (Martin & Jurafsky, 2000; Hahn, Vozila, & Bisani, 2012).

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. ‘Smoothing’ of n-gram statistics is also used to mitigate sparseness and sampling issues. 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 smoothing is represented by the independent combination 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 units 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).

2.1.2. Similarities to Existing Words

In this study, we consider morphological decomposition as a relevant dimension of similarity between pseudowords and existing words. Another major approach to word similarity is the lexical neighborhood. 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. ‘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 and Pisoni (1998), Bailey and Hahn (2001), Hahn and 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 (Marian et al., 2012; Heller, 2014; Storkel, 2004).

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. Given that two words share a morpheme by virtue of similarities in form and meaning, psycholinguistic research on morphological processing offers another window into how the similarities amongst words in the lexicon shape the wordlikeness of pseudowords.

2.2. Importance of Pseudowords, Limitations of Pseudoword Sources

For addressing gradient wordlikeness and for a wide variety of questions in linguistics and psycholinguistics, pseudowords are a critical resource. Judgments of pseudowords and nonwords shed light on the nature of phonological representations and abstractions (Coleman & Pierrehumbert, 1997; Frisch et al., 2000; Hayes, Zuraw, Siptár, & Londe, 2009; Hayes & White, 2013). They are used to investigate word processing and the structure of the lexicon: as foils in lexical decision tasks (Meyer & Schaneveldt, 1971; Taft & Forster, 1975; Forster & Davis, 1984); in ‘wugs’ tasks to assess the productivity of morphological patterns in children (Berko, 1958) and adults (Pierrehumbert, 2006b; Ernestus & Baayen, 2003; Zuraw, 2007). The novel status of pseudowords minimizes lexical frequency effects, and they lack the conventionalized semantic content that real words have (though similarities to real words may in some cases suggest semantic features). Researchers studying categorization and lexical semantics can instead impose semantics, by pairing the pseudowords with images or with discourse contexts that constrain the word meanings (as in Imai, Gentner, & Uchida, 1994; Alegre & Gordon, 1999).

As stimuli, pseudowords have the advantage over real words in that they can be selected to evenly occupy the space of possible words, a space which real words occupy only sparsely. Therefore, pseudowords are better suited to provide sufficient variety of experimental items, even with a several design constraints. For the same reason, pseudowords can better represent the full range of wordlikeness; real words are biased toward the high-probability end of the phonotactic space because they are shaping precisely the forces of wordlikeness being investigated. Pseudowords are also used for studies of encoding and memory during word learning, using measures such as word repetition accuracy (Edwards, Beckman, & Munson, 2004), and lexical priming and cortical response changes (Davis & Gaskell, 2009). Finally, artificial language research necessarily uses pseudowords. For example, the artificial language experiment described in Schumacher, Pierrehumbert, and LaShell (2014) used non-English pseudowords generated with a Welsh phonotactic model in order to encourage participants to be open-minded about whether the language presented would or would not have English-like inflectional morphology.

Research explicitly eliciting Likert-scale wordlikeness ratings of pseudowords, such as the present project, uniformly find that wordlikeness is gradient. This result leads to the conclusion that the cognitive representation of wordlikeness is also gradient. Wordlikeness has also been shown to affect speed and accuracy in repetition and lexical decision tasks, and performance in implicit memory tasks (for examples, see Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Frisch et al., 2000; Muncer, Knight, & Adams, 2014). To examine the gradient properties of wordlikeness, stimuli are needed which comprehensively cover the space of wordlikeness. However, current scored or normed pseudoword sources have important limitations in their coverage of the space of possible words, which are reviewed below. To address some of these limitations, we analyze a new dataset of 8400 pseudowords and nonwords, PseudoLex. In order to explore effects of morphology, PseudoLex systematically spans a broader range of phonotactic likelihood (including impossible nonwords) for monosyllabic and multisyllabic forms, for a broader range of pseudoword lengths (4 to 7 phones), normed by collecting 24 judgments per item.

In order to investigate wordlikeness effects for longer or lower-probability pseudowords, PseudoLex addresses some specific limitations of existing pseudoword sources. These sources often have limited coverage of the space of possible words and phonotactic models. For example, pseudowords may be generated by starting from a real word and constructing a nonword neighborhood around it by inserting, removing, or substituting a single phoneme (e.g., Irvine Phonotactic Online Dictionary presented in Vaden, Halpin, & Hickok, 2009). This approach misses many possible forms, and does not control for overall phonotactic likelihood. Possible forms are also frequently missed when studies with pseudowords are forced to focus on very short stimuli, often only monosyllables (e.g., ARC Nonword Database: Rastle, Harrington, & Coltheart, 2002). Some pseudoword sources use orthography only instead of a phonemic transcription. For languages with opaque spelling systems, this is a step removed from phonology. The resulting pseudowords are useful for reading research, but not necessarily for other types of research.

More general pseudoword sources are often still limited to biphone phonotactic methods, which may fail to capture regularities over larger timescales (e.g., WordGen: Duyck, Desmet, Verbeke, & Brysbaert, 2004). These effects are important in many languages (e.g., English). For example, a biphone model is able to produce a pseudoword like STSTSTS (due to the existence of both ST and TS sequences, in words such as such PASTOR, STOP, CATSUP, PATSY, and CAHOOTS). A triphone model correctly avoids such a string. Other methods can also address the biphone limitations on capturing larger structures, such as the subsyllabic approach of the tool “Wuggy” (Keuleers & Brysbaert, 2010). The simple triphone model used in PseudoLex is able to supply the broad range of stimuli that span high and low acceptability that the current study requires.

2.3. Construction of PseudoLex

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 address the limitations described above and thereby enable more comprehensive modeling of the relationship between wordlikeness and phonotactics.

2.3.1. How PseudoLex Items are Generated

For PseudoLex, statistical phonotactic models are used to generate items. The models are trained on a monomorphemic subset of CELEX. 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 baseline 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 baseline 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.

2.3.2. 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 rarely a 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’.

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

Figure 2-1. Example graphone mappings.

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.

2.3.3. Wordlist

We generated 8400 pseudowords with a CELEX-based corpus of 11382 monomorphemic words in phonemic representation (Baayen, Piepenbrock, & Gulikers, 1995). The input corpus was hand-edited to ensure that words were monomorphemic (Hay et al., 2004), to avoid capturing the boundary phonotactics of compound words (e.g., in the triphone /ɔtd/ from HOTDOG). 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 3); 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 versus 89,871 wordforms), and it has better coverage of American vernacular. Homophones and homographs were not common: for 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 2-1). 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.

Table 2-1. Example stimuli for the 28 cells of the current study. Each cell represents 300 stimuli.

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

2.3.4. Morphological Decomposition

To explore the role of morphological similarities to existing words, the stimuli were analyzed post hoc for suffixation and compounding patterns. 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 decomposition 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*) occurs when the output stem is a pseudoword (e.g. IOD + 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 them can be imagined; 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 2-3 in section 2.6.1.

The analyses generated by the Lancaster stemmer and the compounding analysis may include both spurious and missed decompositions. 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 decomposition 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, 2017), but 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.

2.4. 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.

2.4.1. Methods

2.4.1.1. Participants. The study collected data from 1440 participants via Amazon Mechanical Turk (825 female, 608 male; 7 participants declined to provide gender information). 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, and Barenboym (2011) report higher item variability for an online versus an in-lab task. However, capturing such variability may be a useful step toward understanding the natural range in human cognition. Current lab studies are unduly reliant on Western college undergraduates as participants (Heinrich, Heine, & Norenzayan, 2010), and online data collection makes it possible to recruit a more diverse participant pool (Gosling, Sandy, John, & Potter, 2010).

2.4.1.2. 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 included in the PseudoLex dataset.

2.4.1.3. Procedure. Data were gathered online by recruiting participants through Amazon Mechanical Turk (AMT). Participants chose the experiment “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 how English-like it was 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. Each participant rated 140 pseudowords, composed of 5 items for each of the 28 cells in the design. After completing the 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 baseline 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.

2.4.2. Effects For Replication

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.

2.4.2.1. 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 extends to longer items and to less-probable items. We evaluate both triphone and biphone models. Traditional biphone-only versions fails 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.

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

2.4.2.3. Vocabulary level. Frisch and 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.

2.4.2.4. 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).

2.4.2.5. 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').

2.5. Replication

The phonotactic model of wordlikeness is validated by demonstrating the statistical significance of biphone and triphone scores, subject to relevant controls (described in the previous section). Figure 2-2 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.

Word scores decline systematically with word-length, as discussed above. Figure 2-3 illustrates this pattern for the biphone and triphone scores. Ratings also decline systematically with length, although the within-length variation of ratings (relative to the difference in median between one length and the next) is greater than for the phonotactic scores. This indicates that additional factors besides biphone and triphone scores play a role in the participants' judgments.

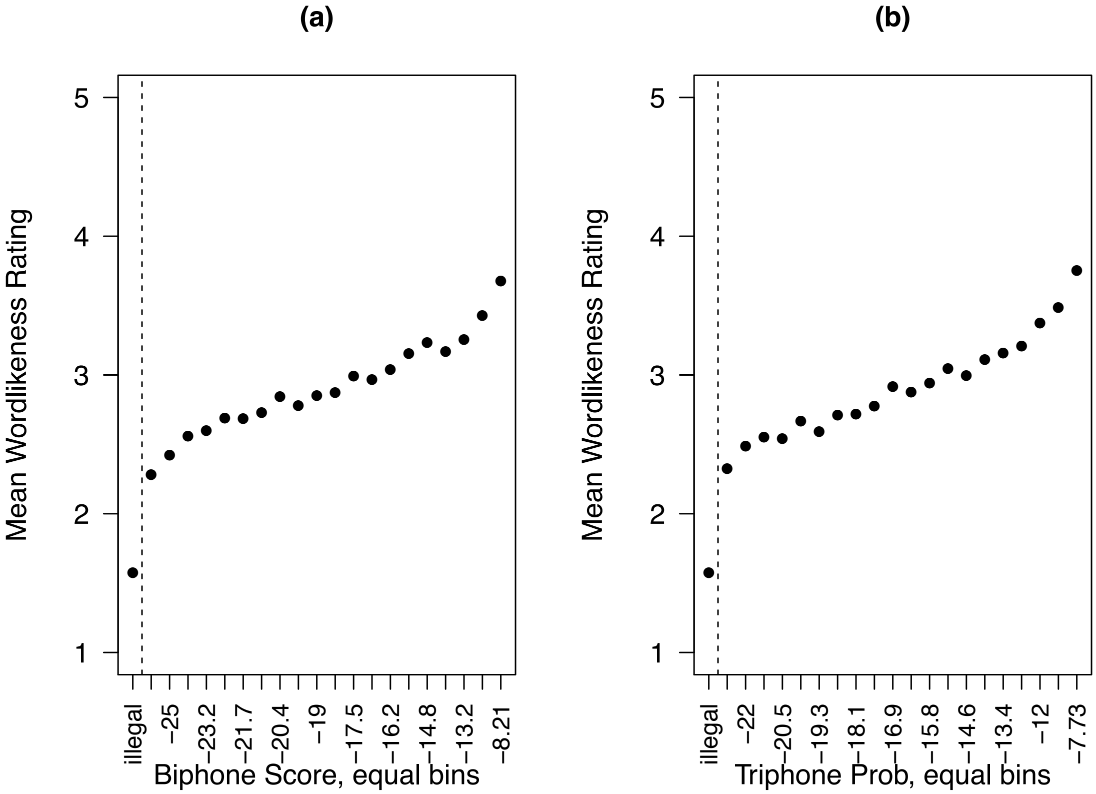


Figure 2-2. Mean wordlikeness rating by log phonotactic probability scores. Bins contain equal observation counts, pooled over all lengths: a) biphone score, b) triphone score. The baseline items (labeled as “illegal” on the x-axis) are rated lower than the lowest-scored legal items. On the average,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

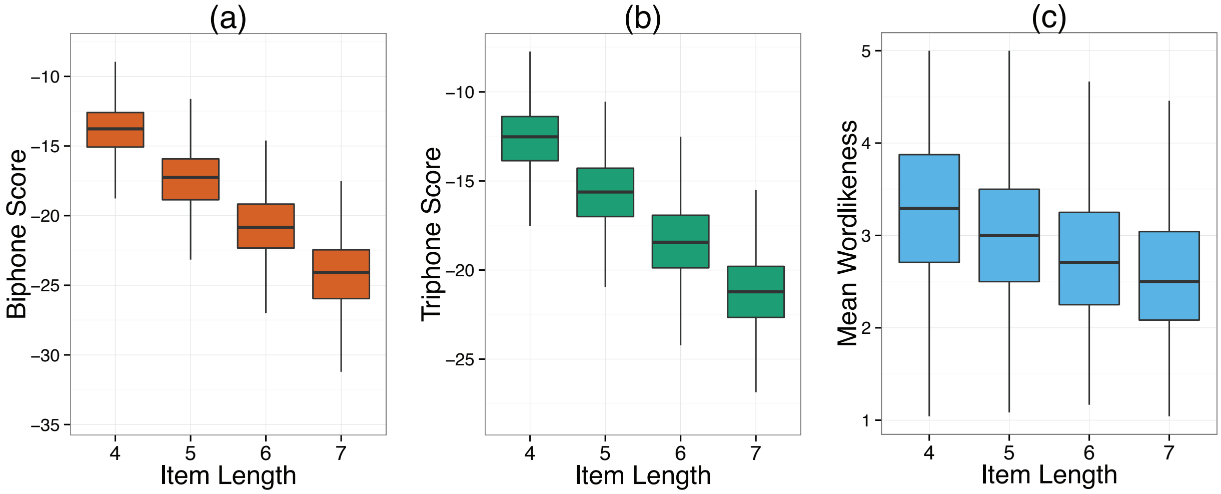


Figure 2-3. Boxplots of phonotactic scores and wordlikeness ratings, separated by word length: a) biphone scores, b) triphone scores, c) wordlikeness ratings.

The relationship of wordlikeness ratings to the replicated effects was evaluated using linear mixed-effects regression (LMER) implemented in R package ‘lme4’ (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2014). 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 the high correlation between word scores and length, we divided the analysis into 4 models by item length. This means that the relationship of length to wordlikeness judgments is not directly statistically evaluated here. 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, a base 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 base 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-value < 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-value < 0.05, X^2 method); these values are reported in Appendix 5. The four resulting models ('Baseline' models) are summarized in Table 2-2; information for factors excluded from a model is marked by '–'. Models in the 'Baseline' set are suffixed with 'A'.

Table 2-2. Model factors in Baseline models.

Baseline 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

2.5.1. 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. This high correlation makes including all four factors in the same LMER models problematic. 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 Baseline models described on Table 2-2, in which bigraph and trigraph orthotactic score factors were substituted for the phonotactic score factors 'biphone' and 'triphone'. These equivalent models are similar overall; to the extent that they differ, the fit of the phonotactic versions is uniformly superior (though only slightly). 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.

2.5.2. 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.

2.5.3. 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 2-2 as 'vocabulary'. Vocabulary level is also involved in significant interactions, reported below.

2.5.4. Lexical neighborhood. The presence of one or more orthographic neighbors provides a significant positive influence on an item's wordlikeness rating (see Table 2-2, '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.

2.5.5. Factor interactions. As shown on Table 2-2, 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).

2.6. Effects of Pseudomorphology

The Baseline models presented on Table 2-2 provide a baseline 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 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.

2.6.1. Morphological decomposition.

Two morphological processes were explored: suffixation and compounding. These are both highly productive in English, whereas prefixation is less productive. Items were coded as having a full suffixation parse, a partial suffixation parse, or no suffixation parse ('suffix_real', 'suffix_pseudo', 'suffix_none'); and as having a full compound parse, a partial compound parse, or no compound parse ('compound_real', 'compound_pseudo', 'compound_none'); these categories are described in more detail above (section 2.3). Of items with any parse, 29% have both suffixation and compound parses. 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 this method because it is highly replicable and minimizes the need for additional assumptions. Example decompositions are shown on Table 2-3; note that some real stems may be unfamiliar to the participants, so some 'real' parses could function as 'pseudo' parses.

Table 2-3. Example decomposition outcomes for stimuli by length and parse type. Morpheme boundaries are marked with '+', and '–' indicates no examples exist.

Item Length	No Parse	Suffixation Parse		Compound Parse	
	None	Pseudo	Real	Pseudo	Real
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

The effect of compound decomposition on the distribution of these intercepts is shown in Figure 2-4; 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 compounds parses).

The pattern is similar for suffixation, though there are key differences. Figure 2-5 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 2-4 and Figure 2-5 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.

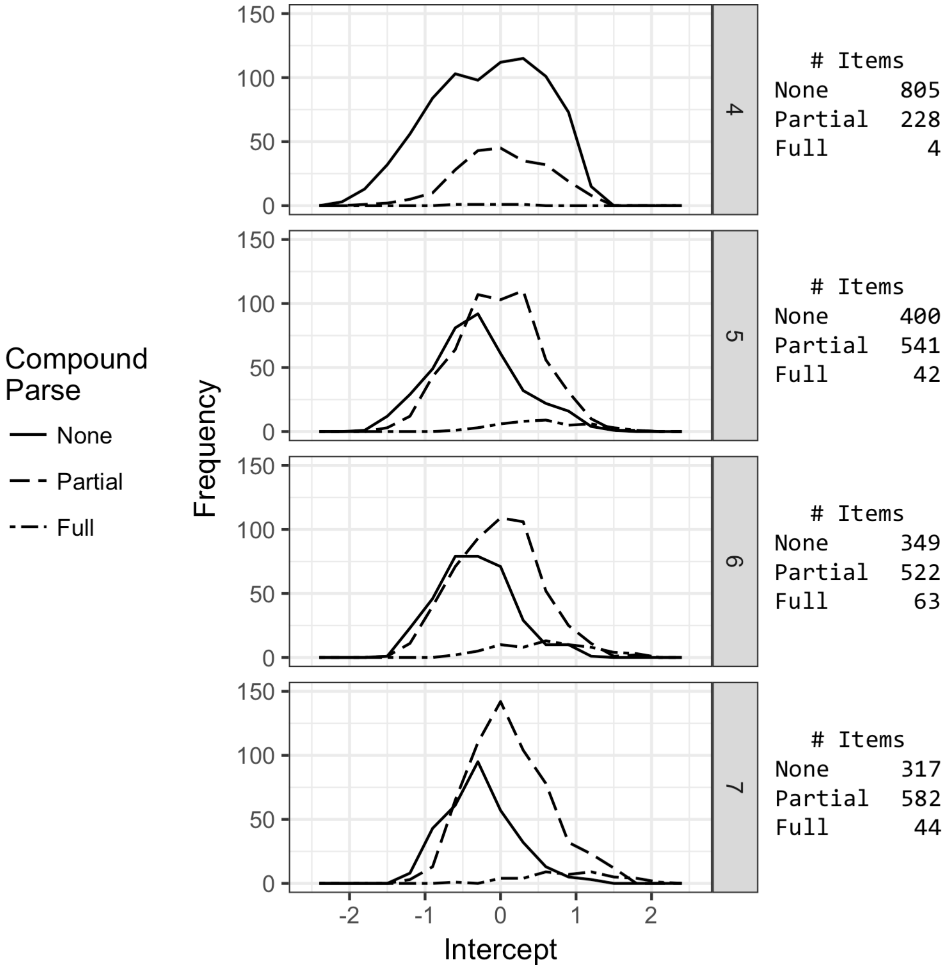


Figure 2-4. Effect of a compound parse on the distribution of wordlikeness intercepts, for pseudowords of length 4 to 7. Each panel shows superimposed histograms of the number of pseudowords having the indicated intercept value. The count of items in each category is given to the right.

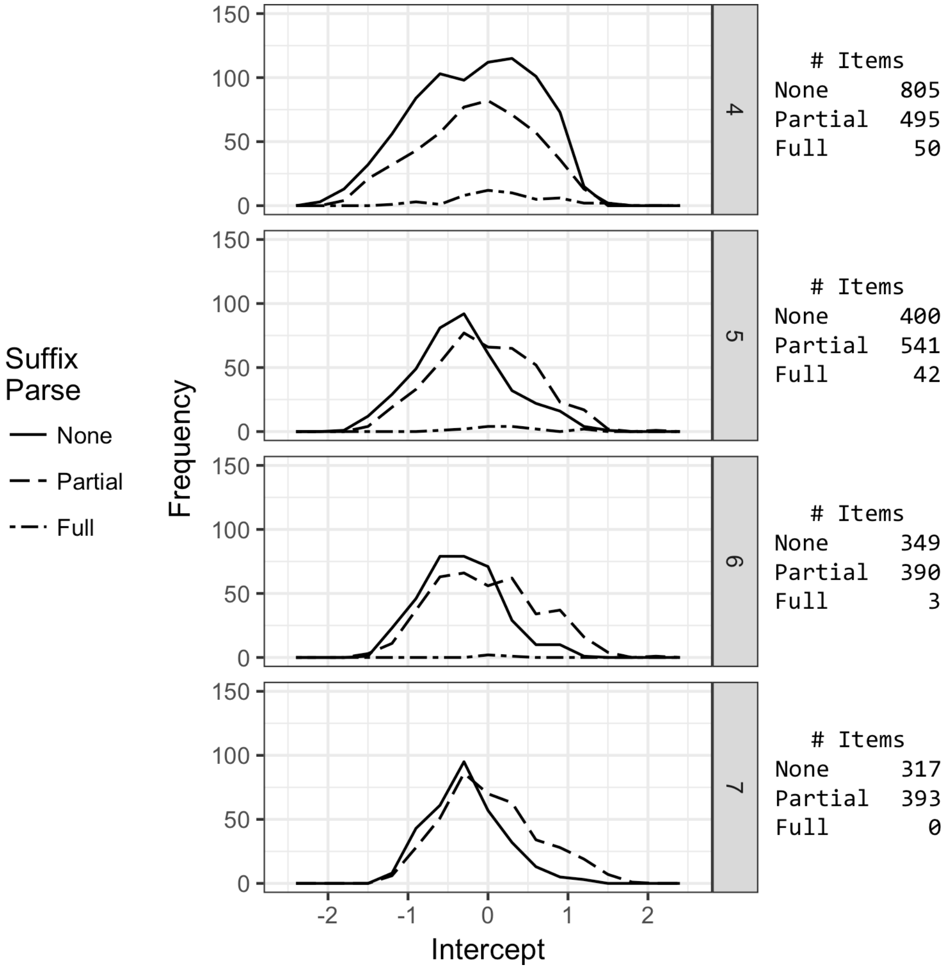


Figure 2-5. Effect of a suffix parse on the distribution of wordlikeness intercepts, for pseudowords of length 4 to 7. Each panel shows superimposed histograms of the number of pseudowords having the indicated intercept value. The count of items in each category is given to the right.

2.6.2. Modeling decomposition effects. 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, a base 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 base models were pruned by removing insignificant factors to yield the final models; see Table 2-4. 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 DUMP + OUSER, but also as a compound of the real stem DUMP with a pseudostem 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 Base 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.

Table 2-4. Model factors in Decomposition models.

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_pseudo	0.404	11.12	0.312	9.62	0.279	9.51	0.345	12.07
compound_real	0.366	1.67	0.975	11.72	0.813	12.61	1.147	15.53
suffix_pseudo	–	–	0.212	7.11	0.243	8.50	0.245	8.75
suffix_real	–	–	0.521	7.87	0.320	2.81	0.354	1.60
neighbors:vocabulary	-0.001	-3.25	–	–	–	–	–	–
neighbors:compound_pseudo	–	–	-0.272	-3.80	–	–	–	–
neighbors:compound_real	–	–	-0.555	-2.27	–	–	–	–
biphone:suffix_pseudo	–	–	–	–	–	–	0.022	2.09
biphone:suffix_real	–	–	–	–	–	–	0.021	0.29
triphone:suffix_pseudo	–	–	-0.032	-2.19	–	–	–	–
triphone:suffix_real	–	–	-0.092	-3.11	–	–	–	–
vocabulary:compound_pseudo	–	–	–	–	0.001	3.31	0.001	3.90
vocabulary:compound_real	–	–	–	–	0.002	3.34	0.002	3.01
vocabulary:suffix_pseudo	–	–	–	–	–	–	0.001	2.26
vocabulary:suffix_real	–	–	–	–	–	–	0.003	1.63

For all lengths, the presence of a partial compound parse ('compound_pseudo') has a significant and positive effect on wordlikeness rating (see factor estimates and t-values on Table 2-4). For all lengths except length 4, the presence of a complete compound parse (‘compound_real’) yields a significant and larger positive effect on wordlikeness; because there are only 9 compound_real 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_pseudo’, ‘vocabulary:compound_real’): 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 2-6.

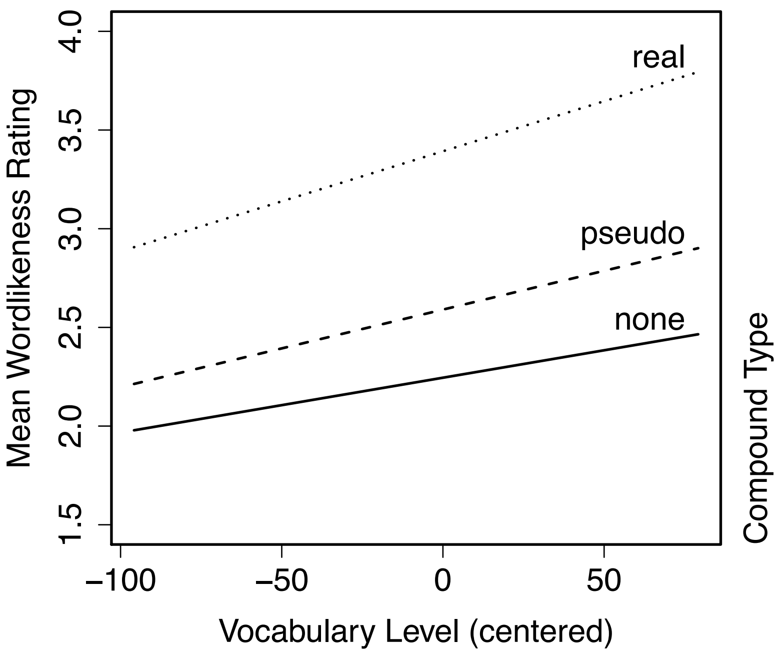


Figure 2-6. Interaction of vocabulary level with compound type as captured in Decomposition models, for pseudowords of length 7. Matches to existing words of English have a greater positive effect on ratings by people who know more words.

Compound parsing also interacts with neighborhood density in the Length 5B model ('neighbors:compound_pseudo', 'neighbors:compound_real'). The positive effect of lexical neighbors on wordlikeness rating is significantly reduced when a partial or real 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 real compounds with and without lexical neighbors rests on only three items that are analyzed as real compounds and have lexical neighbors: CHIPPET, YONNET, and MODGEM. It is far from clear that the embedded words in these items are as psychologically salient as the corresponding real 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_real’ and ‘suffix_pseudo’) 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_pseudo’) yields a significant positive effect on wordlikeness (see factor estimates and t-values on Table 2-4). The presence of a complete suffix parse (‘suffix_real’) has a significant and larger positive effect in the models for Length 5B and Length 6B; note that the t-value of ‘suffix_real’ 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_pseudo’) and with biphone probability (‘biphone:suffix_pseudo’); as with the suffix parse main effect, the specific factor level ‘biphone:suffix_real’ 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 2-7. The cross of the ‘real’ and ‘pseudo’ lines is unlikely to be consequential, because only a small minority of participants have such low vocabulary scores.

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 2-8. The category of ‘real’ suffixations is omitted from Figure 2-8a because there are only 4 such forms, of which two were probably mis-parsed. The category of ‘real’ suffixations is omitted from Figure 8b because all such items also had analyses as compounds, and as a result the effect size distinguishing the ‘pseudo’ items from the ‘real’ 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.

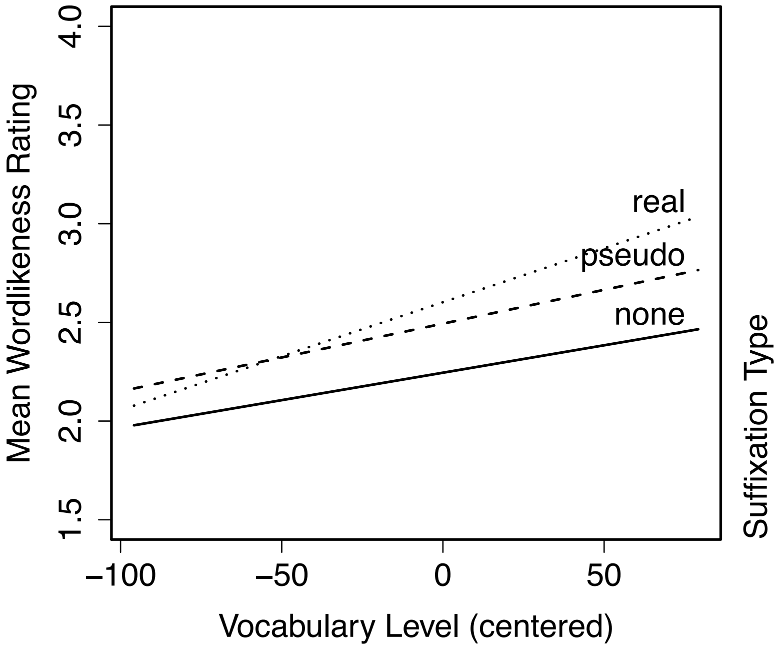


Figure 2-7. Interaction of vocabulary level with suffixation as captured in Decomposition models for length 7. Matches to real suffixes and words of English have a greater positive effect on ratings by people who know more words.

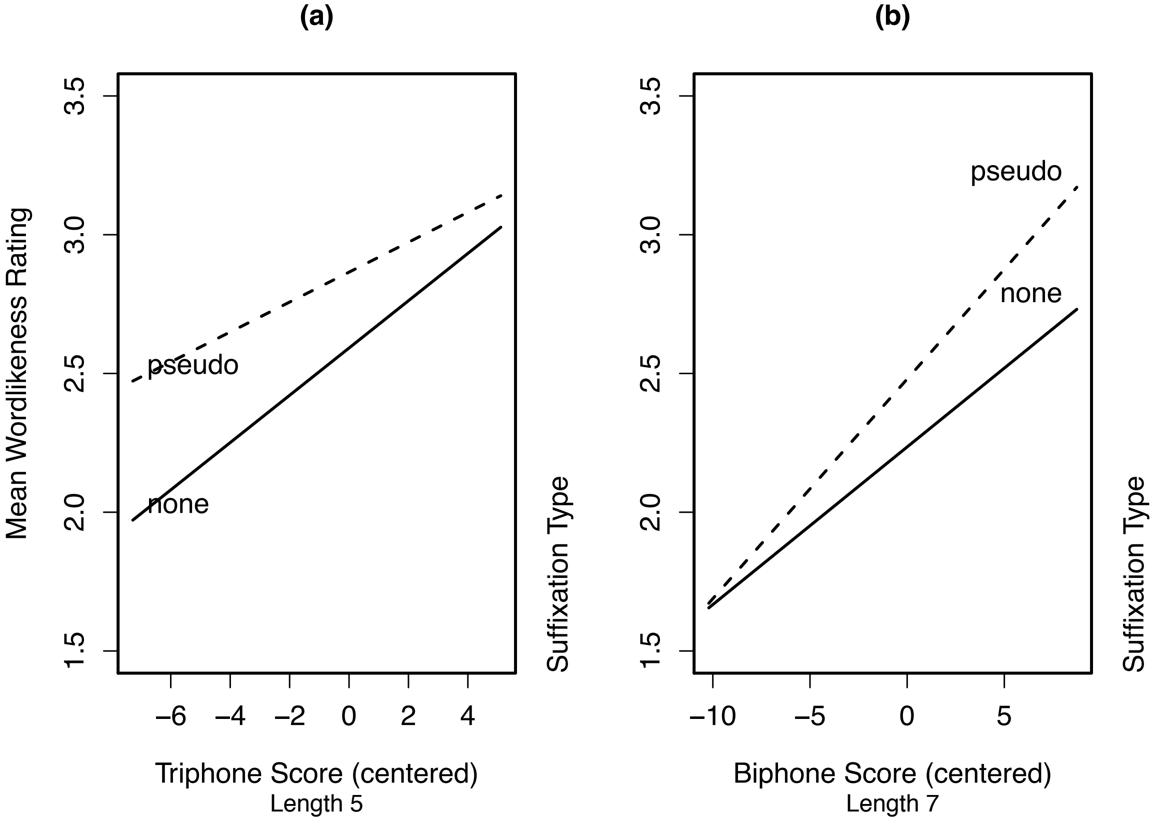


Figure 2-8. Interactions of phonotactic score with suffixation. (a) triphone scores for pseudo-suffixed and non-suffixed items of length 5. (b) biphone scores for pseudo-suffixed and non-suffixed items of length 7.

2.7. Discussion

The wordlikeness results in the Baseline models replicate effects for biphone and triphone likelihood, vocabulary level, word length, and lexical neighborhood. In particular, biphone and triphone scores are effective for 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) when available, while also using more abstract biphone knowledge as needed. In addition, when scores based on orthotactic probabilities were substituted for those based on phonotactic probabilities in the Baseline model, the orthotactic and phonotactic scores were shown to provide effectively identical information for the PseudoLex inventory.

Statistical analysis of lexical similarity in the form of potential morphological decomposition reveals highly significant effects on wordlikeness. Because pseudowords have no established meanings, and because many of the decompositions do not conform to the syntactic and semantic constraints of English morphology, the benefit from such decompositions points to the existence of shallow morphological processing, as suggested by Taft and Forster (1975), Libben et al. (1999), Deacon and Kirby (2004), Hay et al. (2004), Rastle et al. (2004), and Lehtonen et al. (2011). We have shown that evidence of such processing exists even if phonotactic likelihood and other factors are controlled, and it is involved not only in the on-line processing tasks explored by previous researchers, but also 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 fact that the 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.

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 score increased the positive effect of having a suffix parse (see ‘biphone:suffix_pseudo’). This may indicate that it is difficult for a suffix alone to redeem the poor phonotactics of 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_pseudo’). 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 decomposition 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 decomposition 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 and 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 decomposition (Length 6B and Length 7B) and suffix decomposition (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.

2.8. Conclusion

This study provides evidence that people process novel words using their morphological knowledge, in addition to lexical and phonological statistics. The Baseline models replicate the 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 that baseline. Such effects imply shallow morphological decomposition of novel wordforms. The parallel existence of shallow and deep processes has previously been shown for real words, and suggested in other linguistic domains as an efficient, flexible strategy for perception in noisy and variable contexts (Sanford & Graesser, 2006). The enhanced acceptability of parseable pseudowords should give them an advantage in being added to the lexicon over phonotactically legal words of comparable length. The lesser advantage for words with a partial parse suggests that pseudowords containing ‘cran-morphemes’, though viable, should be less readily assimilated than fully decomposable pseudowords.

3. Gradient Typicality Judgments of English Words

3.1. Introduction

The characterization of possible words in a language is a central goal of phonology, and is important to many lines of linguistic inquiry, including child language acquisition, studies of speech errors, and models of lexical innovation. This characterization is made on the basis of speakers' judgments of what forms are acceptable and unacceptable in their language, along with statistical patterns in the lexicon and corpora. In addition to the categorical division between possible and impossible forms, we use the term *typicality* to make a gradient distinction between possible words that are highly typical words in the language, and possible words that are less typical. This gradient typicality is correlated with statistical lexical measures, which are in turn important predictors for speed and accuracy in lexical processing research. Two major approaches to modeling typicality are phonotactic probability and lexical neighborhood density. These two perspectives are often highly correlated, but have been shown to provide different and complementary cues to word typicality. Vitevitch and Luce (2016) survey studies showing that models of lexical neighborhood density are explanatory for a variety of experimental tasks, and that phonotactic probability information can provide additional predictive power, particularly for possible forms that are not words.

However, these effects are not uniform across all tasks and contexts. Johnston and Kapatsinki (2011) show that word learning in adults is improved by a phonotactic novelty preference; i.e., a benefit from the presence of low-probability phonotactic patterns, rather than high-probability ones. This replicates a previous study in which Storkel et al. (2006) not only found the phonotactic novelty effect for adult word learners, but also the familiar pattern that higher neighborhood density improved word learning. Because phonotactic probability and neighborhood density are correlated in the lexicon, the opposition of these two effects is particularly striking. It does not seem to be the case that the adoption of words into the lexicon is governed by this phonotactic novelty effect, though. Pierrehumbert (1994) used human experimental judgments and statistical analysis of triconsonantal clusters in English to argue that the lexicon was shaped by phonotactic pressures: there are thousands of possible triconsonantal clusters that could be formed by the probabilistic combination of attested bigrams, but nearly all attested trigrams ranked in top 200 most likely. In addition, only 50 of these likely trigrams were actually attested, implying that the lexicon was influenced by negative phonotactic constraints, such as widely-attested OCP constraints. The human judgments presented support the conclusions that speakers have and use phonotactic knowledge about the probability of these sequences, as well as knowledge of other phonotactic constraints.

3.1.1. Typicality effects for pseudowords

We use the terms *pseudowords* to describe forms that are possible in a language, but unattested; we reserve the term *nonwords* for forms that are impossible in a language. In their study of pseudowords, Bailey and Hahn (2001) note the issue that probability and neighborhood are correlated in previous work which showed effects of phonotactic probability and of neighborhood density on pseudowords; and that this correlation is found in the lexicon of real words. Bailey and Hahn collected typicality judgments in a set of experiments with pseudoword stimuli specifically designed to decorrelate these two factors. They find that probability and neighborhood density have separate positive gradient effects on typicality judgments. These findings are replicated by Needle, Pierrehumbert, and Hay (in review), who use a large set of pseudoword stimuli which spans a wide range of phonotactic probability and item lengths. They show that gradient typicality judgments are partially explained by phonotactic probability, and that the lexical neighborhood density effect is also present for shorter items; the majority of longer items have no neighbors at all. They further find a positive correlation between typicality ratings and the presence of apparent morphology in the randomly-generated pseudowords.

3.1.2. Typicality effects for real words

Some phonological theories predict that real words would exhibit gradient well-formedness, while others do not. Here, we consider the treatment of both morphologically complex and simple real words (or, more precisely, of attested forms). In traditional Optimality Theory (OT) accounts, words are categorized as ill-formed or well-formed by a ranking of constraints; real words cannot be ill-formed (Prince & Smolensky, 1993). Hayes (2000) provides a modification of OT to account for some gradience in situations where a given form is participating in variation. By allowing the constraint structure itself to vary in strictness and in constraint ranking, certain marginal parses are made possible, and well-formedness will correspond to a weighted combination of the possibilities for a given speaker. Morphological structure plays a central role in the variation Hayes describes, as in the example of the dark and light /l/ in GALE-Y vs GAI-LY. In such a case, Hayes assumes that “in judging a given form, a consultant will normally assign it the highest rating possible under the grammar”. We might then expect that there is no gradience in judgments when a form is not in variation, as in morphologically simple words. Sensitivity to morphological structure is similarly important for Ernestus and Baayen (2003), who are able to capture gradient acceptability judgments in complex Dutch forms. While a traditional OT approach consists of negative constraints, and treats attested words as equally well-formed, Ernestus and Baayen construct a stochastic OT model using both positive and negative constraints in pairs (“mirror-constraints”) to allow gradience for attested and unattested phoneme sequences. Using a conceptually similar approach, Martin (2011) also predicts gradient well-formedness for attested phoneme sequences within and across morphemes in English. Martin’s model uses a Maximum Entropy grammar, which consists of a set of OT-like constraints in which each constraint is assigned a weight. Like Ernestus and Baayen’s mirror-constraint method, Martin uses a broad set of positive and negative constraints to yield gradience for attested and unattested forms.

While these approaches succeed in predicting gradient well-formedness in cases where morphological complexity creates some degree of variation or uncertainty about the single best result, a comprehensive theory will also address simple real words. Hayes and Wilson (2008) use a Maximum Entropy model to determine the constraints on English onsets. The model is trained on a corpus of English onsets, so morphological structure is not involved. The set of constraints is pruned from a space of negative constraints, with preference given to more general and parsimonious constraints. The resulting model assigns gradient well-formedness scores to unattested onsets, but uniformly high scores for real onsets. In their comparison of different models of well-formedness of English onsets, Daland, Hayes, White, Garellek, Davis, and Norrmann (2011) point out that the models which best match human judgments for unattested onsets also perform the worst for attested ones. The Hayes and Wilson model is among these: it performs well for unattested onsets, but assigns all attested onsets the same high score instead of reflecting the gradience of the human judgments. Based on the evidence from Daland et al. (2011), it does not seem that gradient well-formedness is absent for simple words, and present when morphological complexity intervenes. Instead, the Hayes and Wilson results may stem from the different constraint selection and pruning procedure for the Maximum Entropy model: unlike the positive and negative constraints used by Martin, and by Ernestus and Baayen, the constraints of Hayes and Wilson are only negative, and pruned with a preference for simple, general constraints.

It makes sense that well-formedness theories might treat real words as extremely well-formed. There are well-attested effects of being a known word over being a pseudoword, and of high versus low word frequency. This real-word advantage and frequency advantage may in fact be aspects of the same mechanism, given that all and only real words have non-zero frequency. For example, real words in English are often resistant to regularizing pressures within the lexicon; irregular past-tense forms such as DOVE persist despite regular competitors (DIVED). Indeed, it is precisely high-frequency real words which best resist such pressures (Bybee, 1995). Given this, we should ask whether the real-word and frequency advantages are absolute, or if real words show gradient typicality judgments, as pseudowords do. There is good reason to hypothesize that real words actually differ in typicality: real words, like pseudowords, do vary in phonotactic probability and in neighborhood density. These factors have been shown to affect language production and perception; for example, Vitevitch and Luce (2016) note that words with dense neighborhoods are processed more slowly and less accurately. Experimental evidence of gradient typicality for real words across these factors has been limited, as the use of real words can entail a number of additional experimental concerns: e.g., participant vocabulary, word frequency, age of acquisition, or emotional valence.

3.1.3. The current study

We present a extended post-hoc analysis of the data from Bailey and Hahn (2001). The original study analyzed typicality judgments for pseudoword stimuli, but the data gathered included judgments about the real-word items included as fillers. The current analysis combines the judgments for these real words with the pseudoword data to consider whether typicality is gradient for real words, and if that gradience is affected by the same factors as pseudowords. Mixed-effects regression allows real words and pseudowords to be compared across the same subjects; we consider the effects of phonotactic and orthotactic probability, phonological and orthographic neighborhood density, and word frequency. We find that phonotactic and orthotactic probabilities correlate positively with typicality judgments for pseudowords and real words. We also find that orthographic neighborhood density correlates positively with typicality, but phonological neighborhood density does not. For real words, word frequency also has a positive effect on typicality.

3.2. Data and Methods

3.2.1. Design and Stimuli

In Bailey and Hahn (2001), 24 participants (17 female, 7 male) rated items for their typicality as words of English on a scale from 1 to 9. Stimuli were presented in orthographic visual form, and participants were instructed to focus on the sound of each item. The stimuli consist of 328 simple monosyllable items: 69 real words and 259 pseudowords, generated by a syllable-formation grammar. Of the 259 pseudowords, 22 are ‘isolate’ and 237 are ‘near miss’; isolate items have a string-edit distance of 2 from the nearest real word in the lexicon, and near miss items are chosen such that they have a string-edit distance of 1 from both an isolate and some real word in the lexicon.

3.2.2. Analysis

For each item, we calculated orthotactic and phonotactic probability, orthographic and phonological neighborhood density, and log-transformed word frequency. The orthotactic and phonotactic probability values are smoothed trigram probabilities calculated using interpolated Witten-Bell smoothing in the SRI Language Modeling Toolkit (Stolcke, 2002). The orthographic and phonological neighborhood densities are calculated using the CLEARPOND tool (Marian et al., 2012) with CELEX data (Baayen et al., 1995). Word frequency data is from CELEX and log-transformed. Because majority of the stimuli are pseudowords, which have no frequency, log word frequency is binned into 3 categories in the analysis: items with frequency of 0 (all of the pseudowords) are ‘unattested’, attested words falling under the midpoint of the log frequency range (log F < 4.75) are ‘low frequency’, and attested words above the middle of the frequency range are ‘high frequency’. In this post-hoc analysis, real words are only 21% of the stimuli; and unlike the pseudowords, the real words were not carefully constructed to limit the correlation between neighborhood density and phonotactic probability (see Table 3-1).

Typicality judgments are modeled using linear mixed-effects regression (LMER) with the lme4 package (Bates et al., 2015) in the R software environment (R Core Team, 2014). The dependent variable is participant-rated typicality, scaled from 0 to 1. We included random intercepts for each item and participant. Models are selected by beginning with a maximal effects structure, then pruning insignificant effects beginning with the highest order effects.

Table 3-1. For real words, the correlation between phonotactic probability and phonemic neighborhood density are much higher than for the pseudowords. The correlation for orthographic measures are also higher for real words than for pseudowords.

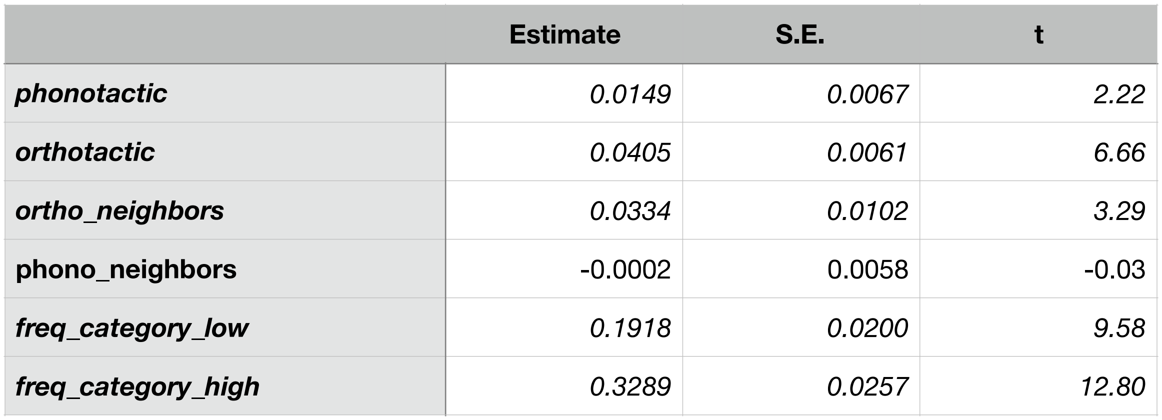
Correlation:	Phonotactics vs. Phono Neighborhood	Orthotactics vs. Ortho Neighborhood
All items	0.56	0.74
Real words	0.63	0.71
Pseudowords	0.31	0.63

3.3. Results

3.3.1. Regression results for pseudowords and real words combined

After pruning the LMER model, no interactions remained (see Table 3-2). There are positive effects on typicality for both phonotactic and orthotactic probability: items are rated higher as their probabilities increase (see Figure 3-1a, 3-1b). There is a positive effect of orthographic neighborhood density, but not of phonological neighborhood density (see Figure 3-1c, 1d); this difference may be due to the visual task, or to orthographic outliers in the stimuli. For word frequency, there are positive effects for items that are low frequency or high frequency, in comparison to the baseline value of unattested items (see Figure 3-2). The estimated effect size for high frequency items is larger than for low frequency, suggesting a frequency advantage for typicality within the subset of real words; this effect is examined in section 3.3.2.

Table 3-2. LMER model summary for all items. Significant effects are italicized.



	Estimate	**S.E.**	**t**
phonotactic	*0.0149*	*0.0067*	*2.22*
orthotactic	*0.0405*	*0.0061*	*6.66*
ortho_neighbors	*0.0334*	*0.0102*	*3.29*
phono_neighbors	-0.0002	0.0058	-0.03
freq_category_low	*0.1918*	*0.0200*	*9.58*
freq_category_high	*0.3289*	*0.0257*	*12.80*

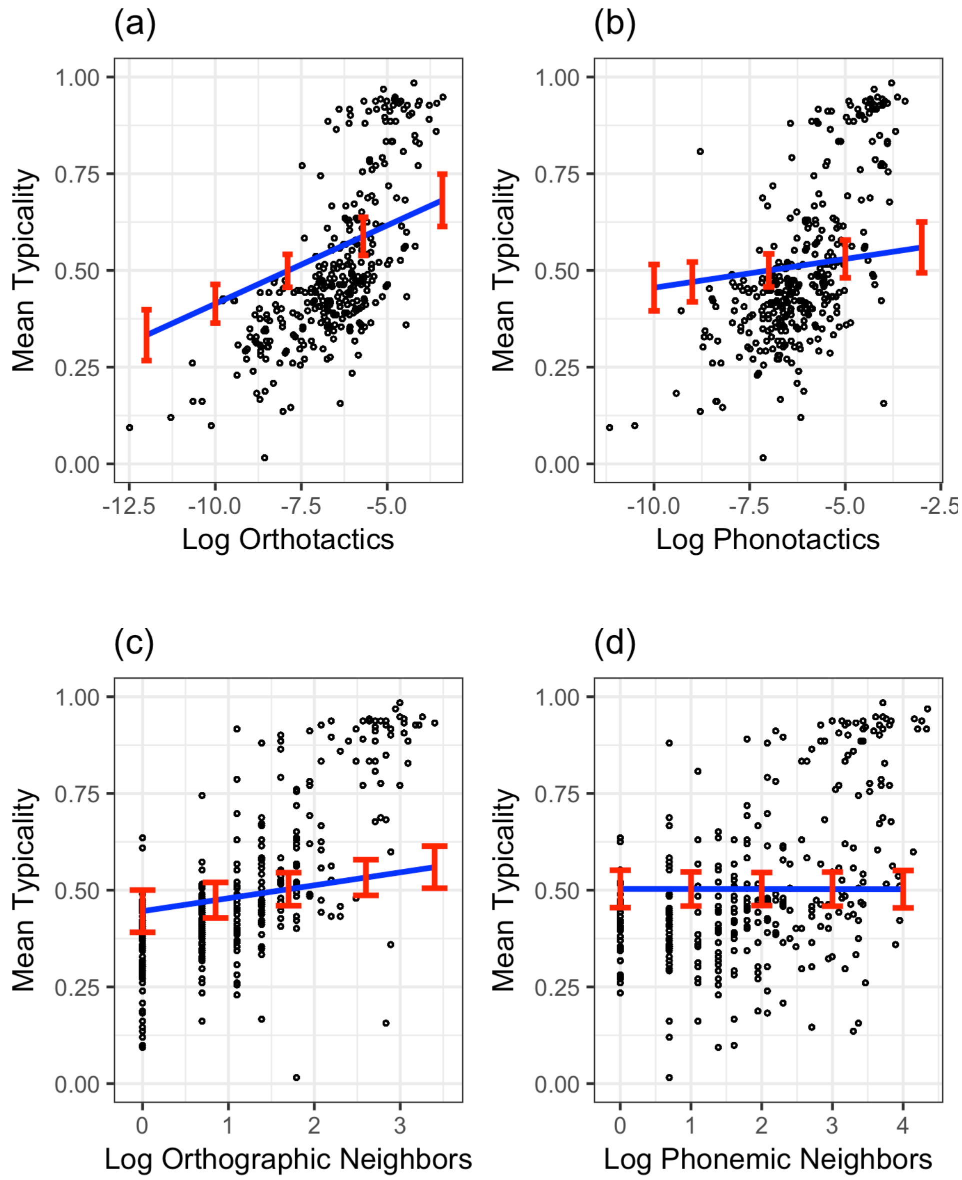


Figure 3-1. Main effects in the LMER model (all items). (a) Effect of orthotactic probability; (b) effect of phonotactic probability; (c) effect of orthographic neighborhood size; (d) effect of phonemic neighborhood size. Points indicate mean typicality judgment per item. Distribution plots are superimposed with model fits: blue lines are linear effect estimates, and red bars show 95% confidence intervals across the domain.

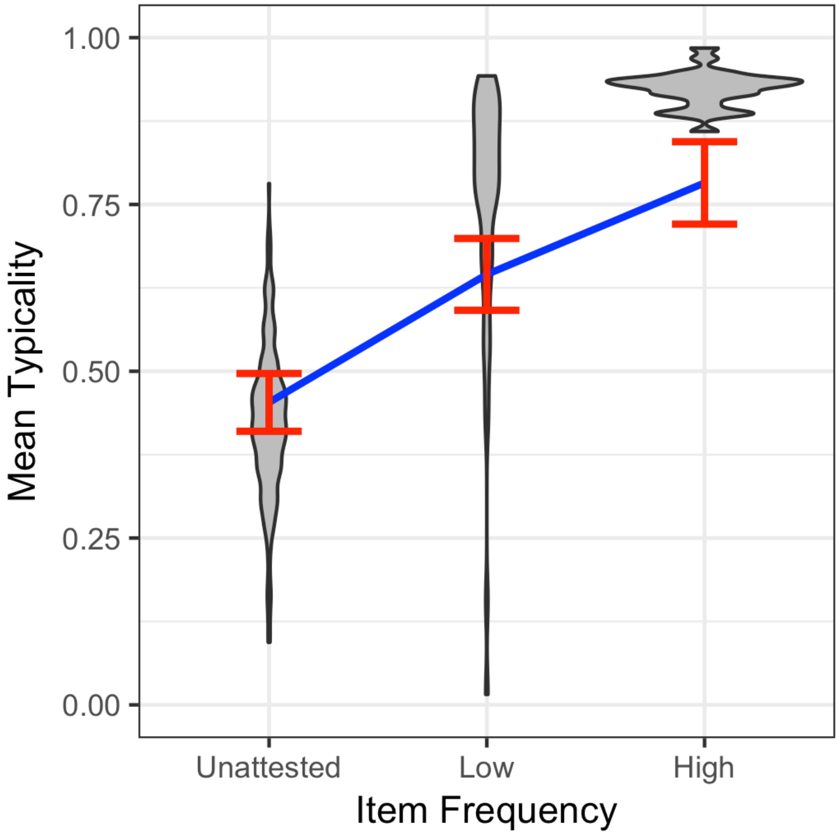
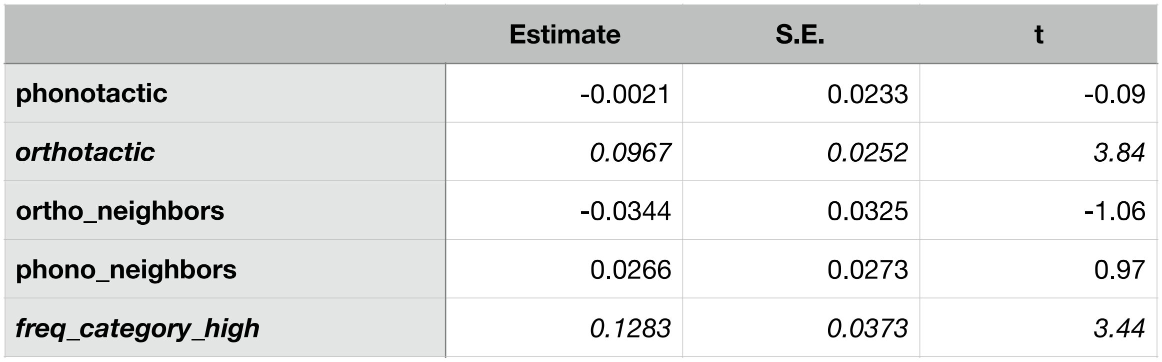


Figure 3-2. Effect of item frequency. Distribution plots are superimposed with model fits: the blue lines is the linear effect estimate, with red bars showing 95% confidence intervals for each category. Mean typicality is lowest for unattested items, higher for low frequency items, and highest for high frequency items.

3.3.2. Regression results for real words only.

The subset of real words only was also analyzed, due to the possibility that the pseudowords could simply overshadow the real words in the model. This data subset is much smaller, and there are stronger correlations between factors (e.g., between phonotactic probability and phonemic neighborhood density). For real words only, we again see a significant positive effect of orthotactic probability: real words are rated as more typical when they have more probable sequences of letters (see Table 3-3). There is also a significant positive effect of frequency: high frequency real words are rated more typical than low frequency words. The effect of high frequency is shown on Table 3-3, using low frequency words as the baseline.

Table 3-3. LMER model summary for real words only. Significant effects are italicized.



	Estimate	**S.E.**	**t**
phonotactic	-0.0021	0.0233	-0.09
orthotactic	*0.0967*	*0.0252*	*3.84*
ortho_neighbors	-0.0344	0.0325	-1.06
phono_neighbors	0.0266	0.0273	0.97
freq_category_high	*0.1283*	*0.0373*	*3.44*

3.4. Discussion

We observe gradient typicality in human judgments of both real words and pseudowords, but we also see an advantage for real words over pseudowords, and for high-frequency words over low-frequency words. Both real words and pseudowords show the positive effect of orthotactic probability. Neighborhood density has a positive effect for typicality of pseudowords, but this effect was not significant in the analysis of the real word subset. However, the post-hoc nature of the analysis limits our ability to estimate relative effects on real words only: not only is the number of real words small, but the words chosen do not reflect the efforts of Bailey and Hahn (2001) to decorrelate phonotactic probability from neighborhood density. Indeed, it is specifically the natural correlation of these factors that led Bailey and Hahn to use carefully-selected pseudowords.

An adequate theory of well-formedness must include both mechanisms of similarity (e.g., probability, neighborhood density, analogy, or morphological analysis), as well as the real-word and frequency advantages. The observation of gradient typicality judgments for real words in addition to pseudowords provides support for linguistic theories in which real words may vary in typicality, such as the probabilistic generative grammars described by Coleman and Pierrehumbert (1997), and as elaborated by Hay et al. (2004). Gradient judgments for real words are also compatible with the constraint-based approach in the Maximum Entropy method used by Martin (2011). These data do not accord with the predictions of Optimality Theory, nor with Hayes and Wilson's (2008) implementation of their Maximum Entropy phonotactic learner. OT accounts of gradient typicality for real words are limited to cases of morphological alternation, but there is not such alternation in the current data. The Maximum Entropy phonotactic learner simply does not predict gradient typicality for attested forms.

The primary observation that typicality judgments are gradient for real words makes an important connection with psycholinguistic theories of lexical processing. The importance of frequency, likelihood, and neighborhood density has been demonstrated in a variety of behavioral tasks (see Vitevitch and Luce, 2016), and these same factors are shown to affect typicality judgments in our analysis. This suggests that typicality judgments are part of a single connected system of lexical processing, along with other tasks of interest to psycholinguistics, such as lexical decision, identification, and naming. This connection brings the study of typicality and well-formedness into mutually beneficial exchange with studies of lexical processing and other areas of psycholinguistics.

4. Explicit Morphological Decomposition of Complex English Pseudowords

4.1. Introduction

In the study of lexical processing, a large body of previous work has shown a variety of factors are involved when a person considers a word, including whether the word is known or familiar, the similarity of the word to the mental lexicon of known words, or morphological decompositions of the word. In particular, Needle et al. (in review) showed that the presence of an apparent morphological parse in pseudowords was associated with higher wordlikeness judgements. In that study, the apparent morphology was determined by automatic dictionary methods. This raises the question of whether the apparent morphology is actually recognized by speakers, and whether they are able to decompose both real words and pseudowords. Though Needle et al. controlled for lexical neighborhood density and phonotactic probability, increased wordlikeness for apparent morphology could be the result of other kinds of broad lexical similarity within the mental lexicon. The task of explicit decomposition instead requires morphological knowledge. The current study addresses the question of whether participants are able to perceive and identify morphemes within both words and pseudowords. This study investigates the ability of speakers of US English to explicitly decompose familiar complex words, and novel pseudowords containing real morpheme endings. We find that participants are highly accurate in decomposing the complex real words, and they are also successful for complex pseudowords. Performance in both of these conditions exceeds baseline levels that could be achieved by chance, or by an optimal phonotactics-only strategy.

4.1.1. Morphological decomposition models

Previous findings on the morphological processing of words have been variable. Marslen-Wilson et al. (1994) use priming evidence to show that pseudoaffixed words (such as BOOTY and CORNER) and semantically opaque complex words (such as DEPARTMENT) are not decomposed; they suggest that semantic processing is a required component of decomposition. This conclusion is challenged by subsequent work: Libben et al. (1999) find evidence that all morphological analyses are made available by a pre-lexical parsing process, and indeed, Libben et al. (2003) show that semantically opaque compounds were decomposed. It was also shown by Rastle et al. (2004) that words like CORNER are decomposed. These results suggest that the morphological processing of real words involves shallow, pre-lexical analysis in addition to the more in-depth semantic analysis suggested in Marslen-Wilson et al. The discrepancy between the findings of Marslen-Wilson et al. and the other decomposition studies mentioned may derive from paradigm differences: while the experiments in Libben et al. (1999), Libben et al. (2003), and Rastle et al. (2004) involve some similar mechanisms of priming and lexical decision for complex words, they are all visual tasks. The experiments reported by Marslen-Wilson et al. use a cross-modal lexical decision paradigm, in which the participant hears a spoken prime at the same time as they see a visual target word. This design is motivated by the timing differences between the modalities: an auditory word unfolds in time, while a visual word is perceived all at once. Differences in linguistic modalities must be considered, but it may also be the case that real morphological effects from the auditory prime are blocked or attenuated due to the change in modality. There is also evidence to suggest that pseudowords are processed similarly to real words: Taft and Forster (1975) find that prefixed pseudowords are decomposed before lexical access occurs, and Deacon and Kirby (2004) show that children's success in reading novel or made-up words is predicted by their general morphological awareness.

4.2. Methods and Data

4.2.1. Experimental paradigm

The study uses a new online experimental paradigm in which participants are shown a series of words and pseudowords, one at a time. This paradigm is used to gather explicit morphological decomposition responses for simple and complex items, as well as familiarity ratings for each item. Each word is presented with a user interface that allows a single marker to be placed between the letters of the word, indicating a decomposition boundary; and a set of buttons to give a Likert-scale rating of familiarity with the word. For each item, the participant responds to two tasks: a) "Split the word into two meaningful parts, if possible." and b) "Rate how familiar you are with this word.". The participant indicates a single position to split the item by clicking between the letters displayed to move the decomposition marker. To give the familiarity response, they click on the Likert scale below the item; see Figure 4-1 for images of example trials. These two tasks may be completed in either order, prior to clicking the 'Next' button to move to the next trial.

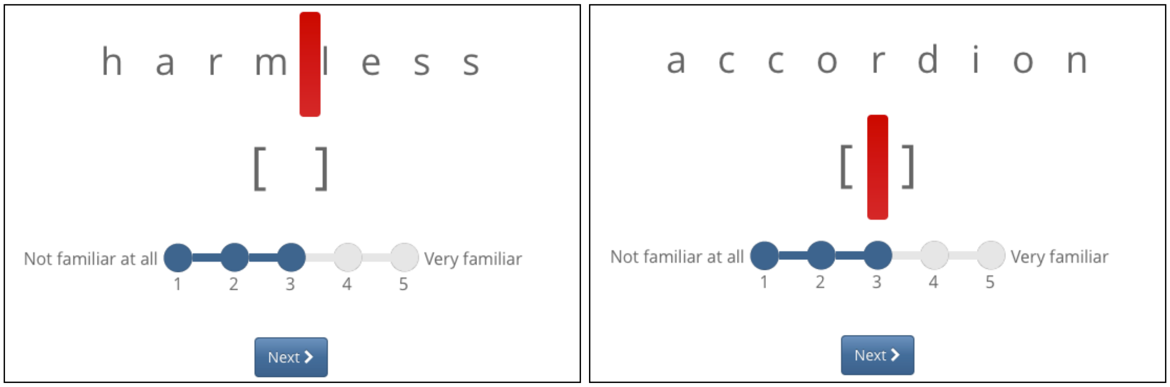


Figure 4-1. Example experiment trials, with decomposition and familiarity response. The left panel shows a decomposition placed and familiarity response of 3. The right panel shows a 'no decomposition' response.

4.2.2. Stimuli

Item stimuli consist of simple real words, complex real words, and complex pseudowords. Target morphemes include both compounding elements and suffixes. The complex pseudowords, designed to be comparable to the real complex words, consist of a pseudo-stem and a real morpheme ending. The stems for these pseudowords were drawn from amongst the 8400 pseudowords that were generated for the norming study presented in Needle et al. (in review). These vary in length and have statistical wordlikeness scores as determined by smoothed phonotactic and orthotactic scores. The stems selected for the present study all had above-median scores. In addition, stems with low ratings (regardless of score) were excluded. Thus, they were all of good phonotactic quality. Three additional criteria were imposed. The length distribution fell in the middle of that for real stems in the study. Stems were selected to have a phonotactically legal transition to the suffix, defined as having a digram probability within the range for the complex real words. Combinations with unanticipated word embeddings were eliminated by hand; for example, EGAUSSAGE was not used as an example of a word with the suffix -AGE because it contains the words GAUSS and SAGE. The complex real words use different morphemes from the pseudowords, and their stems are always able to stand alone (e.g., GRASS in GRASSLAND).

The experiment has 288 items: 108 complex real words, 108 complex pseudowords, and 72 simple real words. Pseudowords have three examples each of 36 morphemes. Complex real items have three examples each of 36 morphemes. During item selection, frequent morphemes and words were preferred. The morphemes used include both suffixes and compounding endings. For suffix-type morphemes, 24 are consonant-initial and 24 are vowel-initial. For compound-type morphemes, all 24 are consonant-initial. It was not possible to find 24 vowel-initial words that satisfied the selection criteria for this experiment and the planned followup related to author gender. The morphemes vary in productivity, and both morphemes and whole words vary in length and frequency. For examples of experiment items, see Table 4-1. Summary statistics on frequency for the items are provided in Table 4-2.

Table 4-1. Example stimuli by category, including compounds and suffixations.

Simple Real	**Complex Real**		**Complex Pseudo**	
	-ful	**+ light**	**-ium**	**+ case**
tennis	graceful	firelight	balnium	snoshcase
straight	lawful	searchlight	vodepium	clumcase
porcelain	handful	daylight	thrafium	pelpcase
72 items	***108 items, 36 real endings***		***108 items, 36 real endings***	

Table 4-2. Summary statistics for frequency of real word stimuli.

	Log Frequency			
	Min	Max	Median	SD
All real words	0	8.8	4.4	1.65
Simple reals	3.9	8.1	5.0	1.01
Complex reals	0	8.8	3.6	1.74

4.2.3. Participants

The study collected data from 216 participants via Amazon Mechanical Turk (94 women, 120 men; two participants declined to provide gender information). All participants reported being English speakers currently residing in the United States. Reported birth years range from 1943 to 1995 (one participant declined to answer). All participants completed the experiment between 2017-6-7 and 2017-6-9. Participants were paid $3 for completing the task, which took up to 30 minutes. Six participants were excluded for insufficient decomposition performance on simple and complex real words combined (d' < 1) (see Figure 4-2).

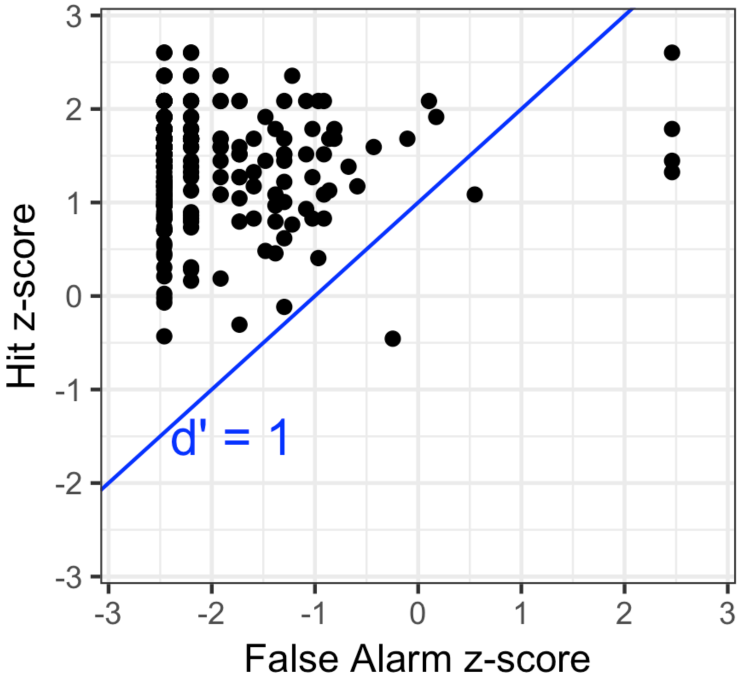


Figure 4-2. Decomposition accuracy for participants. Each participant's performance is a black point. The x-axis shows the relative rate of False Alarms (incorrectly parsing a simple real word), and the y-axis shows the relative rate of Hits (correctly parsing a complex real word). The blue line shows the exclusion criterion of d-prime greater than 1.

4.3. Results

All real words were rated as highly familiar (for complex real words, M = 4.6, SD = 0.47; for simple real words, M = 4.8, SD = 0.13), and all pseudowords were rated as highly unfamiliar (M = 1.3, SD = 0.11). Participants were quite accurate in the decomposition task: the average accuracy for decomposition responses was 96% for simple real words, 88% for complex real words, and 65% for complex pseudowords. The decomposition assumed in constructing the stimuli is treated as the only correct response. This accuracy is impossible by chance, and it is also beyond the performance of an optimal strategy using phonotactic cues to the presence of a word boundary (see 4.3.1). We argue that the accuracy rates observed are the result of participants recognizing morphemes in the complex stimuli.

4.3.1. Decomposition mechanisms

We suggest that our decomposition accuracy results are best explained by a word-matching mechanism in which participants recognize known words and morphemes embedded in the stimuli. The word-matching mechanism is tested by Brent and Cartwright (1996) in a computer simulation of their model of infant and child learning for speech segmentation: they show effective segmentation is achieved by a combination of leveraging known words, with distributional knowledge of probabilistic phonotactic constraints on words. Brent and Cartwright suggest that adults, who have much larger vocabularies, make complete use of the word-matching mechanism in speech segmentation. The use of known words to gain word boundary information is also effectively used by Daland & Pierrehumbert (2011) as part of the array of segmentation learning models they test. These results imply that participants in the current study are not relying solely on phonotactic knowledge, but making considerable use of their morphological and lexical knowledge.

A mechanism of finding known words within complex pseudowords yields the expected decomposition for all compound-type pseudowords (e.g., CASE in PELPCASE) and 24 of the 72 suffixation-type pseudowords (in which the intended suffix matches a real word; e.g., SPURLGATE, SNOFFWORTHY). Our classification of these morphemes as either compounding or suffixing elements is taken from the hierarchical decompositions given in CELEX, but participants may decompose pseudowords without strong ideas about the type of morpheme they perceive. In 56% of the pseudoword stimuli, embedded morphemes match common real words. We suggest that participants are making use of this mechanism, and additionally that it operates over both known words and affixes (e.g., -ER in LONTER). The inclusion of affixes may help explain the high mean accuracy participants showed for the suffixed real-word stimuli, of which only 33% contained real embedded words.

We compare participant performance against a phonotactic boundary strategy to test the hypothesis that participants succeed in the explicit decomposition task using morphological knowledge. For both phonemic and orthographic bigrams at the location of the expected decomposition, we compared the likelihood of a boundary being present versus absent (taking the ratio of the log likelihoods) (cf. Daland & Pierrehumbert, 2011). The boundary likelihood ratio is derived from orthographic and phonemic bigram statistics in the 10931 CELEX monomorphemes, a list made as discussed in Hay et al. (2004) by hand-checking the lexical entries in the CELEX lexicon (Baayen et al., 1995). Monomorphemes are used so that the bigram statistics accurately reflect words without internal boundaries. Boundary likelihood ratio is defined as the probability that a boundary is present, divided by the probability that a boundary is not present. To estimate the probabilities for bigrams with boundaries, we make the simplifying assumption that words can combine freely.

The orthotactic version of the baseline strategy had similar overall performance to the phonemic version (37% versus 35%), so we present only the orthotactic results here. For the complex real words, the expected accuracy of the orthotactic strategy is 42%, less than half of the observed accuracy rate (88%). This discrepancy is overwhelming, and it is also possible that participants simply have explicit morphological knowledge for these known words, so we focus on the pseudoword results. For complex pseudowords, the orthotactic strategy correctly suggests parsing for 32 out of 100 items (32% accuracy); 8 of the 108 items are excluded because the boundary bigram statistics were unavailable. As for the complex real words, this rate is less than half of the average observed accuracy for pseudowords (M = 65%, SD = 27%). The distribution of individual performance on pseudowords has a strong negative skew, with 85% of participants exceeding the orthotactic strategy (see Figure 4-3).

Though participants are not relying solely on orthotactic information, there is some evidence that this factor does contribute to decomposition accuracy, as described in Brent and Cartwright's combined model. Perception of embedded morphemes should be easier when the morpheme boundary is more likely (i.e., when a matching cue is present). Participants were more likely than average to correctly parse items when the orthotactic strategy would be correct (M = 76%, SD = 10%); and the reverse is true for incorrectly-predicted items (M = 58%, SD = 19%). This pattern suggests that participants are sensitive to the presence or absence of the orthotactic cue. The effect of a negative orthotactic prediction must be of limited strength, however, because participants agreed with the orthotactic strategy's 'no parse' response only 33% of the time, and we may assume that some of those could be explained by participants deliberately choosing to ignore a perceived morpheme because of how they understand the semantic aspect of the task instructions (see section 4.4).

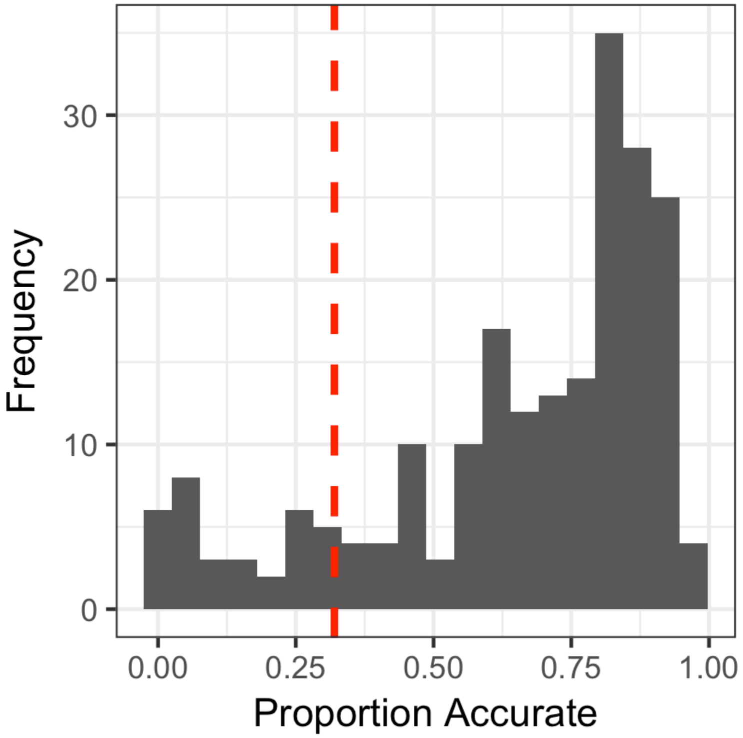


Figure 4-3. Histogram of participant accuracy rates for pseudoword items. The red dashed line marks the 32% accuracy predicted by the orthotactic cue baseline.

4.3.2. Automatic suffix recognition

Needle et al. (in review) used automatic methods to predict the morphemes participants might have been perceiving in their set of 8400 pseudowords: compound items were predicted by a word-matching dictionary method, and suffixed items were predicted using the standard Lancaster stemmer. However, they raised questions about the accuracy of the Lancaster stemmer in predicting realistic suffix parses, and suggested that it most likely underestimated the extent of suffixations that participants would recognize for these items. We show that the Lancaster stemmer has low accuracy for the suffixed pseudowords in this study: the stemmer produces the expected parse for only 20 of the 72 suffixed pseudowords (28%). Participants performed more than twice as well on these items (M = 60%, SD = 27%). In addition, the incorrect stemmer outputs are not reasonable alternatives to the expected decomposition, even though this was often the case for human errors. For example, participants often produced reasonable parses such as PSYCHO- and -LOGY for PSYCHOLOGY*,* where -OLOGY was expected. The stemmer frequently recognized no suffix, or a trivial suffix that a human would not find reasonable (e.g., -E, -TE); in other cases, the stemmer suggested a shorter real suffix embedded in the expected suffix (e.g., -ILE in -MOBILE). For example stemmer errors, see Table 4-3.

Table 4-3. Example performance of the Lancaster stemmer on pseudowords.

Pseudoword	Expected Suffix	Stemmer Suffix	Accuracy
bloudage	**age**	**age**	**Correct**
kliftarian	**arian**	**arian**	
lonter	**er**	**er**	
blouchify	**ify**	**ify**	
skirking	**ing**	**ing**	
balnium	**ium**	**ium**	
sporchling	*ling*	*ing*	*Incorrect (too short)*
flugmobile	*mobile*	*ile*	
spurlgate	*gate*	*ate*	
lemphitis	*itis*	*is*	*Incorrect (trivial)*
drenlike	*like*	*e*	
glumphette	*ette*	*te*	
gloffless	*less*	—	*Incorrect (none)*
droofoid	*oid*	—	
scloinosis	*osis*	—	
snoffworthy	*worthy*	—	

4.4. Discussion

We find that people are able to accurately decompose complex real words and pseudowords. The real word stimuli were highly familiar, so it is likely that participants knew their meanings; participants were also aware that the pseudowords were highly unfamiliar, and lacked meaning in the lexicon. Despite the unfamiliar and meaningless status of the pseudowords, most participants were able to provide accurate morphological decompositions for them. In fact, feedback comments show that some participants deliberately chose not to decompose pseudowords because of the mention of meaning in the experiment's instructions, implying that the decomposition performance results are conservative. This finding contrasts with the results of Marslen-Wilson et al. (1994), instead providing additional support for models of lexical processing which are not exclusively semantic (Libben et al., 1999; Libben et al., 2003; Rastle et al., 2004), and do not require exhaustive morphological matching (Taft & Forster, 1975).

This finding is relevant for theories of lexical processing and language change: we show that people are sensitive to morphology in novel or unknown words. This is a good processing strategy for English, in which morphologically complex words are common. Morphological productivity is a dominant process in lexical innovation, for old and new endings: we see many words with new and reanalyzed morphemes like -HOLIC, -CATION, CYBER-, -PEDIA, or SMART-; and established morphemes as in BLOATWARE, BLUEJACK, CLICKBAIT, BOTNET, and HUMBLEBRAG. If people perceive and process morphemes in pseudowords, their previous experience with those morphemes may influence their processing in other ways.

Specifically, indexical associations with morphemes may extend to novel words with those morphemes, affecting the uptake and spread of new words. We test for the effects of morpheme-mediated gender associations in a followup study (see Chapter 5).

5. Gendered Associations of English Morphology

5.1. Introduction

Morphological systems arise from experience with words as encoded in the lexicon. Both statistical and episodic information about words leave traces in mental representations (see reviews in Pierrehumbert, 2006a and 2016). Lexical statistics are known to be important in morphological learning, and learning in turn relates to change over time (Bybee, 1995; Bybee & Thompson, 1997; Komarova & Nowak, 2001; Daland, Sims & Pierrehumbert, 2007). However, there remains much unexplained variability in how people acquire and extend morphological patterns. In particular, lexical statistics alone fail to predict why some rare patterns become much more prevalent over time (Bauer, 2001). A factor that may contribute to this variability is social-indexical information. In the domain of allophonic variation, some variants become conventionally associated with different social characteristics. People can provide cues to their social identities and personae when they chose to produce these variants (see review in Eckert, 2008). This process provides an avenue for innovations to take hold, as people imitate people they admire or identify with (Labov 2001). Indexical associations have been documented for whole words (R. Lakoff, 1973) and for morphosyntactic patterns such as number and tense marking (Rickford & Rickford, 2000). The extent of such associations for derivational morphemes and compounding elements is less clear. These could in principle be excellent vehicles for social-indexical information, because they encompass a large number of different forms with rather unrestricted semantics. One can easily imagine that semantically similar affixes, such as -ITY versus -NESS, might be used preferentially by different groups. Some groups might use an affix where others use a compound or periphrastic (as in ROOMETTE versus SLEEPING COMPARTMENT). Here, we present a quantitative experimental study on the relationship of speaker gender to derivational morphology and compounding patterns. Speaker gender is a highly salient aspect of the linguistic context that has played a central role in sociolinguistic theory. We show that people have significant success in associating English words with speaker gender. Their implicit knowledge generalizes to gender associations of novel words (pseudowords), such as THRAFIUM and PELPCASE, that appear to be morphologically complex but have no established meaning. Our experimental protocol combines a morphological decomposition task with a social judgment task. By analyzing the combined results, we are also able to shed light on the cognitive architecture that is responsible for the generalization of gender associations to novel complex word forms.

5.1.1. Social-indexical information

Sociolinguistic variation arises in language when groups within a linguistic community develop different patterns of expression. Simple differences in linguistic experience can go towards explaining why people in one group may speak differently from people in another, but it does not provide the full story. Some—but not all—aspects of sociolinguistic variation enter general awareness, and are conventionally associated with groups of people, or with the stereotypical attributes of these groups (e.g., with attributes such as coolness, toughness, or sensitivity). When this happen, the variation has become indexical. It can be used by speakers to convey social information concurrently with their propositional message. Indexicalization thus requires the variation not merely to exist, but also to be represented in the cognitive systems of speakers and listeners.

Social-indexical variation in the domain of phonetic variation has been intensively studied. Building on the findings of sociolinguistic fieldwork, cognitive encoding of such variation has been revealed in a variety of experimental tasks. Purnell, Idsardi, & Baugh (1999) find that listeners are quite successful in identifying standard, African-American, and Chicano dialects of American English based on variation in the form of the word HELLO. Clopper & Pisoni (2004b) find that listeners are able to classify speakers into regional dialect groups. Hay, Warren & Drager (2006) find that the apparent social class of the speaker influences the perception of words that are phonetically ambiguous in the context of a merger in progress. Hay & Drager (2010) show that phonetic category boundaries are impacted by subtle priming of the Australian versus New Zealand dialects. Other studies have shown that lexical encoding and memory are compromised for dialects that are low-status or non-standard, even when word recognition has not been affected (Sumner & Samuel, 2009; Clopper, Tamati, & Pierrehumbert, 2016). Turning to production, German, Carlson, and Pierrehumbert (2013) describe an imitation experiment in which American English speakers learning the allophones of /t/ and /r/ of a Glaswegian English speaker generalize the target patterns to other words. They retain the ability to generalize the pattern a week later when their knowledge of Glaswegian dialect is re-activated by hearing speech recordings that do not contain any examples of the target patterns. This behavior clearly involves a cognitive association between the Glaswegian speaker or dialect, and the allophonic pattern. Gender is one of the most salient types of social-indexical information. Gendered associations for phonetic patterns are widely documented, affecting both perception (Johnson, 2006) and production (Foulkes & Docherty, 2006). Gender is of particular interest in models of language variation and change, because women often demonstrate earlier participation in emerging sound changes, at least in English, which is the most studied language (Eckert, 1989; 2008).

The observation that men and women differ in general patterns of word use goes back to R. Lakoff (1973). Large-scale quantitative studies supporting this observation include Boulis and Ostendorf (2005), which analyzed telephone conversations, online forum postings, and web pages; and Mihalcea and Garimella (2016), which analyzed a blog posts. In a historical corpus study, Nevalainen, Raumolin-Brunberg, & Mannila (2011) report gendered associations for whole words (YE vs. YOU), syntactic patterns (-ING OF vs. -ING), and also for affixes (-TH vs. -S). Because people tend to associate with others who share their interests, status, and expertise, such gender differences are correlated with differences in register and topic. In a study of different registers, Plag, Dalton-Puffer, & Baayen (1999) find that some affixes (e.g., -ITY, -NESS, -ION, -IZE) are more productive in writing than in speech; Bucholtz (1999; 2001) in turn discusses Greco-Latinate forms as part of a constellation of language variables used by the "nerd" community of practice at Bay City High School, a social label that reflects not only intellectual interests, but also gender and race.

For gendered social meanings to exist, gender differences in observed usage must be present. However, the presence of these usage differences is itself not sufficient to imply gendered social meanings. Therefore, observing gendered differences in morphemes may not mean that these morphemes are being used to carry social meanings. Indeed, Nevalainen et al. (2011), suggest that gendered differences may be explained by strong social divisions, not by gendered social meanings per se: "Women tended to lead vernacular changes, whereas men were the leaders of processes related to educated and professional written usage" (p. 4). Citing Labov

(2001), they suggest that abstract features like morphemes (in contrast to whole words and phonetic features) may be unlikely to be strongly associated with social meanings. Two recent experiments, however, provide indirect evidence that this skepticism may not be entirely justified. Using the Asch "social pressure" paradigm, Beckner, Rácz, Hay, Brandstetter, and Bartneck (2016) show that in a past tense formation task, people are influenced by other people but not by humanoid robots, indicating that social judgment acts a filter in morphological processing. Using an artificial language paradigm, Rácz, Hay, & Pierrehumbert (2017) investigate the learnability of interlocutor gender as determinant of variability in the form of the diminutive affix, finding that this contextual condition is as learnable as a phonological condition. Such reflexes of social factors in cognition for morphology put us one step further towards uncovering social-indexical meanings for morphological patterns. Adopting the methodology of socio-phonetics, we here address the issue more directly. First, we use corpus statistics to identify differences between men and women in the usage of words and morphemes. Then, we carry out a gender identification experiment using male-dominated, female-dominated, and gender-neutral forms. In a novel protocol, the identification task is combined with an explicit morphological decomposition task. The results have important consequences for the influence of social information on word formation and change in the lexicon.

5.1.2. Structure of the mental lexicon

This investigation into the relationship between social-indexical information and morphemes takes place in the context of active debate over the nature of the mental lexicon and morphological systems which derive from it. If our goal is to determine at what levels and to what units indexical information may attach, then competing ideas about the lexicon impose different constraints. Under ‘multiple-route’ models (e.g., in Hay & Baayen, 2005), morphemes, simple words, and complex words are all reified as lexical entries. Lexical entries for complex words may be accessed either directly or through the morphemes that comprise them. Phonotactic cues, frequency relationships, and semantic transparency all affect which route is more likely to succeed first, and the strength of the morphological boundary in a complex word is a gradient function of the access history. In fully analogical models, both simple and complex words are stored in the mental lexicon, and novel complex words are generated or parsed on-demand based on similarities amongst known words (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2010; Dawdy-Hesterberg & Pierrehumbert, 2014; Rácz, Pierrehumbert, Hay, & Papp, 2015). Words and morphemes may also be undifferentiated, as in the NDL (Naïve Discriminative Learner) model of Baayen, Hendrix, and Ramscar (2013). In this model, the concepts for affixes and roots have the same status, and letter sequences (e.g., trigrams) are linked directly to these concepts. This means that the meaning of the phrase A BRITISH PROVINCIAL CITY is encoded as a set of concepts (A, BRITAIN, ISH, PROVINCE, IAL, CITY). In the NDL, complex words are epiphenomenal results of patterns of association between phonological material and categories of meaning. In the Item-and-Process approach (Haspelmath & Sims, 2013), morphologically complex words are created by rules that add or modify simpler word forms. This is the standard approach in generative phonology, receiving a statistical implementation in the MGL (Minimum Generalization Learner) developed by Albright and Hayes (2003).

All these models would need to be augmented in some manner to support social-indexical associations. In interpreting our results, we will discuss simple model extensions, in which anything that appears in the ontology for a model is a potential host for a social-indexical association. For example, both morphemes and words in the multiple-route model might potentially host associations. In the MGL model, both stems and rules might be associated with social factors. It exceeds the scope of the present paper to consider more complex extensions that might potentially obtain social-indexical effects indirectly.

5.1.3. Morphological decomposition

In classical linguistic theory, the morpheme is the minimal unit of association between form and meaning, and complex words can be decomposed into two or more morphemes. A confluence of findings, reviewed in Hay & Baayen (2005), come together to indicate that the classic theory is oversimplified, and that the decomposability of complex words is variable and gradient. But decomposability remains consequential. Hay & Baayen (2001) address the observation that the type frequency of a morpheme is a surprisingly poor predictor of its productivity, showing that the prediction can be improved by assuming that complex words that are more frequent than their stems (such as STAIRS and GOVERNMENT) are accessed as wholes, and therefore do not contribute to the effective type frequency for the suffixes they exhibit. Hay (2002) shows that English suffixes are generally ordered with more decomposable suffixes outside of less decomposable ones. Hay et al. (2004) show that participants make use of statistical word-boundary parsing in order to make wellformedness judgments of pseudowords. The judgement is based on the best available parse. People respond as if an internal word boundary is present in pseudowords that contain consonantal sequences that are unattested or rare within monomorphemic words. However, it does not automatically follow from such results that morphological decomposition plays a role in social-indexical processing of speech. Insofar as real words have social associations, the extent to which social associations also accrue for their morphological components is not known. It is also not known whether social associations of known words generalize to novel words, and still less whether any such generalization occurs through overall similarities in word form, or through more structured morphological parsing.

The current study is a step toward untangling these questions. It answers the call of both Pierrehumbert (2006a) and Foulkes & Docherty (2006) to improve on traditional statistical models of language by incorporating social effects. Factors that need to be incorporated include intra- and interspeaker variability, social interpretation, and the cognitive mechanisms that connect these aspects. The study considers the gender association effects of whole words and morphemes, for simple real words, complex real words, and complex pseudowords. To evaluate the gender associations of morphemes, it focuses on a set of derivational suffixes and compounding elements that differ (according to a corpus study) in their rates of use by men versus women. Indexicality is evaluated by asking participants to perform a judgment task in which they associate word forms with faces of men and women. Participants also give an explicit decomposition for each word (or respond that no decomposition is needed), alongside the gender association response. Our analyses consider gender responses in conjunction with both the accuracy of morphological decomposition, and the objectively available support for morphological decomposition.

5.2. Methods

5.2.1 Corpus statistics

We selected the British National Corpus to survey gender bias for words and suffixes. It includes material from a variety of different genres for which the gender of the author can be determined. For this study, we used the written portion of the British National Corpus, and included only those documents that could be attributed to men or women authors. The British National Corpus written subset contains 3,141 documents, for a total of 87,953,932 words; after filtering for author gender, there are 378 documents from women (13,451,416 words) and 844 documents from men (28,659,100 words). We note that the corpus has more material written by men authors than by women authors. This may have improved the statistical estimates for man-biased forms. In addition, information currently available in the British National Corpus limits us to considering gender in terms of a man–woman binary. In this corpus, as in everyday life, author gender is correlated with the topic of discussion. More than half of the 'imaginative' content domain is written by women, making their relative representation over twice that of men. However, men are overrepresented in the other 9 content domains, especially 'natural science' (2300%) and 'commerce' (500%). In this study, we lack the information to tease apart these variables.

Following Mihalcea and Garimella (2016), we calculate the gender bias of each word as the ratio of use frequency by women versus men authors. This is expressed below as a log ratio. Negative values mean that the word is man-biased; women use the word less than men. Positive values mean than the word is woman-biased. The results broadly replicate Mihalcea and Garimella (2016) in finding that a large number of words display little gender bias, but a certain number are used much more by one gender than by the other. These provide targets for the experimental stimuli. Morpheme gender bias values were calculated as the ratio of grouped use frequencies for complex words sharing the final morpheme as determined from CELEX decompositions; e.g., the calculation for -LAND includes GRASSLAND, DREAMLAND, and so on. For compounds, the value is determined solely from appearances of the compounding element as the second element in a compound word, because some frequent compounding elements have diverged semantically from their meanings as isolated words. It may not be surprising that gender bias was present for a variety of compounding elements, and it also proved to be present for a variety of suffixes.

5.2.2. Presentation

The study uses a new online experimental paradigm in which participants are shown a series of words and pseudowords, one at a time. Each word is presented with a user interface to allow a single marker to be placed between the letters of the word, indicating a decomposition boundary; and accompanied by a pair of named face images. For each item, the participant responds to two tasks: a) “Split the word into two meaningful parts, if possible.” and b) “Which author most likely used this word?”. The participant indicates a single position to split the item by clicking between the letters displayed to move the decomposition marker. To give the gender response, they click directly on the face of either the man or the woman shown above the item; see Figure 5-1 for images of example trials.

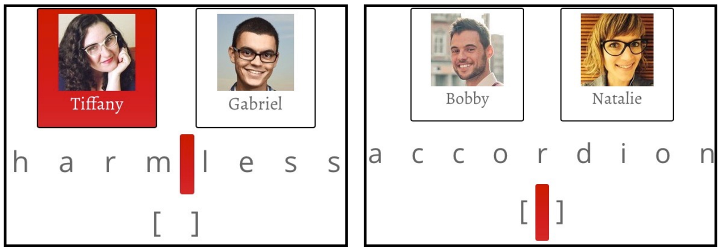


Figure 5-1. Example experiment trials, with gendered faces and decomposition responses. The left panel shows the left face selected and a decomposition placed. The right panel shows a ‘no decomposition’ response, with a face not yet selected.

This paradigm is used to gather explicit morphological decomposition responses for simple and complex items, as well as the implicit gender associations for each item. These two tasks may be completed in either order, prior to clicking the ‘Next’ button to move to the next trial. The explicit decomposition paradigm was previously validated using a baseline experiment (216 participants, 288 items each), in which participants gave morphological decomposition responses in addition to Likert ratings for item familiarity. The items in the baseline experiment are the same as those in the current study. All real words were rated as highly familiar, and all pseudowords were rated as unfamiliar. In the baseline experiment, the average accuracy for decomposition responses was 96% for simple real words, 88% for complex real words, and 65% for complex pseudowords (taking the “correct” decomposition to be the one assumed in constructing the stimuli). Accuracy is similar for the current experiment: 96% for simple real words, 86% for complex real words, and 65% for complex pseudowords (for full analysis of parsing, see 5.3.1.).

5.2.3. Stimuli

5.2.3.1. Author faces. Author faces were created using public domain images of 6 women and 6 men. This experiment used only faces appearing to be white adults between 25 and 40 years old (see Figure 5-2). Each of the 12 images was assigned a name based on the most popular names by gender in the United States since 1917 (Social Security Administration, 2016). We consider these names to have stable gender associations and familiarity for participants of varying ages. Each name is among the 10 most popular names for the 100 year interval, and all names are in the top 200 most popular for Americans born in the 1980s (which corresponds to the face age range). None have ambiguous gender. Names for each image are held consistent across all subjects. All different-gender pairs of men and women were used to make 36 distinct pairings, and were presented in 2 orders (man–woman, woman–man) for a total of 72 face pair orderings.

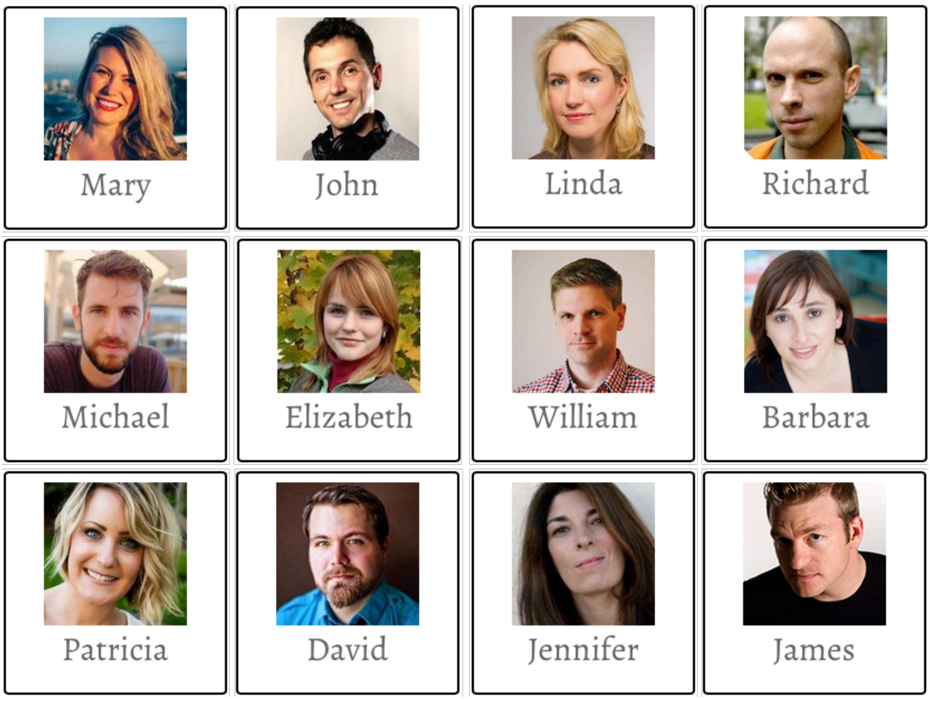


Figure 5-2. Gendered faces of 6 women and 6 men.

5.2.3.2. Items and script design. Item stimuli consist of simple real words, complex real words, and complex pseudowords. Each real word has a whole-word gender bias value. In each complex word, the second morpheme has a morpheme gender bias value. Target morphemes include both compounding elements and suffixes. The complex pseudowords, designed to be comparable to the real complex words, consist of a pseudo-stem and a real morpheme ending. The stems for these pseudowords were drawn from amongst the 8400 pseudowords that were generated for the norming study presented in Needle et al. (in review). These vary in length and have statistical wordlikeness scores as determined by smoothed phonotactic and orthotactic scores. The stems selected for the present study all had above-median scores. In addition, stems with low ratings (regardless of score) were excluded. Thus, they were all of good phonotactic quality. Three additional criteria were imposed. The length distribution fell in the middle of that for real stems in the study. Stems were selected to have a phonotactically legal transition to the suffix, defined as having a digram probability within the range for the complex real words. Combinations with unanticipated word embeddings were eliminated by hand. For example, EGAUSSAGE was not used as an example of a word with the suffix -AGE because it contains the words GAUSS and SAGE. The complex real words use different morphemes from the pseudowords, and their stems are always able to stand alone (e.g., GRASS in GRASSLAND).

The experiment has 288 items: 108 complex real words, 108 complex pseudowords, and 72 simple real words. Simple real words are balanced by whole-word gender bias: 24 woman-biased, 24 neutral, and 24 man-biased. Pseudowords are balanced by morpheme gender bias, with three examples each of 36 morphemes: 12 woman-biased, 12 neutral, and 12 man-biased. Complex real items are balanced for both whole-word gender bias and morpheme gender bias: 12 woman-biased, 12 neutral, and 12 man-biased morphemes; within each morpheme, there is one woman-biased, one neutral, and one man-biased whole-word example. During item selection, frequent morphemes and words were preferred. The morphemes used include both suffixes and compounding endings. For suffix-type morphemes, 24 are consonant-initial and 24 are vowel-initial. For compound-type morphemes, all 24 are consonant-initial. It was not possible to find 24 vowel-initial words that both occur frequently in compounds and exhibit strong gender bias. The morphemes vary in productivity, and both morphemes and whole words vary in length and frequency. For the Complex Real condition, it was necessary to reach farther down the word frequency scale than for the Simple Real condition to obtain enough items with strong gender biases. For examples of experiment items, see Table 5-1. Summary statistics on characteristics of the items are provided in Table 5-2.

Table 5-1. Example stimuli by category, including compounds and suffixations.

Simple Real	**Complex Real**		**Complex Pseudo**	
	-ful	**+light**	**-ium**	**+case**
tennis	graceful	firelight	balnium	snoshcase
straight	lawful	searchlight	vodepium	clumcase
porcelain	handful	daylight	thrafium	pelpcase
72 items	***108 items, 36 real endings***		***108 items, 36 real endings***	

Table 5-2. Summary statistics for real word stimuli: whole-word gender bias and frequency.

	Gender Bias				**Log Frequency**			
	Min	**Max**	**Median**	**SD**	**Min**	**Max**	**Median**	**SD**
All real words	-4.3	2.4	-0.03	1.25	0	8.8	4.4	1.65
Simple reals	-2.1	2.3	-0.1	0.99	3.9	8.1	5.0	1.01
Complex reals	-4.3	2.4	-0.03	1.0	0	8.8	3.6	1.74

5.2.4. Participants

The study collected data from 216 participants via Amazon Mechanical Turk (111 women, 101 men; four participants declined to provide gender information). All participants reported being English speakers currently residing in the United States. Reported birth years range from 1948 to 1996 (three participants declined to answer). All participants completed the experiment between 2017-5-1 and 2017-6-1. Participants were paid $3 for completing the task, which took up to 30 minutes. Six participants were excluded for insufficient decomposition performance on simple and complex real words combined (d' < 1).

5.3. Results

We first evaluate gender associations of whole-word and morpheme gender bias for simple real words, complex real words, and complex pseudowords (Regression Models 1, 2, and 3). Then, we consider the role of explicit morphological decomposition: does explicit parse accuracy necessarily imply morpheme awareness, or could performance instead be explained by phonotactic or orthotactic cues (Model 4)? Do participants need to correctly parse morphemes to be influenced by the gender associations of those morphemes (Model 5)? Finally, if explicit parsing is not required, does it nonetheless improve gender response accuracy?

The effect of word and morpheme bias on gender responses was analyzed using logistic mixed-effects regression with the function glmer implemented in R package lme4 (Bates et al., 2015) in R (R Core Team, 2014). For all regression models reported here, each continuous measure is centered in the models: whole-word gender bias, morpheme gender bias, and log word frequency. Final models are pruned: for each model, analysis begins by including all relevant fixed effects and their interactions, as well as slopes and intercepts for each random effect. None of the models converged properly with random slopes included, so the first step of pruning in each case was to remove random slopes (leaving random intercepts only).

Insignificant terms are removed from the models one by one, with higher-order (interaction) terms removed first. Details of model pruning are described for each model, below.

It is necessary to split the gender response analysis into 3 models: simple real words have whole-word bias values only (Model 1), complex real words have both whole-word and morpheme bias values (Model 2), and pseudowords have morpheme bias values only (Model 3). For whole-word and morpheme gender bias values, woman-biased values are greater than 0, and man-biased values are less than 0. The response variable for Models 1, 2, and 3 is gender response, in terms of log-odds. Summaries for Models 1, 2, and 3 are given on Table 5-3. In figures showing effects from these models, log-odds estimates are transformed and shown in terms of probability of women responses chosen, from never (0) to always (1).

The model for simple real items (Model 1) contains a fixed effect for whole-word gender bias, and random intercepts for each participant, item, and face image. No interactions were pruned from the final model. There is a significant effect of whole-word gender bias ($\beta = 0.42$, SE = 0.076, $z = 5.5$, $p < 0.0001$). Participants are more likely to choose woman responses as the item becomes more woman-biased (see Figure 5-3a). For complex real items (Model 2), the model contains fixed effects for whole-word gender bias, morpheme gender bias, and log word frequency; interaction terms for word gender bias with morpheme gender bias, and for word gender bias with log word frequency; and random intercepts for each participant, morpheme, item, and face. During pruning, the three-way interaction between word gender bias, morpheme gender bias, and log word frequency was removed first; then, the two-way interaction between morpheme gender bias and log word frequency was removed. Word frequency is taken from the COBUILD corpus via CELEX. There is a significant positive effect of word gender bias (β =

0.31, SE = 0.041, *z* = 7.5, *p* < 0.001). Model 2 shows the same pattern for whole words as Model 1 (see Figure 5-3b): responses for real words reflect their gendered statistics, with people more likely to choose woman responses as the complex real items become more woman-biased. There is also a significant positive effect of word frequency (β = 0.092, SE = 0.031, *z* = 3.0, *p* = 0.0027), meaning that higher-frequency words are associated more with women. The main effect of morpheme gender bias is not significant (β = 0.031, SE = 0.12, *z* = 0.26, *p* = 0.79).

There are two significant interactions affecting word gender bias. The interaction of word gender bias with word frequency is significant (β = 0.086, SE = 0.027, *z* = 3.2, *p* = 0.0016), such that the effect of word gender bias on gender response is weaker as frequency decreases. Experience with a word is needed for a gender association effect to obtain, and more experience supports better learning of the association. The interaction of word gender bias with morpheme gender bias is also significant (β = 0.20, SE = 0.080, *z* = 2.5, *p* = 0.013): the influence of word gender bias increases as morphemes are more woman-biased (see Figure 5-5). That is, amongst words containing more woman-biased morphemes, the man-biased whole words were judged to be more man-biased, and the woman-biased words were judged to be more woman-biased. We view this interaction with considerable caution, because it does not arise naturally in any current model of the mental lexicon. Insofar as the effect proves to be reliable, we speculate that it might arise indirectly from the correlation of gender bias with register and topic in the experimental stimuli. Overall, the man-biased morphemes are more typical of formal prose and the woman-biased morphemes are more typical of colloquial language. The gender association of a whole word might be more salient—and thus easier to learn—in conversational contexts than in formal prose.

The model for complex pseudowords (Model 3) contains a fixed effect for morpheme gender bias, and random intercepts for each participant, morpheme, item, and face image. Complex pseudowords were significantly more likely to be associated with women faces when the morpheme group was more woman-biased (β = 0.22, SE = 0.06, z = 3.5, p < 0.001). Figure 5-4 compares the morpheme gender bias effect for complex pseudowords versus complex real words.

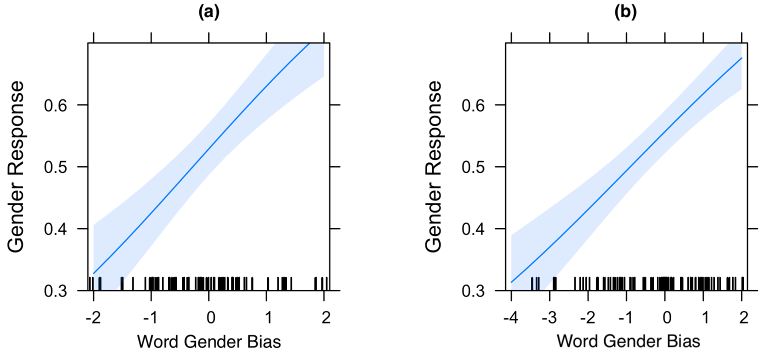


Figure 5-3. Probability of woman gender response correlates with whole-word gender bias for: (a) simple real words (Model 1) and (b) complex real words (Model 2). Shaded regions indicate pointwise 95% confidence intervals for normal distributions.

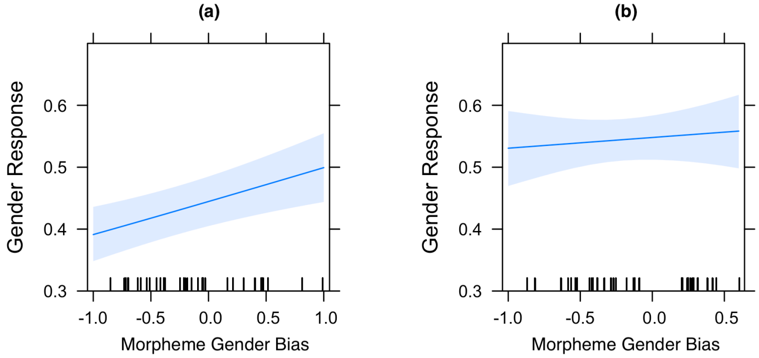


Figure 5-4. Probability of woman gender response correlates with morpheme gender bias for (a) complex pseudowords (Model 3); not for (b) complex real words (Model 2).

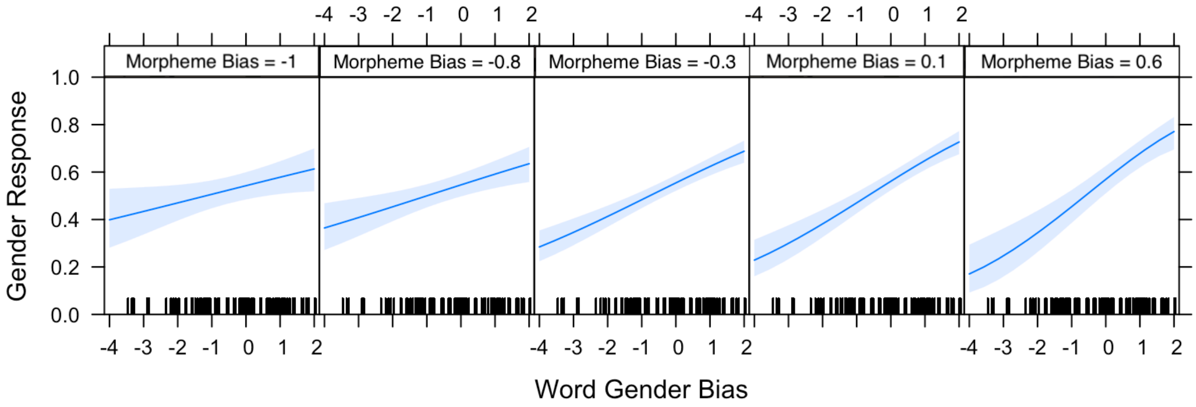


Figure 5-5. Interaction of morpheme gender bias with word gender bias for gender response (Model 2). The word gender effect is stronger for more woman-biased morphemes.

Table 5-3. Regression model summaries for Models 1, 2, and 3.

Model 1:				
gender_response ~ word_gender + (1|workerId) + (1|item) + (1|face_1) + (1|face_2)				
	Estimate	**Std. Error**	**z-value**	**p(>|z|)**
word_gender	*0.42*	*0.076*	*5.5*	*< 0.001*
Model 2:				
gender_response ~ word_gender + morph_gender + log_freq + word_gender:morph_gender + word_gender:log_freq + (1|workerId) + (1|morph) + (1|morph:item) + (1|face_1) + (1|face_2)				
word_gender	*0.31*	*0.041*	*7.5*	*< 0.001*
morph_gender	0.031	0.12	0.26	0.79
log_freq	*0.092*	*0.031*	*3.0*	*0.0027*
word_gender:morph_gender	*0.20*	*0.080*	*2.5*	*0.013*
word_gender:log_freq	*0.086*	*0.027*	*3.2*	*0.0016*
Model 3:				
gender_response ~ morph_gender + (1|workerId) + (1|morph) + (1|morph:item) + (1|face_1) + (1|face_2)				
morph_gender	*0.22*	*0.063*	*3.5*	*< 0.001*

5.3.1. Decomposition accuracy

We now turn to the accuracy of decomposition responses, as a prerequisite to considering the role played by explicit morphological decomposition in the gender associations. A decomposition is judged accurate for complex items if it exactly parses the intended morpheme, and for simple items if the response is ‘no decomposition’. As in the baseline experiment to used validate the paradigm, decomposition accuracy rates are above chance for simple real words, complex real words, and complex pseudowords. Performance on complex words also exceeds a baseline using phonotactic boundary statistics (described below).

Using logistic mixed-effects regression, we evaluate the contribution of phonological information to decomposition accuracy; specifically, we test the possibility that participants are parsing items based on the statistical cues to the presence of a word boundary, without any perception of morphemes. For both phonemic and orthographic bigrams at the location of the expected decomposition, we compared the likelihood of a boundary being present versus absent (taking the ratio of the log likelihoods) (cf. Daland & Pierrehumbert, 2011). The boundary likelihood ratio is derived from orthographic and phonemic bigram statistics in the 10931 CELEX monomorphemes, a list made as discussed in Hay et al. (2004) by hand-checking the lexical entries in the CELEX lexicon (Baayen et al., 1995). Monomorphemes are used so that the bigram statistics accurate reflect words without internal boundaries. Boundary likelihood ratio is defined as the probability that a boundary is present, divided by the probability that a boundary is not present. To estimate the probabilities for bigrams with boundaries, we make the simplifying assumption that words can combine freely.

For morphological decomposition accuracy, Model 4 contains fixed effects for orthographic boundary likelihood ratio, phonemic boundary likelihood ratio, and for lexicality (real word or pseudoword); and random intercepts for each participant, morpheme, and item (see Table 5-4). Model 4 includes response data for complex real words and complex pseudowords only; 7% of pseudoword data and 8% of complex real word data are excluded because boundary ratio statistics are not available for the expected boundary. In Model 4, there is a significant positive effect of lexicality: participants parse pseudowords less accurately than complex real words overall ($\beta = 2.1$, SE = 0.34, $z = 6.2$, $p < 0.0001$) (see Figure 5-6c). This lexicality effect may mean that participants gain a boost from recognizing two morphemes (the stem and the affix) instead of only the affix; or that they have explicit morphological knowledge of the familiar real words. The strength of the orthographic cue is significantly associated with decomposition accuracy ($\beta = 0.41$, SE = 0.093, $z = 4.4$, $p < 0.0001$) (see Figure 5-6a). Participants parse complex pseudowords more accurately when the expected boundary is orthographically likely. The effect of phonemic boundary cue is not significant ($\beta = 0.11$, SE = 0.083, $z = 1.3$, $p = 0.21$) (see Figure 5-6b).

The usefulness of boundary likelihood is limited: the orthographic boundary cue leads to a correct parse in only 37% of complex items (42% for complex real words, 32% for complex pseudowords), but participants gave correct parses for 88% of the complex real words and 65% of the complex pseudowords. This discrepancy means that participants are making significant use of other information for morphological decomposition, such as recognition of morphemes per se.

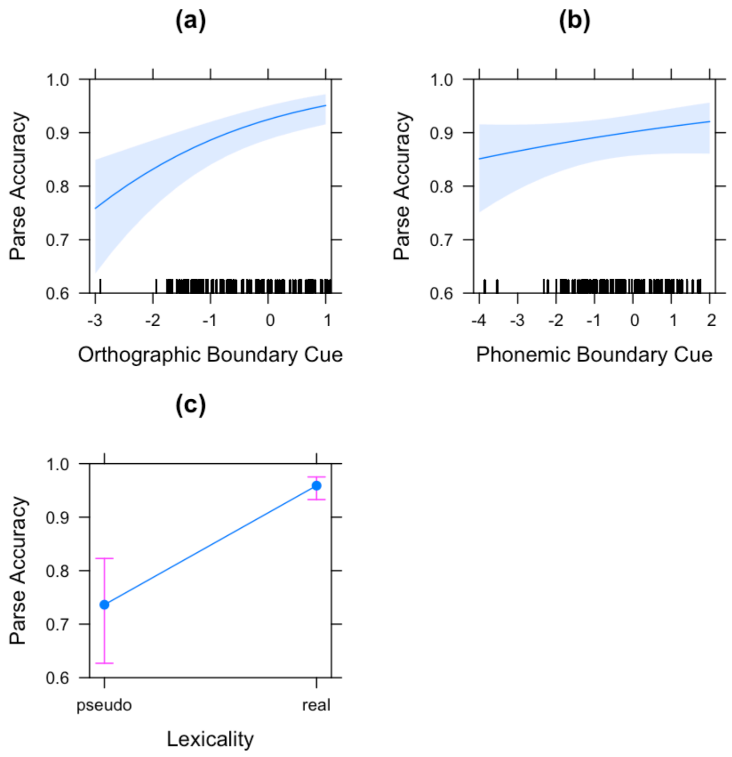


Figure 5-6. Effects of boundary cue and lexicality on decomposition accuracy (Model 4). (a) Probability of morphological parse accuracy correlates with orthographic boundary cue. (b) Phonemic boundary cue is not significant. (c) Accuracy is higher and less variable for complex real words.

Table 5-4. Regression model summary for Model 4.

Model 4:
accurate ~ boundary_ortho + boundary_phono + lexicality + (1|workerId) + (1|morph) + (1|morph:item)

	Estimate	**Std. Error**	**z-value**	**p(>|z|)**
boundary_ortho	*0.41*	*0.093*	*4.4*	*< 0.001*
boundary_phono	0.11	0.083	1.3	0.21
lexicality = ‘real’	*2.1*	*0.34*	*6.2*	*< 0.001*

5.3.2. Gender accuracy in relation to decomposition of pseudowords

As shown in Models 2 and 3, the morpheme gender bias had a significant effect only for pseudowords. How did this effect come about? The decomposition analysis shows that people have moderate success in decomposing pseudowords, and it suggests that they are using both morphological awareness and orthotactic cues. We can now ask whether morphological parsing influences whether participants give the expected gender response. Is gender response accuracy higher when participants made accurate decompositions? Is it higher for word forms that objectively contained stronger cues for the decomposition? Applying to pseudowords only, Model 5 predicts gender response accuracy as a function of relevant stimulus and decomposition-response effects (Table 5-5): fixed effects are included for orthographic boundary cue, phonemic boundary cue, morpheme gender bias magnitude (the absolute value of morpheme gender bias), and parse accuracy (true or false); and random intercepts for participant, item, morpheme, and face image. Actual parse accuracy as well as both boundary cues are included to cover the possible case that participants had poor awareness of morphological parsing information that nonetheless implicitly affected their gender responses. All three cues can be included in a single model because they are not excessively correlated.

There is no significant effect of orthographic boundary cue ($\beta = 0.026$, SE = 0.043, $z = 0.60$, $p = 0.40$) (Figure 5-7a), phonemic boundary cue ($\beta = 0.028$, SE = 0.050, $z = 0.57$, $p = 0.57$), or parse accuracy ($\beta = -0.035$, SE = 0.042, $z = -0.84$, $p = 0.40$). These results suggest that participants' gender response accuracy is not affected by whether they decomposed the items, whether consciously or implicitly. In contrast, the magnitude of morpheme gender bias is a highly significant predictor of gender response accuracy ($\beta = 0.57$, SE = 0.13, $z = 4.5$, $p <$

0.0001) (Figure 5-7b); participants are more likely to choose the gender face that matches the expected morpheme gender as the gender bias increases. Figure 5-8 displays this effect for the specific morphemes in the experiment. The figure also allows us to consider direct sound symbolism for gender as a factor in Model 5. In previous studies, some gender associations have been reported for specific phones, including high/front vowels with feminine and low/back vowels with masculine (Babel & McGuire, 2012; Wu, Klink, & Guo, 2013). Such associations are not apparent in the figure, and a more detailed statistical analysis (for which we omit the details) did not yield any significant results.

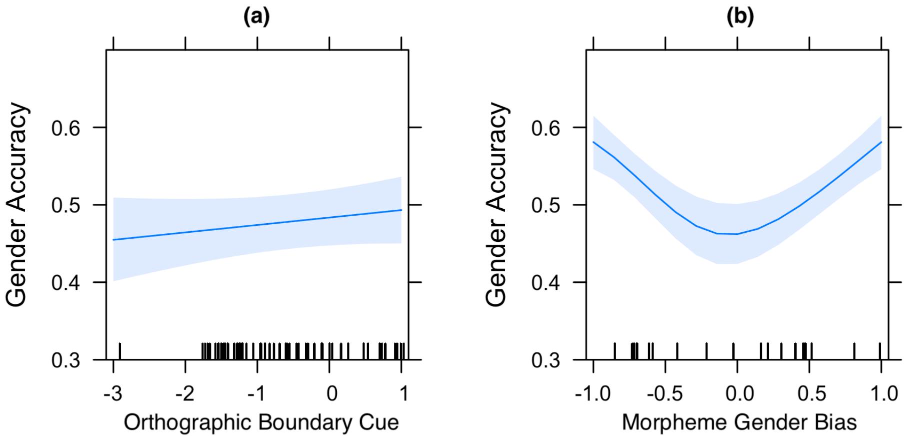
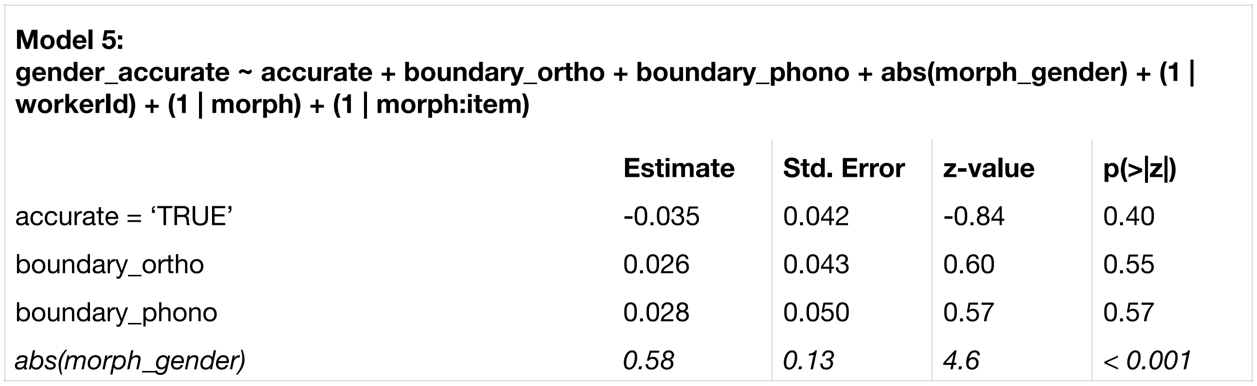


Figure 5-7. Gender accuracy effects of orthographic boundary cue and morpheme gender bias. More extreme morpheme bias increases gender accuracy, but orthographic boundary cue has no effect.

Table 5-5. Regression model summary for Model 5.



Model 5:
gender_accurate ~ accurate + boundary_ortho + boundary_phono + abs(morph_gender) + (1 | workerId) + (1 | morph) + (1 | morph:item)

	Estimate	**Std. Error**	**z-value**	**p(>|z|)**
accurate = ‘TRUE’	-0.035	0.042	-0.84	0.40
boundary_ortho	0.026	0.043	0.60	0.55
boundary_phono	0.028	0.050	0.57	0.57
abs(morph_gender)	*0.58*	*0.13*	*4.6*	*< 0.001*

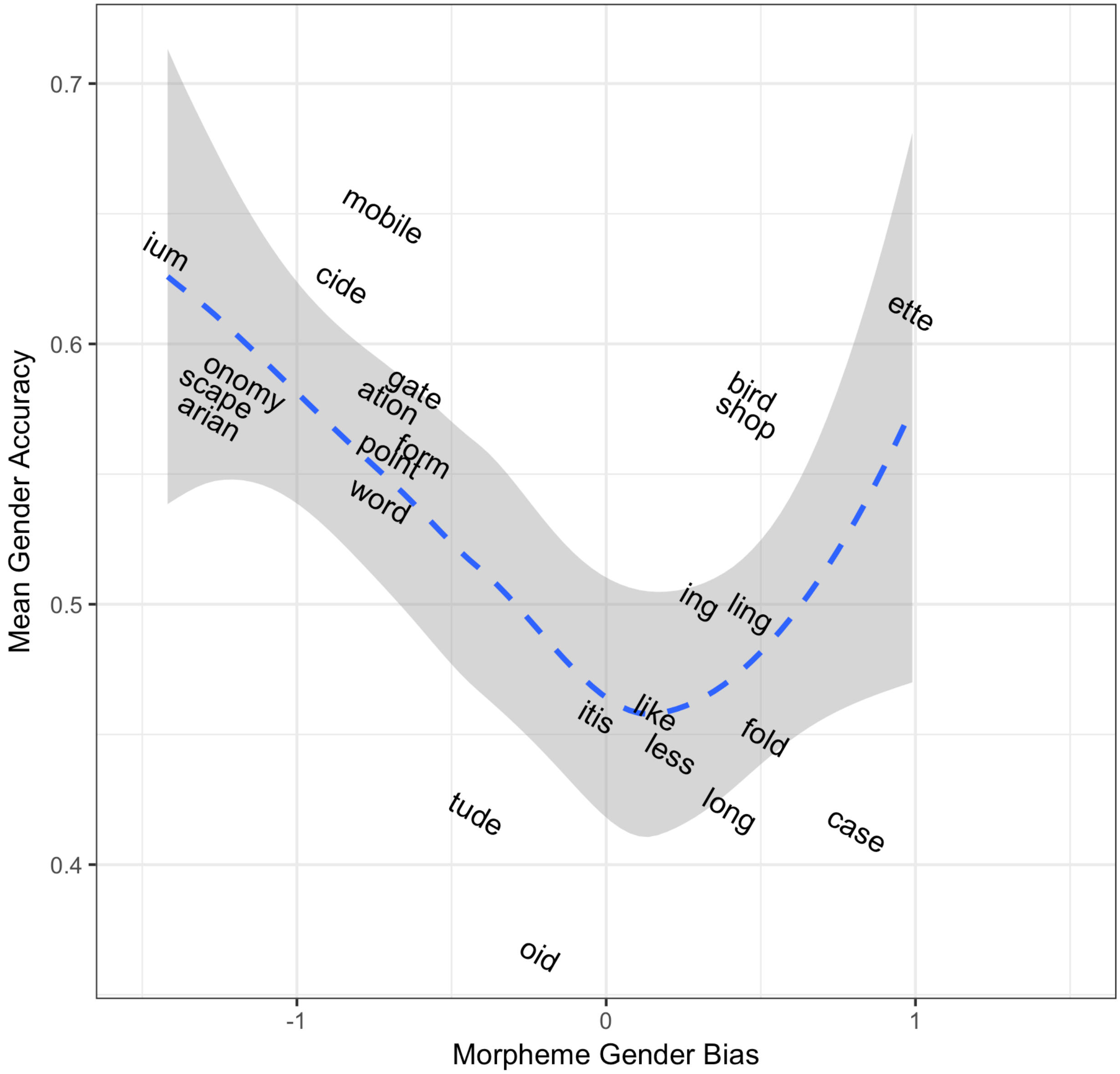


Figure 5-8. Gender accuracy effect of morpheme bias; accuracy is higher for morphemes that have more extreme bias toward either men (values less than zero) or women (values greater than zero). The LOESS fit with 95% confidence interval is shown by the blue dashed line and grey shading.

5.4. Discussion

We found that that speakers have social-indexical associations between words and gender. The effect of word gender bias on gender responses is influenced by two relevant interactions: with word frequency, and with morpheme gender bias. We also found gender associations for morphemes within pseudowords, but not for real words. Our results for whole words extend related findings such as Quina, Wingard, & Bates (1987) and Bucholtz (1999; 2001). For both simple and complex real words, participants reliably matched the gender bias of the whole word as estimated from corpus statistics, suggesting that their intuitions are the result of gendered experience with these words. The interaction of word frequency with gender association for complex words supports this hypothesis: even though all of the real words in this study were rated as highly familiar in the baseline study, the gender association is strongest for the most frequent words, and disappears for the rarest words. This result is reminiscent of the pattern found by Clopper and Pisoni (2004a) for dialectal experience: their listeners were better able to associate speakers with regional dialects when they had more exposure to relevant speech variation. The effect of word frequency on gender association was not significant for the simple word model, which may be explained by the different frequency ranges for simple and complex real words: at 150, the median simple word frequency is higher than 75% of complex real words (for which the median is 37). Additional work with rarer simple word stimuli might show the same disappearance of the gender association effect.

We did not see an effect of morpheme gender bias for complex real word stimuli. Instead, we understand the significant interaction between morpheme gender bias and word gender bias from the perspective of the word gender bias main effect: the effects of word gender bias are more polarized in words containing more woman-biased morphemes. This means that the effect of morpheme gender is not cumulative with the effect of word gender, but enhances the word gender effect. We view this interaction with extreme caution due to both the size and the nature of the effect, which is not predicted by any of the morphological theories considered. We suggested that this pattern, if it is a real one, might arise as an artifact of the different communicative situations in which the various words in our study are encountered: formal and textual, versus casual and face to face.

For the pseudoword stimuli, where whole word knowledge does not exist, we see a main effect of morpheme gender bias on gender responses. Our results show that participants significantly associated pseudowords with gendered faces that matched the gendered corpus statistics for their component morphemes. In addition, participants more frequently matched the predicted gender bias when that bias was stronger. The morpheme gender association effect obtains regardless of participants' accuracy in explicitly parsing the pseudowords, or of the presence of partially useful orthotactic cues to the presence of an internal word boundary. That is, the ability to identify the morpheme itself does not change the gender association effect. These results indicate that people are influenced implicitly by the presence of real morphemes in unknown words. They contrast with the outcomes observed for real words, in which gender associations of morphemes were not significant as a main effect.

Our results present us with questions about the role of knowledge about whole words in contrast to knowledge of word parts. For real words, whole-word knowledge affects gender responses, to the exclusion of word-part knowledge. For pseudowords, only word-part knowledge is available, and it affects gender responses. However, explicit decomposition results do not control or influence this effect, so word-part knowledge appears unrelated to decomposition responses in these tasks. In section 1.2, we summarized four different current theories about morphological representation and processing in the mental lexicon: models related to multiple-route, general analogy, probabilistic rule application, and the NDL. In light of the results presented, we can engage more deeply with these theories and consider how readily they can be extended to encompass the socio-indexical patterns that we found.

We can consider the multiple-route approach as a more sophisticated successor to a simple always-decompose model. An always-decompose model is consistent with the pattern of results The pattern of results is not consistent with such a model, in which people are always recognizing the morphemes per se and retrieving stored gender information about each morpheme. Under such a model, we would have expected judgments of complex real words to reflect the gender associations of the parts (even if gender associations of the whole word also play a role in the people's judgments). We might also have expected the morpheme effect to be stronger when the whole word frequency is low (that is, too low for a whole-word gender association to obtain). However, there was no significant effect of morpheme gender bias for complex real items, regardless of whether the whole word frequency was high or low. Under this decomposition theory, we would further have expected that morphological decomposition would feed into gender judgments for pseudowords: the association with gender would be stronger when the gendered morpheme was identified in the parse. However, this expectation is not fulfilled. A multiple-route approach offers a better, though not complete, explanation of our results. The phonological or orthographic representations of words are recognized either as wholes, or by decomposition into constituent morphemes. Both routes lead to activating the meaning of the whole word, and the question of which route wins is subject to concerns such as the relative frequencies of the whole word and its morphemes. Both whole words and morphemes are present in the mental lexicon, with their own associated information. This information includes the gender-biased experience assumed in the current study: both morphemes and whole words can have gendered associations. If we assume that activation of morphemes should be reflected in the gender response task, it predicts a pattern of results in which pseudowords reflect morpheme gender bias because they are decomposed and processed by parts, as no whole-word option is available. A morpheme gender effect could be observed for the highly-decomposable complex real word stimuli; specifically, those words with lower whole-word frequency and high-frequency morphemes would be decomposed and processed by parts. In contrast, for real words accessed by the whole-word route, the associations of the morphemes would not be activated. However, this prediction depends on the assumption that complex pseudowords are aggressively decomposed, and that many complex real words are not decomposed during lexical access. These assumptions are not well supported by our decomposition results. Our real words were highly decomposable, and people did decompose them (with parsing accuracy above 85%). The pseudowords were less reliably decomposed and the decomposition was not predictive of gender responses. The lack of morpheme gender influence for real words might be explained by the nature of the gender response task, which can be considered to be a slow or high-level task. This means that participants have plenty of time to activate the relevant meaning representation for complex real words, regardless of the route used, so their gender responses reflect only knowledge about that meaning representation. This view is compatible with our pseudoword gender results, though it may require a considerable disconnect between the implicit process of morphological decomposition during lexical access, and the information about morphological decomposability collected with our protocol.

General analogy models as presented in Nosofsky (1988, 1990), Daelemans et al. (2010) and Dawdy-Hesterberg and Pierrehumbert (2014) readily capture our results, with the proviso that analogical forces determine judgments about unknown words, but have only a weak influence on judgments for known words (as claimed in Daland et al., 2007). The mechanism for pseudowords to have apparent gender associations under such an approach is comparable to that proposed in Johnson (2006) for social identity correlates of allophonic variation to emerge. If the gender effect for pseudowords does not come from explicit gender information known about the morphemes themselves, general analogy provides an alternative mechanism: pseudowords inherit the implicit associations of similar real words. In this case, similarity derives in part from sharing a morpheme: the unknown word GLONITIS would be similar to BRONCHITIS, ARTHRITIS, etc., so it would get the gender association of the overall group. This mechanism depends on whole word gender associations, which have been demonstrated previously and which this study replicated. This mechanism is reminiscent of results in Nation and Cocksey's (2009) study on semantic interference. They found semantic interference from sub-word orthographic matches (e.g., HIP in SHIP) when the sub-word took beginning, middle, or final position in the word, or even when the sub-word involved phonological mismatches to the target (e.g., for the letter 'H' in HIP and SHIP). The semantic associations in that experiment clearly result from overall word similarities and not from morphological decomposition.

The Albright and Hayes (2003) MDL model could capture gendered associations of morphemes by probabilistically associating gender with morphological rules, effectively capturing the results for pseudowords that were actually decomposed. For real complex words, it would be necessary to add the proviso that knowledge about the whole word takes priority over predictions from the rule system. While this proviso is not clearly stated in Albright and Hayes (2003), it is independently necessary to explain why real words have highly stable inflectional morphology even if they belong to groups of word forms whose morphology varies. For example, the past tense of KEEP is KEPT and the past tense of BEEP is BEEPED; only a novel word such as FLEEP exhibits instability (e.g., FLEEPED, FLEPT). The challenge for this model would be to explain why the gendered associations for pseudowords were found to be unrelated to the decomposition judgments, or to the cues for decomposition.

The NDL model of Baayen et al. (2013) is very different from the other approaches presented. Under this theory, there are no lexicon representations for whole words or morphemes at the orthographic or phonological level, but only at a semantic level. Instead, phonological sequences (e.g., triphones) are probabilistically associated with meanings. In his study of semantic effects in English compounds, Kuperman (2013) argues that only the fullest form (i.e., the whole word) matters in such a model, so that any semantic influence of the parts is suppressed or not accessed. This account correctly predicts that the gender responses for the simple and complex real word stimuli will reflect only the gender bias of the whole word. It follows that, for pseudowords, the fullest available form is the morpheme, so the gender response would reflect the morpheme, as in our results. In addition, the NDL explains the gender effect for pseudowords without recourse to decomposition, which means that it accords with our results showing no link between explicit decomposition and morpheme gender bias effects.

To summarize, modifying any current model of the lexicon to capture our results involves ensuring that knowledge about whole words takes priority over a compositional analysis, to the extent that such knowledge is available: not only is whole-word knowledge stronger than word-part knowledge, but word-part knowledge is suppressed when whole-word knowledge is available. Given this proviso, which is often motivated independently by the existence of irregular morphological forms, the results are most readily captured by assuming that social-indexical effects in morphology operate through a general analogical mechanism. While morphological parsing is known to be relevant within the phonology and morpho-syntax, such structured processing may be confined to these parts of the linguistic system. These findings leave several avenues to explore: attention should be paid to rarer real words, to more lower-level or faster experimental tasks, and to pseudowords made of only real morphemes. The interaction between word bias and morpheme bias points toward a new stimulus set that controls for word register and socioloinguistic context, with the exciting possibility that communication mode plays an important role in the encoding and association of indexical information with whole words and morphemes.

6. Conclusions

In order to integrate the findings presented in Chapters 2, 3, 4, and 5, we begin here with a brief summary of the results of these studies. The results in Chapter 2 replicate the finding that pseudowords are not categorically atypical (i.e., not wordlike) as words, and instead show gradient typicality ratings for pseudowords which extend beyond simple and monosyllable stimuli, to longer items which may show apparent morphological complexity. When this broad range of pseudowords is considered, we found that the well-known factors of phonotactic probability and lexical neighborhood density have significant, separate contributions to typicality. In addition, the presence of apparent morphemes in these meaningless stimuli had an additional, separate contribution to typicality. These results support theories of morphological processing in which decomposition is not limited to attested or semantically-transparent words. Below, we will consider how these three factors might be treated as different approaches to the broad concept of lexical similarity, and ask if stronger integration of these mechanisms is possible.

In Chapter 3, we found that real words are not categorically typical, in the same way that Chapter 2 replicated the finding that pseudowords are not categorically atypical. We found gradient typicality judgments for real worlds as well as pseudowords in this study, and that typicality judgments for both real words and pseudowords are affected in similar ways by similar factors: phonotactic probability and lexical neighborhood density. There also appeared to be a strong lexicality bias: real words are judged more typical than similar pseudowords. This bias might be explained in terms of lexical frequency, such that high frequency words receive near-ceiling mean ratings, while the pattern for low frequency words suggests that typicality trends down as frequency approaches zero. This pattern suggests that the same mechanisms are involved for typicality judgments for both real words and pseudowords.

The explicit decomposition experiment in Chapter 4 provided evidence that people are very successful at decomposing complex real words and complex pseudowords, with performance that far exceeds what is predicted by a phonotactic baseline. This finding provides a possible mechanism for the morphological effect on pseudoword typicality described in Chapter 2. We suggest that the strong performance observed, even for meaningless pseudowords, implies that people are applying morphological knowledge through a process of morpheme-spotting. Familiarity judgments gathered during the same trials show that this decomposition was not affected by the fact that people almost unanimously understood that the real words were highly familiar, but the pseudowords were highly unfamiliar. We might take this result in comparison to Chapter 2 to suggest that people are correctly differentiating the tasks of judging typicality versus familiarity: they are able to recognize the constituent morpheme, and understand that this makes the word more typical, but not more familiar.

While Chapter 4 showed that people have morphological knowledge and use it for unfamiliar pseudowords, the study in Chapter 5 built on the decomposition results to demonstrate that people also have indexical knowledge related to morphological units. The gender association results in Chapter 5 replicated the finding that people have such associations for whole real words, and presented the new finding that people have gender associations for complex pseudowords, based on their real constituent morphemes. In addition, gender associations for complex real words were not affected by their constituent morphemes, which requires our morphological models to explain the dominance of whole-word knowledge over word-part knowledge in this task.

Together, these findings contribute to three major points of discussion: the connection between typicality judgments and lexical processing effects for real words and pseudowords; the various measures for the concept of lexical similarity; and the way indexical knowledge is perceived, encoded, and accessed. We also connect these findings to consequences for language innovation and change.

6.1. Tasks, typicality, and lexical processing

Based on the results presented in the preceding chapters, we argue for the useful connection between high-level tasks (e.g., typicality judgments) and the low-level tasks commonly used to study lexical processing (e.g., accuracy and latency in lexical decision, naming, shadowing, priming, etc.). These tasks are generally called low-level because they happen over short time-scales, often below the level of consciousness. In contrast, the tasks in Chapters 2 and 3 are high-level because the participants are not under time pressure, and are conscious of what they are doing. We have shown that both of these approaches depend on similar measures, especially frequency, phonotactic probability, lexical neighborhood density, and morphological complexity. However, lexical processing research is nearly synonymous with the kind of low-level tasks mentioned (see Vitevitch & Luce, 2016), and the study of well-formedness is more often concerned with typicality judgments of the kind presented here. Focusing as they do on similar underlying measures, these separate threads could each benefit from their apparently separate methods: gradient typicality reflects the influence of frequency and similarity on perception, and low-level responses can be used to inform models of well-formedness in fine detail. In a similar way, we argue that an integrated treatment of real words and pseudowords is fruitful for understanding both. Both categories are subject to gradient typicality, affected by similar factors. Chapter 3 discusses how theories of gradient well-formedness are improved by addressing gradience for both real words and pseudowords. The results presented in Chapter 3 remind us that words have no a priori status as real words or pseudowords; instead, the results suggest a continuum running from high frequency words, to low frequency, to zero frequency (unknown words or pseudowords). This is an important consideration in light of the fact that individuals vary widely in the size and contents of their personal vocabulary (for words and for morphemes). At the same time, lexical processing studies already recognize the ways that both real words and pseudowords show gradience in accuracy and latency for a variety of tasks. Indeed, the gradience in processing results for real words was noted as part of the motivation for the study in Chapter 3, providing an example for the mutual relevance of typicality and lexical processing studies.

The relevance of task differences is clear for both practical and theoretical concerns. Practically, low-level tasks may require specialized equipment and careful control of the experimental environment, while high-level tasks can be more robust to larger studies with noisier conditions. One trade-off for this is that low-level tasks can reveal fine-grained responses (especially in terms of time-course) that are not apparent in high-level tasks. For example, the pattern of gender association results in Chapter 5 can be compatible with a multiple-route model if activation of morphemes for real complex words simply is not reflected in the high-level gender response task; in this situation, the gender response is based on the whole-word semantics, and the route to activate that representation does not affect the response. In this case, the morpheme may be activated, but the details of the task mean that this activation is not reflected in the response. Kuperman (2013) makes a similar point when he notes that morphemes with negative emotional valence increase response latency (in a low-level task), but those morphemes do not influence a high-level task like asking for the denotation of the whole word.

6.2. Lexical similarity

Many of the presented findings can be tied to the fundamental concept of typicality as lexical similarity, but doing so leaves many loose threads, as this simple idea has been operationalized in a variety of ways. Evidence shows that typicality is affected by phonotactic probability, lexical neighborhood density, and morphological complexity. Lexical frequency also plays a role in typicality, though the distinction between token frequency and type frequency must be included, as it is in some generalized similarity models (e.g., the Generalized Neighborhood Model in Bailey & Hahn, 2001). The issue is that, while these measures are often correlated, they operate over different time-scales for the words. Simple neighborhood approaches are strongly dependent on length, so that short words have many neighbors (even if they differ by a large proportion of phonemes or letters), while longer words may have no neighbors at all; in addition, phonotactically impossible words can have many neighbors. Phonotactic measures are based on short sub-word scales, but they also compare words to the lexicon as a whole; unlike a neighborhood approach, phonotactics cannot capture the similarity relationship between, e.g., BUFFING and BLUFFING. As we see in Chapters 2, 4, and 5, morphological analysis brings a different though still incomplete perspective, operating over potentially long sub-word scales: BLACKBIRD, BLUEBIRD, and MOCKINGBIRD are similar for people by virtue of their shared morpheme, but they differ in terms of neighbors and phonotactic probability.

This set of related but different similarity approaches suggests that a suitably sophisticated integrative model might be superior, jointly considering the empirical similarity effects of these factors. However, this task is outside the scope of this dissertation, and appears to be a difficult one. There are independent motivations for the separate operation of the different factors mentioned, suggesting that integration is not straightforward. Phonotactic effects, which imply knowledge that is abstracted from the lexicon as a whole, are important in studies of many aspects of perception and production, including second language acquisition, speech errors, and accentedness. In fact, phonotactic information might be considered to precede the lexicon itself, in phonotactic models of infant learning that make early use of phrase boundary phonotactics (Daland & Pierrehumbert, 2011); and phonotactic accounts of segmentation are not explicitly related to lexical similarity. Neighborhood and word-network competition effects provide unique explanations for patterns in phonological and semantic processing, and the results in Chapters 4 and 5 seem to be best explained by morphological knowledge. The decomposition results in Chapters 4 and 5 show that people are very capable of recognizing morphemes (even in pseudowords) while still correctly ignoring spurious form-only matches in real words, but the disconnect between decomposition and gender association in Chapter 5 suggests that the gender effect could be based on pseudoword similarity to morpheme-based clusters of real words by an analogical process. A combined system would need to satisfy many observed patterns, such as the tight relationship between morpheme-based clusters (e.g., BLACKBIRD, BLUEBIRD, and MOCKINGBIRD), but also the typicality ‘veto power’ shown for rare and impossible phonotactic units (e.g., for BNICK). At this point, similarity between words appears to operate over distinct types of links (phonological, morphological, semantic, indexical), and similarity over the lexicon as a whole is best captured by phonotactic approaches.

6.3. Indexical information and the lexicon

The results in Chapter 5 show that indexical information (specifically, gender association) is present within the lexicon. The discussion in Chapter 5 considers some possible structures that can explain the pattern of results reported: while it seems clear that gender information is stored with whole words, there are different ways to derive the morpheme gender bias effect observed for pseudowords. One way is that gender information is stored directly with morphemes, as for whole words. If this is the case, such information is suppressed when whole-word gender information is available, and people must be perceiving morphemes even when their explicit decomposition responses do not reflect this. These requirements are not so far-fetched, given the many things that might happen during processing that are not always reflected in high-level tasks. Alternatively, it could be that morphemes do not have such information directly associated with them. Instead, the pseudoword morpheme effect could result from the influence of a cluster of real words sharing the same morpheme (so that the morpheme effectively carries the average gender bias of the whole cluster). This mechanism requires that information from morpheme-related clusters are available during processing. Regardless of the mechanism by which the whole-word and morpheme gender effects arise, these results show that gender information is present. This leads us to the question of how gender information got there.

Information about gender associations might be stored within the lexicon in various ways, but this information comes from experience. Language is a primarily social phenomenon, so indexical information is often salient in language use. In face to face interactions, it can be impossible to avoid making guesses about the gender, race, class, or age of interlocutors (even if these guesses may be inaccurate), just as it can be difficult to ignore other aspects of the interactional context (e.g., location, activities, nearby objects). We need not even assume that these contextual details are highly salient to influence language in use: Hay and Drager (2010) showed that the mere presence of community-associated stuffed toys influenced accent perception. This sets a low bar for socially-relevant contextual information to be perceived and integrated into linguistic processing. At the same time, we can draw on the suggestion from Kuperman (2013) that attentional focus is subject to strong influences from socially-relevant information; Kuperman raises the argument that emotionally-negative morphemes capture attention during lexical processing because of their general psychological importance. The sociolinguistic literature makes it clear that social concerns (e.g., gender) are of great psychological importance, and indeed are directly relevant to language use. Sociolinguistic lexical processing studies show significant social effects of accent, dialect, and identity; these factors matter in low-level tasks, while still being tightly linked to semantic and indexical considerations. For example, a conversation between a parent, their child, and an adult friend would involve very different lexical and phonetic considerations than a conversation between only the parent and adult friend. Some of the results in Chapter 5 suggest that modality also plays an important role in the perception and encoding of indexical information: we proposed that words which most often appear in formal texts (instead of informal face to face conversations) have weaker gender associations. This might be because textual language experience lacks much of the context discussed above: visual and aural cues to gender (as well as race, class, age, etc.).

6.4. Language innovation and change

In the natural social use of language, high-level and low-level tasks proceed in continuous parallel functioning, as people are not only recognizing words in sentences, but also attending to meaning, indexical information, and ad hoc language innovation. At any moment, a person might be confronted with an unknown word, which they need to consider in terms of typicality (including morphology, phonotactics, etc.), but also within the social and environmental context (who is speaking, to whom, where, etc.). This is the context in which language innovation and language change occurs, so all of the factors mentioned must be jointly considered. We can consider a minimal model of language innovation, in which new words must be created, shared, and perceived, starting a cycle of adoption and transmission between language speaker-listeners. The mechanism depends on the crucial step when a listener is perceiving and judging a novel word. This lexical processing is known to be affected by factors such as phonotactic probability and lexical neighborhood density, which are correlated with gradient typicality. These factors of typicality are based on the listener's lexical knowledge as derived from linguistic experience. Experience also gives speaker-listeners rich contextual knowledge, including indexical information about the social context words are used in. This indexical knowledge in turn induces social associations for potential new words. Chapter 5 showed that the processing of pseudowords is affected by indexical associations in morphology, in addition to formal similarity as measured by phonotactic probability or neighborhood density. This suggests that is necessary to integrate knowledge from the sociolinguistic literature, recognizing that lexical processing takes place within the daily social lives of individual speaker-listeners. The single axis of acceptability may instead be complicated into a set of sociallycontextualized judgments: is a given potential word typical in conjunction with the observation that the speaker is a woman, and a member of a specific community? Does a listener consider an encountered word typical given that the tone of the discourse is erudite jargon, or fashionable slang? It may be impossible to abstract away from these concerns; instead, they provide opportunities for richer understanding of linguistic cognition. Experimental work is needed to explore the effects of other aspects of speaker-listeners' indexical information, including categories well-known to sociolinguistics: age, class, community membership, professional sphere, and other factors of identity. At the same time, the findings presented in this dissertation must be tested with more complex statistical models, in conjunction with other experimental tasks. In particular, integration requires that the results presented be verified in those paradigms commonly used for lexical processing.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/ experimental study. *Cognition*, **90**(2), 119-161. doi:10.1016/S0010-0277(03)00146-X

Alegre, M., & Gordon, P. (1999). Rule-Based versus Associative Processes in Derivational Morphology. *Brain and Language*, **68**, 347-365.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, **56**(3), 329-347. doi:10.1177/0023830913484896

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

Babel, M., & McGuire, G. (2012). Gendered sound symbolism and masking effects in speech processing. In *Proceedings of Interspeech-2011* (pp. 2001-2004). International Speech Communication Association.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1), 1-48.

Bauer, L. (2001). Morphological productivity. Cambridge University Press.

Beckner, C., Rácz, P., Hay, J., Brandstetter, J., & Bartneck, C. (2016). Participants conform to humans but not to humanoid robots in an English past tense formation task. *Journal of Language and Social Psychology*, **35**(2), 158-179. doi:10.1177/0261927X15584682

Berko, J. (1958). *The child's learning of English morphology* (Doctoral dissertation, Radcliffe College.).

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Boulis, C., & Ostendorf, M. (2005, June). A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 435-442). Association for Computational Linguistics. doi:10.3115/1219840.1219894

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, **61**(1-2), 93-125.

Bucholtz, M. (1999). "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, **28**(2), 203-223.

Bucholtz, M. (2001). The whiteness of nerds: Superstandard English and racial markedness. *Journal of Linguistic Anthropology*, **11**(1), 84-100. doi:10.1525/jlin.2001.11.1.84

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, **10**(5), 425-455. doi: 10.1080/01690969508407111

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 23, No. 1, pp. 378-388).

Clopper, C. G., & Pisoni, D. B. (2004a). Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change*, **16**(1), 31-48. doi:10.1017/S0954394504161036

Clopper, C. G., & Pisoni, D. B. (2004b). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, **32**(1), 111-140. doi:10.1016/S0095-4470(03)00009-3

Clopper, C. G., Tamati, T. N., & Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. *Journal of Phonetics*, **58**, 87-103. doi:10.1016/j.wocn.2016.06.002

Coleman, J., & Pierrehumbert, J. B. (1997). Stochastic Phonological Grammars and Acceptability. In *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, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed), *Attention and Performance VI* (535-555). Hillsdale, NJ: Erlbaum.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2010). *TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide*. ILK Research Group Technical Report Series no. 10-01.

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

Daland, R., & Pierrehumbert, J. B. (2011). Learning Diphone−Based Segmentation. *Cognitive Science*, **35**(1), 119-155. doi:10.1111/j.1551-6709.2010.01160.x

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, **28**(2), 197-234.

Daland, R., Sims, A. D., & Pierrehumbert, J. (2007, June). Much ado about nothing: A social network model of Russian paradigmatic gaps. In A*nnual Meeting - Association for Computational Linguistics* (Vol. 45, No. 1, p. 936).

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

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1536), 3773-3800.

Dawdy-Hesterberg, L. G., & Pierrehumbert, J. B. (2014). Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience*, **29**(10), 1268-1282. doi:10.1080/23273798.2014.899377

Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics*, **25**, 223-238.

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 488-499.

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, **1**(3), 245-267. doi:10.1017/S095439450000017X

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, **12**(4), 453-476. doi:10.1111/j.1467-9841.2008.00374.x

Edwards, J., Beckman, M. E., & Munson, B. (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.

Ernestus, M. T. C., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, **79**(1), 5-38.

Feldman, L. B., Milin, P., Cho, K. W., del Prado Martín, F. M., & O'Connor, P. A. (2015). Must analysis of meaning follow analysis of form? A time course analysis. *Frontiers in Human Neuroscience*, **9**.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**(4), 680-698.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, **34**(4), 409-438. doi:10.1016/j.wocn.2005.08.002

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

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

Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in English and Arabic. *Typological Studies in Language*, **45**, 159-180.

German, J. S., Carlson, K., & Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics*, **41**(3), 228-248. doi: 10.1016/j.wocn.2013.03.001

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (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, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, **29**(2), 228-244.

Hahn, S., Vozila, P., & Bisani, M. (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, U., & Bailey, T. M. (2005). What makes words sound similar?. *Cognition*, **97**(3), 227-267.

Haspelmath, M., & Sims, A. (2013). *Understanding morphology (2nd ed.).* Routledge, Taylor Francis Group. London.

Hay, J. (2002). From speech perception to morphology: Affix ordering revisited. *Language*, **78**(3), 527-555. doi:10.1353/lan.2002.0159

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, **9**(7), 342-348. doi:10.1016/j.tics.2005.04.002

Hay, J., & Baayen, H. (2001). Parsing and productivity. In *Yearbook of Morphology 2001* (pp. 203-235). Springer Netherlands.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, **48**(4), 865–892. doi:10.1515/ling.2010.027

Hay, J., Pierrehumbert, J. B., & Beckman, M. E. (2004). Speech Perception, Well-Formedness, and the Statistics of the Lexicon. *Papers in Laboratory Phonology VI*, Cambridge University Press, Cambridge UK, 58-74.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, **34**(4), 458-484. doi:10.1016/j.wocn. 2005.10.001

Hayes, B. (2000). Gradient well-formedness in Optimality Theory. *Optimality Theory: Phonology, syntax, and acquisition*, **88**, 120.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, **44**(1), 45-75.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, **39**(3), 379-440.

Hayes, B., Zuraw, K., Siptár, P., & Londe, Z. C. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, **85**, 822-863.

Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, **33**, 61- 83.

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

Imai, M., Gentner, D., & Uchida, N. (1994). Children’s theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, **9**(1), 45-75.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, **34**(4), 485-499. doi:10.1016/j.wocn.2005.08.004

Johnston, L. H., & Kapatsinski, V. (2011). In the beginning there were the weird: A phonotactic novelty preference in adult word learning. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 978-981). City University of Hong Kong, Hong Kong, China.

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

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

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, **42**(3), 627-633.

Komarova, N. L., & Nowak, M. A. (2001). The evolutionary dynamics of the lexical matrix. *Bulletin of Mathematical Biology*, **63**(3), 451-484. doi:10.1006/bulm.2000.0222

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. *Frontiers in Psychology*, **4**, 203.

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

Labov, W. (2001). Principles of linguistic change. Vol. 2. Social factors. Oxford, UK: Blackwell.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, **2**(1), 45-79. doi:10.1017/S0047404500000051

Lehtonen, M., Monahan, P. J., & Poeppel, D. (2011). Evidence for Early Morphological Decomposition: Combining Masked Priming with Magnetoencephalography. *Journal of Cognitive Neuroscience*, **23**(11), 3366-3379.

Libben, G., Derwing, B. L., & Almeida, R. G. (1999). Ambiguous Novel Compounds and Models of Morphological Parsing. *Brain and Language*, **68**, 378—386.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound facture: The role of semantic transparency and morphological headedness. *Brain and Language*, **84**, 50-64.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *Bioscience*, **51**, 341-352.

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

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

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

Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, **7**(8): e43230. doi:10.1371/journal.pone.0043230

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, **101**(1), 3.

Martin, A. (2011). Grammars leak: Modeling how phonotactic generalizations interact with the grammar. *Language*, **87**(4), 751-770.

Martin, J. H., & Jurafsky, D. (2000). *Speech and language processing*. Prentice-Hall.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, **90**, 227–234.

Mihalcea, R., & Garimella, A. (2016). What men say, what women hear: Finding gender-specific meaning shades. *IEEE Intelligent Systems*, **31**(4), 62-67. doi:10.1109/MIS.2016.71

Muncer, S. J., Knight, D., & Adams, J. W. (2014). Bigram frequency, number of syllables and morphemes and their effects on lexical decision and word naming. *Journal of Psycholinguistic Research*, **43**(3), 241-254.

Nation, K., & Cocksey, J. (2009). Beginning readers activate semantics from sub-word orthography. *Cognition*, **110**(2), 273-278. doi:10.1016/j.cognition.2008.11.004

Needle, J. M., Pierrehumbert, J. B., & Hay, J. B. (in review). Phonotactic and morphological contributions to wordlikeness: Analysis of a new English pseudoword lexicon.

Nevalainen, T., Raumolin-Brunberg, H., & Mannila, H. (2011). The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change*, **23**(1), 1-43. doi:10.1017/S0954394510000207

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**(1), 54-65. doi: 10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, **34**(4), 393-418. doi: 10.1016/0022-2496(90)90020-A

Novak, J. R., Minematsu, N., & Hirose, K. (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, J., Yang, D., Minematsu, N., & Hirose, K. (2011). *Initial and Evaluations of an Open Source WFST-based Phoneticizer*. The University of Tokyo, Tokyo Institute of Technology.

Paice, C. D. (1990). Another Stemmer. *ACM SIGIR Forum*, **24.3**, 56-61.

Pierrehumbert, J. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In Keating (ed.) *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, Cambridge: CUP, pp. 168-188.

Pierrehumbert, J. 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.

Pierrehumbert, J. B. (2006a). The next toolkit. *Journal of Phonetics*, **34**(4), 516-530.

Pierrehumbert, J. B. (2006b). The statistical basis of an unnatural alternation. In L. Goldstein, D. H. Whalen, & C. Best (eds.), *Laboratory Phonology VIII, Varieties of Phonological Competence*. Mouton de Gruyter, Berlin, 81-107.

Pierrehumbert, J. B. (2016). Phonological representation: beyond abstract versus episodic. *Annual Review of Linguistics*, **2**, 33-52. doi:10.1146/annurev-linguistics-030514-125050

Plag, I., Dalton-Puffer, C., & Baayen, H. (1999). Morphological productivity across speech and writing. *English Language & Linguistics*, **3**(2), 209-228.

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, **18**(1), 10-30. doi:10.1177/0261927X99018001002

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar*. Cambridge, MA: Blackwell.

Quina, K., Wingard, J. A., & Bates, H. G. (1987). Language style and gender stereotypes in person perception. *Psychology of Women Quarterly*, **11**(1), 111-122. doi:10.1111/j.1471-6402.1987.tb00778.x

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, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social Salience Discriminates Learnability of Contextual Cues in an Artificial Language. *Frontiers in Psychology*, **8**. doi:10.3389/fpsyg.2017.00051

Rácz, P., Pierrehumbert, J. B., Hay, J. B., & Papp, V. (2015). Morphological emergence. *The Handbook of Language Emergence*, 123-146. doi:10.1002/9781118346136.ch5

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

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, **55A**, 1339-1362.

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

Rickford, J. R. & Rickford, R. J. (2000) *Spoken Soul: The Story of Black English*. John Wiley and Sons.

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

Schumacher, R. A., Pierrehumbert J. B., & LaShell, P. (2014). Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language. In *Proceedings of the 36th Meeting of the Cognitive Science Society* (CogSci2014), Austin, TX: Cognitive Science Society.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (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.

Social Security Administration. (2016, May 5). Top Names Over The Last 100 Years. Retrieved from https://www.ssa.gov/OACT/babynames/decades/century.html.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

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

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

Stump, G. T. (2017, July). The significance of “potentiation” for morphological theory. Presented at the *Morphological Typology And Linguistic Cognition workshop at the 2017 Linguistic Institute*, Lexington, KY.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, **60**(4), 487-501. doi: 10.1016/j.jml.2009.01.001

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

Taft, M., & Forster, K. I. (1975). Lexical Storage and Retrieval of Prefixed Words. *Journal of Verbal Learning and Verbal Behavior*, **14**, 638-647.

The British National Corpus, Version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

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

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

Vitevitch, M. S., & Luce, P. A. (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, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, **2**, 75-94.

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

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Language and Speech*, **40**(1), 47-62.

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

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, **45**(4), 1191-1207.

Wu, L., Klink, R. R., & Guo, J. (2013). Creating gender brand personality with brand names: The effects of phonetic symbolism. *Journal of Marketing Theory and Practice*, **21**(3), 319-330. doi:10.2753/MTP1069-6679210306

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

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Language*, **83**, 277-316.

Appendix 1: 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.

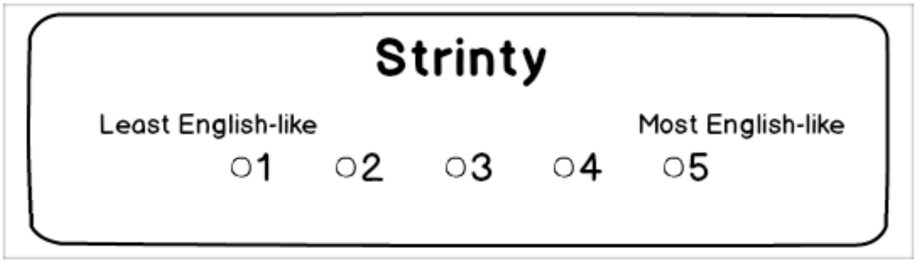
Focus on how the word sounds, not on the spelling.

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:



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 2: 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.

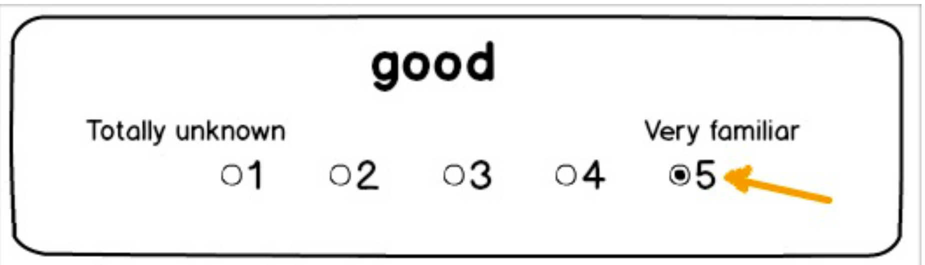
4 = familiar; I have definitely seen or heard this word, I think I know what it means, and I would use this word.

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:



good

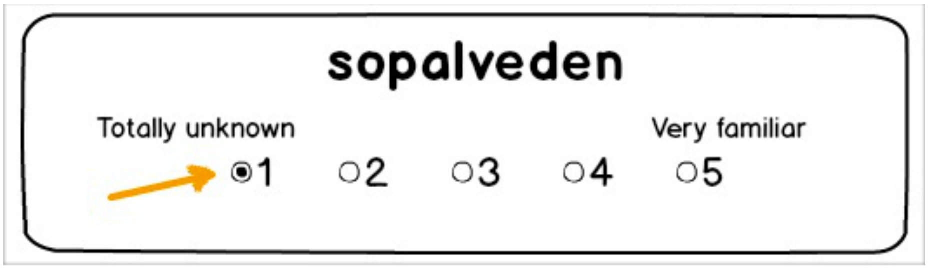
Totally unknown   Very familiar

○1 ○2 ○3 ○4 ◉5

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:



sopalveden

Totally unknown Very familiar

◉1 ○2 ○3 ○4 ○5

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

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

Appendix 3: 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 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. Appendix 4 provides examples of errors that the stability filter removes. In an initial batch of 120000 candidate items, 38073 items passed the filter (32%).

Appendix 4: Examples of P2G/G2P instability results.

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

Appendix 5: Extended Model Comparison Statistics

Drop-1 model comparison statistics for Baseline models. Unavailable comparisons indicated by '-'.

Baseline 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

Appendix 6: List of experiment stimuli, “baseline_appendix_6.txt”.

The file contains the complete list of stimuli for this study. The format is tab-delimited plain text, with columns for: item (“item”), morpheme group (“morph”), real or pseudoword status (“lexicality”), simple or complex item (“complexity”), compound or suffixation (“structure”), and orthographic boundary ratio (“boundary_cue_ortho”).

Appendix 7: List of experiment stimuli, “gender_appendix_7.txt”.

The file contains the complete list of stimuli for this study. The format is tab-delimited plain text, with columns for: item (“item”), morpheme group (“morph”), real or pseudoword status (“lexicality”), simple or complex item (“complexity”), compound or suffixation (“structure”), morpheme gender bias category (“morph_gender”), whole word gender bias category (“token_gender”), and orthographic boundary ratio (“boundary_cue_ortho”).