Running head: LEXICAL DYNAMICS

# Lexical Dynamics for Low-Frequency Complex Words: A Regression Study Across Tasks and Modalities

R. Harald Baayen,[1] Lee H. Wurm,[2] and Joanna Aycock[2]

[1]Radboud University Nijmegen and

Max Planck Institute for Psycholinguistics

[2]Wayne State University

Abstract

In this study we examine the word recognition process for low-frequency morphologically complex words. One goal of the study was to replicate and expand upon findings suggesting facilitative effects of morphological relatives of a target word. A second goal was to demonstrate the need for a reinterpretation of root and surface frequency effects, which traditionally have been taken as indicators of parsing-based and memory-driven processing, respectively. In a first study, we used the same stimuli across auditory and visual lexical decision and naming. Mixed-effects statistical modeling revealed that surface frequency was a robust predictor of RTs even in the very low end of the distribution, but root frequency was not. Also, the nature of the similarity between a target and its lexical competitors is crucial. Measures gauging the influence of morphological relatives of the target were facilitative, while measures gauging the influence of words related only in form were inhibitory. A second study analysing data from the English Lexicon Project, for a large sample of words from across the full frequency range, supports these conclusions. An information-theoretical analysis of root and surface frequency explains why surface frequency must be the most important predictor, with only a marginal role for root frequency.

Lexical Dynamics for Low-Frequency Complex Words:

A Regression Study across Tasks and Modalities

This study investigates the dynamics of lexical processing for low-frequency morphologically complex words. It is well-known that lexical processing in visual and auditory comprehension is not simply a matter of holistically matching the stimulus against templates stored in memory. Processing is modulated by words that provide partial matches with the input, both in visual comprehension (Coltheart, Davelaar, Jonasson, & Besner, 1977; Bowers, Davis, & Henley, 2005) and in auditory comprehension (Marslen-Wilson, 1984, 1989; Zwitserlood, 1989). In what follows, we refer to partially matching words as lexical candidates. The aim of the present paper is to contribute to our understanding of the probabilistic dynamics governing the selection of the appropriate candidate from sets of lexical candidates for low-frequency morphologically complex words. More specifically, our interest focuses on enhancing our understanding of established measures for storage and computation, and on replicating recently proposed and developing new quantitative measures gauging the effect on lexical processing of various kinds of lexical candidates.

Of all possible partially matching lexical candidates, those words that are morphologically related to the target word have attracted the least research. For instance, in the large regression study of Balota, Cortese, Sergent-Marschall, Spieler and Yap (2004), no morphological variables were considered. Furthermore, some central concepts in the processing literature, such as neighborhood density and uniqueness points, are very limited with respect to their application to complex words. The count of lexical neighbors, the words at a Hamming distance of 1 from the target as formulated by Coltheart et al. (1977) works well only for short words. Most longer words, and therefore most derived words and compounds, have no neighbors under this definition because as word length is increased, lexical similarity neighborhoods become extremely sparse very quickly. For longer words, therefore, the standard operationalization of neighborhood density becomes vacuous. Turning to auditory comprehension, we note that the construct of the uniqueness point in cohort theory (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1984) is calculated on lexical lists from which suffixed continuations of the target word have been removed. Without this restriction, a word's uniqueness point would shift past word offset for any word with morphological continuation forms, and the concept would become useless. But with the restriction, the construct of the uniqueness point leaves unanswered what the effect might be on lexical processing of a word's morphological continuation forms.

A great many studies have addressed morphological representation and processing by means of the priming paradigm. These studies (e.g., Stanners, Neiser, Hernon & Hall, 1979; Marslen-Wilson, Tyler, Waksler, & Older, 1994; Feldman, 2000) have clarified that priming effects for morphologically related words tend to be larger than for words that share either only form or only meaning. Spuriously embedded words such as *tar* in *guitar*, have attracted considerable interest both in visual and auditory research.

In auditory comprehension, the meanings of initial embeddings like *boy* in *boycott* become temporarily available (e.g., Prather & Swinney, 1977). The evidence for non-initial embeddings (*tar* in *guitar*) is less unequivocal (compare, e.g., Luce & Cluff, 1998; Luce & Lyons, 1999; Shillcock, 1990; Vroomen & de Gelder, 1997; with, e.g., Pitt, 1994; see also McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995). Recent work has emphasized the importance in the speech signal of subphonemic cues which differentiate between free-standing *ham* and embedded *ham* in *hamster* (e.g., Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003; Kemps, Wurm, Ernestus, Schreuder, & Baayen, 2005).

In visual word recognition, non-morphological embeddings like *corn* in *corner* and *broth* in *brothel* are currently under intense scrutiny (Seidenberg & Gonnerman, 2000; Gonnerman & Andersen, 2002; Rastle, Davis, & New, 2004; Diependaele, Sandra, & Grainger, 2005). The key issue is whether the special status of *-er* as a suffix of English is crucial for explaining potentially different magnitudes of priming for *corn* and *corner* compared to *broth* and *brothel.*

Of particular interest is a recent study by Bowers et al. (2005). Using the semantic categorization task, they observed that people are slower to reject *chat* as an article of clothing than they are to reject it as a human body part. The authors argued that this is because the word *hat* partially matches the target word *chat*. Upon reading *chat* the meaning of *hat* is co-activated, and pulls the reader toward a "Yes" response for the clothing category. These data argue against a strictly modular cascaded architecture for lexical processing in which form-driven processes culminate in the selection of a single candidate, the meaning of which would subsequently be activated in semantic memory. Instead, from the start, the meanings of many partially matching lexical candidates all seem to be taken into consideration during visual comprehension, just as in auditory comprehension the meanings of lexical competitors have been found to be immediately available (Zwitserlood, 1989; cf. also Salverda et al., 2003). In auditory comprehension, partial matches are defined primarily by the dynamics of cohort reduction, dynamics which we now know are tuned to fine phonetic detail in the speech signal. In visual comprehension, only linear order needs to be preserved: Bowers et al. (2005) have shown that non-contiguous but sequentially ordered matches (*sip* for *ship, procuring* for *pouring*) are considered along with contiguous substrings (*hat* in *chat*). From this body of evidence we infer that there are no compelling reasons to suppose that specifically morphological pattern matching processes would be required for understanding complex words. Instead, a word's morphological constituents are made available on the basis of general pattern matching, wherever and whenever there is some evidence for their presence in the input. As a consequence, the candidate set from which the target has to be selected is large and noisy.

In spite of the complexity of selecting the target from its competitor set, readers and listeners seldom become aware of competitors such as *corn* in *corner* or *hat* in *chat*. Clearly, lexical competitors are weeded out very effectively. The hypothesis explored in the present study is that this is accomplished predominantly on the basis of two kinds of information in lexical memory: Syntagmatic information on the one hand, and paradigmatic information on the other (cf. De Saussure, 1916 (1966); Hay & Baayen, 2005). Syntagmatic information in memory is contextual in nature, and concerns our knowledge that *hat* is an improbable constituent when preceded by a *c,* but a probable constituent when followed by *ful*, *less*, or *band*. Paradigmatic information in memory concerns our knowledge about the morphological networks in which constituents are embedded, e.g., that *hat* is a constituent in *hatful*, *hatless, hatband, hatpin*, and *hatter,* and that *ful* is a constituent in words such as *thankful, armful, blissful, bottleful,* etc.

The importance of the paradigmatic relations between morphologically complex words is emerging from work on the morphological family size effect. A word's morphological family is defined as the set of compounds and derived words in which that word appears as a constituent (*hatful*, *hatless, hatband, hatpin*, and *hatter* for *hat)*. The morphological family size is the cardinality of this set (5). This measure has been found to correlate negatively with visual lexical decision times (Baayen, Feldman, & Schreuder, 2006; Bertram, Baayen, & Schreuder, 2000; de Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002; de Jong, Schreuder, & Baayen, 2003; Dijkstra, Moscoso del Prado Martín, Schulpen, Schreuder, & Baayen, 2005; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004a; Schreuder & Baayen, 1997), and also with auditory lexical decision times (Wurm, Ernestus, Schreuder, & Baayen, 2006). The first goal of the present study is to provide further evidence that a word's paradigmatic entanglement facilitates selection from the set of lexical candidates. We do so by contrasting consistent facilitation for measures quantifying the probability weight of paradigmatically related lexical candidates with consistent inhibition for measures gauging the probability weight of non-morphological lexical candidates.

The second goal of the present study is to show that the traditional interpretation of root and surface frequency effects is incorrect. Frequency effects for surface forms and frequency effects for roots have traditionally been interpreted as evidence for memory-driven retrieval and for parsing-mediated comprehension respectively. (In what follows, we use root frequency to refer to the lemma frequency of the base word of a complex form.)

A large body of literature is concerned with the level at which the experimentally observed frequency effects arise. One theory holds that complex words are initially parsed into their constituents, and that these constituents subsequently provide access to a representation for the word itself (Taft 1994, 2004; Taft & Forster, 1975). Frequency effects for roots would bear witness to the activation of the root constituents, and frequency effects for full forms would be the hallmark of subsequent activation of the words' own representations. Another possibility that has been considered primarily for inflected words is that they are initially processed through full-form driven template matching, and that their full-form access representations subsequently activate separate representations for roots and inflectional categories, representations which in turn feed the compositional processes of semantic and syntactic interpretation (Caramazza, Laudanna, & Romani, 1988). In this approach, the allocation of frequency effects to discrete ordered levels of processing is reversed. A range of intermediate parallel dual route models has also been proposed (Baayen, Dijkstra, & Schreuder, 1997; Bergman, Hudson, & Eling, 1988; Laudanna, Cermele, & Caramazza, 1997; Schreuder & Baayen, 1995) in which lexical access typically proceeds in parallel through access representations for the constituents on the one hand, and through access representations for full forms on the other hand.

Traditional theories with discrete representational units and a 'magic moment' of word recognition separating 'prelexical' and 'postlexical' processes contrast with connectionist proposals such as the triangle model (Joanisse & Seidenberg, 1999; Seidenberg & Gonnerman, 2000), in which memory is superpositional and in which banks of orthographic, phonological and semantic units interact. In the triangle model, a strong surface-frequency effect is indicative of item-specific learning, and a strong root frequency effect shows the network has learned to generalize across a word's derivations and inflections. Item-specific learning implies a greater dependency on 'rote', and root-based learning implies greater dependency on 'rules'.

In the present paper, we challenge the traditional interpretation of surface frequency and root frequency effects as straightforward diagnostics of storage (either at the access level or at more central levels of representation) and computation respectively. We do so on the basis of an examination of the processing of low-frequency complex words in English. The choice of low-frequency words is motivated by several considerations.

First, there is a growing body of literature suggesting that age of acquisition (Brysbaert, Lange, & Wijnendaele, 2000; Carroll & White, 1973; Ellis & Lambon Ralph, 2000; Gerhand & Barry, 1999; Morrison & Ellis, 2000) may be driving a substantial part of the word frequency effect. By selecting our stimuli well outside the range of frequencies where a substantial confound with age of acquisition might be at issue, we sought to maximize the likelihood of observing the combinatorial dynamics of morphological processing and to minimize potential effects of maturation.

Second, the traditional interpretation of surface and root frequency effects makes the straightforward prediction that decompositional, root-driven processing should become increasingly important as surface frequency decreases. Complex words with a high frequency of use would have their own representation either at the access level (Caramazza et al., 1988; Schreuder & Baayen, 1995) or at a more central level (Taft, 1979, 2004). But in the lower frequency ranges, effects of surface frequency should be either weak or absent altogether. Conversely, effects of root frequency should be most clearly detectable in the lower frequency ranges. Experimental evidence reported by Alegre and Gordon (1999) and Gordon and Alegre (1999) suggests that indeed English readers do not have separate representations for lower frequency words. They propose that regularly inflected English words only have separate individual representations when their frequency exceeds a threshold value of 6 occurrences per million. Below this threshold, processing proceeds only on the basis of a word's constituent morphemes. If the traditional interpretation of root and surface frequency effects is correct, and if the findings of Alegre and Gordon replicate, a survey of derived and notably regularly inflected words with surface frequencies well below the threshold of 6 per million should reveal consistent effects of root frequency and no effects of surface frequency.

However, and this brings us to the third reason for focusing on low-frequency complex words, there is evidence from Dutch that surface frequency effects are far more pervasive than the study of Alegre and Gordon (1999) would lead one to expect. For instance, Baayen et al. (1997) observed frequency effects for regular Dutch noun plurals well below the threshold of 6 per million. Follow-up studies on inflectional morphology in Dutch (Baayen, Schreuder, de Jong, & Krott, 2002; Baayen, McQueen, Dijkstra, & Schreuder, 2003) also point to pervasive effects of full-form frequency for regular inflections in both visual and auditory lexical decision. At the same time, effects of root frequency turned out to be far more elusive (Bertram, Schreuder, &

Baayen, 2000). Although all these studies on Dutch attempt to explain the experimental data using the traditional interpretation of root and surface frequency effects as diagnostics of rule and rote, the results reported in the aforementioned study by Bowers et al. (2005) seriously challenge this interpretation.

Recall that Bowers and colleagues observed, using the semantic categorization task, that *hat* is detected in *chat.* A perusal of the CELEX lexical database (Baayen. Piepenbrock, & Gullikers, 1995) shows that the string *hat* is embedded in only 5 morphologically complex words, but in 64 other words (such as *chat*) in which it is not a morphological constituent. The type-based probability that the string *hat* represents the stem *hat* is a mere 0.07, and the corresponding token-based probability is even lower (0.0001). To this we should add that these probabilities of encountering the stem *hat* are overly optimistic as we did not count words such as *heather* and *charter* in which *hat* is a discontinuous substring. What this example illustrates is that knowledge of the presence of the string *hat* without any information about the context in which this string occurred is completely useless. Although *hat* is apparently detected in *chat* or *heather*, the meaning of *hat* does not reach awareness when either of these words is presented - it is ultimately rejected as a valid constituent, because the surrounding context (the single "c" or the embedding characters in *heather*) does not support a valid morphological parse.

When a truly morphologically complex word such as *hatless* is read, *hat* is likewise detected, but now its interpretation as a piece of clothing is licensed by its right context, the denominal derivational suffix *-less*. We therefore propose to understand the surface frequency of *hatless* as an estimate of the joint probability of *hat* and *-less,* i.e., as a contextually conditioned, syntagmatic probability that supports *hat* as a real meaningful constituent. The root frequency of *hat*, by contrast, simply estimates the unigram probability of *hat* in the context of a preceding and a following space character. In other words, we argue that surface frequency effects do not reflect the activation of dedicated representations at either the access level or at more central processing levels. Instead, we believe that surface frequency effects reflect the syntagmatic combinatorial properties of morphological formatives. Moreover, we think this combinatorial information is stored in lexical memory precisely because it provides crucial information for distinguishing between true (*hat* in *hatless*) and false (*hat* in *chat*) positives.

This new interpretation of root and surface frequency predicts that, even in the lower frequency ranges, where differences between surface frequencies are much smaller than in the higher frequency ranges, effects of surface frequency should still be observed. However, there should be no effect of root frequency, as the root frequency measure estimates a contextually inappropriate probability for complex words.

In order to achieve the goals of this study we have made use of a regression design, following Balota et al. (2004), Ford, Marslen-Wilson, and Davis (2003), and Baayen et al. (2006). This allows us to simultaneously examine a much larger array of psycholinguistic processing variables than has been examined (and thus controlled) in previous factorial studies. A number of our measures for gauging lexical competition are based on Shannon's entropy H,

$$H = \sum_{i=1}^{C} p_i \log_2\left(\frac{1}{p_i}\right),$$

where $C$ is the cardinality of a given lexical set (e.g., a set of lexical competitors, or a set of morphologically related words). The probability $p$ ranges over all the words in the set under consideration, and is estimated as a given word's relative frequency in the set. The entropy of a set is a measure of its average amount of information, and can be thought of as a token-weighted type count. A higher entropy value indicates that there are more lexical candidates in the set, or candidates that are more similar in frequency (leading to a smaller probability of identification for the target word), or both.

We used the same stimuli across four common word recognition tasks: auditory and visual lexical decision and naming. This allows us to compare the consequences of the different temporal dynamics of reading and listening, as well as the consequences of differences in the depth of semantic processing: The lexical decision task requires substantially deeper semantic processing than does word naming (Baayen et al., 2006; Balota et al., 2004).

The current study also addresses the issue of sequential longitudinal effects in word recognition experiments. These longitudinal effects are not of interest to us by themselves, rather, we view them as sources of noise that we want to bring under statistical control in our analyses. We included trial number as a predictor as a means for capturing effects of learning or fatigue during the course of the experiment. Furthermore, it has long been known that adjacent response times in choice tasks such as lexical decision are positively correlated with each other (see Sanders, 1998, for a review). Taylor and Lupker (2001) recently found that RTs on a given visual naming trial depended on whether the preceding trial was easy or difficult (defined either as a word vs. a nonword, or a high- vs. a low-frequency word). We therefore also developed a measure that quantifies the correlational structure of the reaction times with the reaction times at the four preceding trials in the experiment. In the following we provide an overview of the predictors that we took into account in our experiment.

## Predictors

**Control variables.** The **voicing of the initial phoneme** (see Bates, Devescovi, Pizzamiglio, & D'Amico, 1995; see also Bates & Liu, 1996) and the **place of articulation of the initial phoneme** (coded as either front, mid, or back) were included primarily as controls for the voice key in word naming. However, given the results of Balota et al. (2004), they might also affect lexical decision latencies. We therefore assessed the variables' importance across all four tasks, and retained them in the statistical model only when significant. We included the **trial**

**number** (i.e., the position of a stimulus in the randomized presentation lists) in an effort to minimize variance attributable to practice or fatigue effects. In order to assess and control sequential dependencies in the RTs (Taylor & Lupker, 2001), we included three variables (**preceding trial principal components,** or **PCs**) that capture aspects of how quickly the participant responded on the previous four trials. Across all four tasks there were positive significant correlations between a given RT and the previous four RTs, all of which were pairwise correlated as well. In order to bring spill-over effects from preceding trials under statistical control, we orthogonalized the four vectors of preceding RTs by means of principal components analysis (cf. de Vaan, Schreuder, & Baayen, 2006), and included the first three principal components as predictors, thereby avoiding problems with multicollinearity.

**Form variables.** One variable for a word's form is its length in letters or phonemes. Recent studies suggest that the effect of **item length** is nonlinear rather than linear (Baayen, 2005; New, Ferrand, Palliert, & Brysbaert, 2006; Baayen et al., in 2006). The acoustic duration of each (spoken) item**,** in msec, was divided into two subcomponents to be used as predictors. First was length from acoustic onset to the uniqueness point (**length to UP**), in msec. The UP is the theoretically earliest moment at which a risk-free commitment can be made to a word candidate (see Marslen-Wilson, 1984; Marslen-Wilson & Welsh, 1978). In other words, it is the first position at which all lexical candidates that are not morphologically related are inconsistent with the acoustic input. The UP was defined as the middle of the prototypical segment of the particular phone in question (following Radeau, Mousty, & Bertelson, 1989). This point was located using both visual and auditory criteria, with the help of a waveform editor. We expect a larger effect of the UP location in auditory lexical decision than in auditory naming. This is because the later the UP is in a word, the more onset-aligned competitor words it will generally have; and research discussed above suggests that these competitors are co-activated along with the target word, in terms of both form and meaning. We also considered the duration from the UP to item offset, in msec (**UP-to-offset**). While our expectation was these durational variables would have their strongest effects in the auditory tasks, there is some evidence of effects of acoustic structure in non-auditory tasks as well (Balota et al., 2004; Grainger, Diependaele, Spinelli, Ferrand, & Farioli, 2003; Harm & Seidenberg, 2004; Miellet & Sparrow, 2004; Newman & Connolly, 2004).

The traditional N-count measure of Coltheart et al. (1977) was zero for most of our complex words, and is therefore not considered. Instead, we developed two new entropy measures designed to capture the orthographic similarity effects reported by Bowers et al. (2005). **Embedded entropy** was computed across words contained contiguously within one of our target words (e.g., *ring* in our target word *pouring*). We also calculated an entropy over the set of any longer words in which our target was embedded, but in which there were intervening letters (e.g., *procuring* contains our target word *pouring,* as does *posturing).* We refer to this second variable as **matrix entropy.** Embedded and matrix entropies, both measures for the amount of non-morphological noise in the set of lexical candidates, are predicted to be positively correlated with reaction times in visual lexical decision.

**Frequency variables. Surface frequency** and **root frequency** values were taken from the CELEX lexical database of 17.9 million tokens (Baayen et al., 1995). Surface frequency is the wordform frequency from CELEX, and root frequency is the lemma frequency of the root morpheme. In both cases we summed over homographs. If the traditional interpretation of these variables is correct, we should observe no effect of surface frequency and a clear effect of root frequency. If our inference from the study by Bowers et al. (2005) is correct, root frequency should be ineffective and surface frequency should be the crucial predictor across tasks and modalities. We note that we defined root frequency as the lemma frequency of the base and not as the summed frequency of the base across all the derived words in which it occurs as previous research on the family size effect (see below) has shown that the number of different word types rather than the cumulated tokens contributed by these types is the crucial measure to consider.

**Semantic variables.** We included **semantic transparency ratings** as way of addressing the issue of whether the compositionality of a complex word has any effect on its recognition. For each item, the transparency rating used was the mean value given by fifty participants on an eight-point scale (see Wurm, 1997, 2000; Wurm & Ross, 2001 for more details about the rating procedure). Mean transparency ratings for the individual items ranged from 2.32 (*hooker)* to 7.84 *(dictionaries).* Transparency ratings averaged over the affixes ranged from 3.0 (*em-*) to 6.4 (-*s*). We also considered the **number of meanings** of a target word through its number of synsets as listed in WordNet (Beckwith, Fellbaum, Gross, & Miller, 1991; Fellbaum, 1998; Miller, 1990; cf. also Baayen et al., 2006). For the present low-frequency words, this measure was not predictive of RTs in any of our four tasks, so we will not discuss it further.

**Morphological variables.** As a first measure of morphological paradigmatics we considered the **morphological family size.** Like word length, this variable may enter into a nonlinear relation with RT (Baayen et al., 2006). The family size count considers a word's paradigmatic relations with derived words and compounds. We do not expect to find effects of family size in naming, as the family size effect is known to be semantic, whereas word naming is rather insensitive to meaning. We assessed the paradigmatic relations with a word's inflectional variants by means of a separate measure, **inflectional entropy** (Moscoso del Prado Martín, Kostić, & Baayen, 2004b; Baayen et al., 2006). It can be thought of as a token-weighted inflectional family size measure, and we expect it to have a facilitative relationship with response times (Moscoso et al., 2004b; Tabak, Schreuder, & Baayen, 2005; Baayen et al., 2006). Since for our data, inflectional entropy is calculated including frequencies below the putative threshold of 6 per million proposed by Alegre and Gordon (1999), their threshold theory predicts that we should not find an effect of inflectional entropy. This can be seen especially in the subgroup of noun stimuli, where inflectional entropy would always equal 0 if frequencies less than 6 were excluded from the calculations.

We also computed an entropy measure designed to capture aspects of the temporally unfolding speech signal for the spoken tasks. **Cohort entropy** was calculated across any words that matched a given target word from its onset all the way to its offset. It quantifies the amount of information carried by a target word's continuation forms (such as *thankfulness* for our target word *thankful*). Previous work (Wurm et al., 2006) suggests cohort entropy to be facilitatory.

A further morphological variable that we considered was **affix type.** We included equal numbers of derivationally prefixed, derivationally suffixed, and inflectionally suffixed words. Inflectional affixes are fully productive, and the semantic compositionality of inflected items is uniformly high. Derivational affixes vary widely in their productivity, and many derived words express idiomatic shades of meaning (e.g., Gonnerman, 1999). We were interested in determining whether RTs would reflect primarily a distinction between prefixing and suffixing, or a distinction between inflection and derivation. We also wanted to know whether the answer to this question was the same across all four tasks.

For the subset of derived words, we considered a final morphological measure, **affix productivity**. Several quantitative measures have been developed in the linguistic literature. We considered the P measure introduced in Baayen (1991,1994, 2001; Baayen & Renouf, 1996), which gauges the probability that the affix will occur in a neologism. We obtained for each affix the value of P listed in Hay and Baayen (2002). Their calculations were performed for bimorphemic complex words, so this measure dovetails with the degree of morphological complexity of the derived words in our experiments. The P measure gauges the processing costs of an affix: Affixes with higher values for P should reveal reduced processing latencies. We expect to observe this effect in visual lexical decision, for two related reasons. First, P can be thought of as a measure of how "possible" a target is given its affix, and hence is a measure of lexicality. Second, because P expresses the likelihood of morphological well-formedness given the affix, the conditioning information, i.e., the affix, should be available as soon as possible, a condition met only for visual presentation.

## Experiment 1: Lexical Decision and Naming

### *Method*

*Participants.* One-hundred eighty-four participants were assigned to perform auditory naming, auditory lexical decision, visual naming, or visual lexical decision. Participants were undergraduates from the psychology subject pool at Wayne State University. All were native speakers of English with normal hearing and normal or corrected-to-normal vision. Participants received extra credit in a psychology course for their participation. Forty-six people performed visual naming, 45 performed auditory naming, 52 performed visual lexical decision, and 41 performed auditory lexical decision.

*Materials.* The critical items were 70 prefixed derivations, 70 suffixed derivations, and 72 suffixed inflections. For the inflected words, we used 72 of the stimuli used by Alegre and Gordon (1999; Gordon & Alegre, 1999). For the derived words, we selected words from the CELEX lexical database of 17.9 million tokens (Baayen et al., 1995) that were marked as morphologically complex, and composed of one affix plus one root morpheme. We further restricted the set to include only those words that had surface frequencies less than six occurrences per million.

The suffixed inflections were all fully productive and regular, while the derived items spanned the range of productivity (from 0.001 for the prefixes *be-* and *dis-,* and the suffix *-ly,* to 0.015 for the prefixes *over-* and *sub-*). The critical items are listed in the Appendix, and Table 1 shows summary information on key variables.

We also included 108 monomorphemic filler items (mostly adjectives). For lexical decision we included an equal number of pseudowords. These were matched pairwise to the words on length in letters, orthographic neighborhood size, orthographic neighborhood frequency, and the number of words matching on each bigram (as well as the average frequency of those words). Affixes carried by the real words were used on the pseudowords, in proportions that matched the real words. For the auditory conditions these items were read by a native speaker of English unfamiliar with the purposes and hypotheses of the study, digitized at a sampling rate of 20 kHz, low-pass filtered at 9.8 kHz, and stored in individual computer files. The spoken pseudowords that were needed for auditory lexical decision were spoken versions of the visual pseudowords described above.

*Procedure.* For the naming conditions, participants were tested individually in a sound-attenuating booth. Participants were instructed to repeat each word as quickly and accurately as possible. A microphone was positioned approximately 15 cm in front of each participant. For lexical decision, participants were tested in groups of one to three. They were instructed to press one button for real words and one button for pseudowords, as quickly and accurately as possible.

For the visual conditions, stimuli were displayed in the center of a computer monitor in a random order. For the auditory conditions, stimuli were played binaurally over headphones at a comfortable listening level. A different random order was used for each participant or group of participants.

Response latencies were measured from the onset of each stimulus. A practice set of 20 words was used prior to the main experiment to familiarize participants with the procedure.

*Data Analysis.* We analyzed the data set with a linear mixed-effect analysis of covariance with log RT as the dependent variable and participant and items as crossed random effects (Bates, 2005; Bates & Sarkar, 2005; Faraway, 2006; Pinheiro & Bates, 2000). In our analyses, the frequency, family size, and productivity variables were transformed logarithmically in order to remove the skewness in their distributions and to minimize the effect of atypical outliers. Finally, we explored potential nonlinearities by allowing quadratic terms into the statistical models. Only significant nonlinearities were retained and reported.[1]

*Results and Discussion*

Error rates were 6.0%, 3.8%, 13.9%, and 14.0% for the visual naming, auditory naming, visual lexical decision, and auditory lexical decision tasks. Response times for these trials were not included in the statistical analyses. We begin with an overall analysis that included task as a four-level factor, and in which we allowed our predictor variables of interest to interact with task. The interaction results from this overall analysis are shown in Table 2.

As expected, Task had a large and significant effect ($F[3, 34415] = 54.8847$, $SS = 2{,}616{,}651$, $p < .0001$). More importantly, task interacted with all other variables, as shown in Table 2. For ease of interpretation, we fitted a separate multilevel model of covariance to each of the four datasets, using a stepwise backward variable selection procedure. The resulting models are summarized in Tables 3-6. For each resulting model, we carried out bootstrap validation on the by-item means with 200 bootstrap runs and obtained estimates of the explained variance that are more conservative and provide a more realistic measure of the predictivity of the model for unseen data (Harrell, 2001).[2] These bootstrap-adjusted $R^2$ values are listed with Tables 3-6. For all models the optimism (the difference between the raw $R^2$ and the bootstrap-adjusted $R^2$) was small. This ensures that we are not overfitting the data.

Figures 1 and 2 show the partial effect of each significant continuous predictor when the other predictors are held constant at their medians, for the naming and lexical decision tasks, respectively. In each figure, black lines show significant effects when stimuli were presented visually and gray lines show significant effects when stimuli were spoken.[3] Because the y-axes in a given figure use the same range of values for each predictor, the figures provide immediate insight into the relative effect sizes of the predictors. The lexical decision tasks showed more significant effects than the naming tasks; and auditory naming showed fewer even than visual naming. In what follows we will discuss the predictivity of our variables across the four tasks.

**Control variables.** The **voicing of the initial phoneme** was significant in two of the tasks. In visual naming, voiced onsets were associated with slower RTs, while in auditory lexical decision, voiced onsets were associated with faster RTs. These results contrast with those of Balota et al. (2004) for monosyllabic words. They observed that a voiced onset was associated with slower *visual* lexical decision times but *faster* visual naming times. The voicing variable is probably distributed quite differently over our complex words and the monosyllabic (and monomorphemic) words of Balota et al. (2004); and in their study, voicing was embedded in a much larger set of phonetic variables. As voicing is only a control variable in the present study, we leave this issue for clarification by future research.

A second control variable for the voice-key was **place of articulation** (coded back, mid, or front). Surprisingly, place of articulation was not predictive for naming, but was significant in auditory lexical decision. Items with a back place of articulation were processed faster than those with either a front or mid place of articulation. This unexpected effect requires replication before meriting interpretation.

**Trial number** was significant in three of the four experiments. It was associated with longer response times in both naming tasks, indicating that subjects slowed down as they proceeded through the experiment (an effect of fatigue). There was no trial effect in visual lexical decision, and a small learning or practice effect in auditory lexical decision.

The **previous-trial PC**s revealed evidence of sequential RT dependencies across all four tasks, as expected given the literature. What was unexpected was the magnitude of the effects, which was sometimes found to be hundreds of milliseconds (see Figures 1 and 2). Based on the work of Taylor and Lupker (2001), we had expected only very small effects. We conclude that this is a source of noise that is worth taking out of the error term.

**Form variables. Word length** in letters was found to have significant nonlinear effects in both visual tasks, consistent with other findings in the literature (e.g., New et al., 2006). The auditory duration variables had significant effects in both auditory tasks. Words with later **UP**s had longer response latencies, measured from word onset, in both the auditory lexical decision and auditory naming tasks. This is as predicted by the cohort model from which the UP construct came (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1984, 1989). Note that a later UP implies prolonged lexical competition that does not involve morphologically related words, and that we therefore observe inhibition.

A follow-up analysis of the auditory data showed that the UP effect was stronger in auditory lexical decision than in auditory naming, as predicted ($F(1,17390) = 3.95$, $SS = 70,804$, $p < .05$). This underscores the importance of semantics in lexical decision and supports the idea that the semantics of the onset-aligned cohort competitors are activated (e.g., Zwitserlood, 1989).

Surprisingly, the auditory UP location also had a significant effect in *visual* lexical decision, where later UPs were associated with faster RTs. In the light of Johnson and Pugh's (1994) study on visually-driven cohort effects, we decided to calculate the visual uniqueness points, expressed as letter positions, for our stimuli. We re-ran the statistical model shown in Table 5, substituting the visual UP for the auditory UP. These visual UPs were not predictive of RTs ($B$ = -5.5584, $SE\ B$ = 3.4779, $p$ = .11). We also ran a version of the analysis that included both the auditory and the visual UPs. Again, the visual UPs were not predictive ($B$ = -1.970, $SE\ B$ = 3.6891, $p$ = .59), whereas the auditory UPs remained significant ($B$ = -0.0873, $SE\ B$ = 0.0341, $p$ = .01). Thus we can rule out the possibility that the UP effect is confounded with a purely orthographic cohort effect.

Why is a late UP helpful with visual input, but inhibitory with auditory input? Possibly, a late UP is an index of lexicality, in the sense that words that become unique very early are phonotactically more idiosyncratic and thus less word-like than words that share a larger initial sequence of phonemes with other words. Greater lexicality, particularly when it is evident early in a spoken word, seems to provide support for faster "yes" responding in the lexical decision task; but in auditory processing, this greater lexicality goes hand in hand with prolonged lexical competition, as the perceptual system must winnow down the candidates to one (see Wurm et al. (2006) for a discussion of the differing weightings of such facilitative and inhibitory effects during the time-course of spoken word recognition). In visual processing, which has fast access to the full word, there is evidently a benefit from this additional source of lexicality information that is available in auditory memory.

The other portion of a word's duration, **UP to offset**, led to longer response latencies in both auditory tasks. The more acoustic information there is beyond the UP, the longer participants wait to respond. Participants wait past the UP to determine whether any disconfirming evidence is subsequently encountered, or to determine which particular inflectional variant or other morphological continuation form is being presented. Such waiting would seem to be a necessity in naming, in which the precise form must be uttered (including any suffixes), but the data show that there is a similar effect in auditory lexical decision.

The two entropy measures that we designed to capture the similarity effects reported by Bowers et al. (2005) for the visual semantic decision task were predictive for our visual lexical decision latencies. Greater entropies correlated with longer response latencies for both **embedded entropy** (e.g., *ring* in the target word *pouring*) and **matrix entropy** (e.g., *posturing* for the target word *pouring*), as expected for measures that gauge the probability of inappropriate lexical candidates. These two measures can be thought of as extensions of Coltheart's N to longer, morphologically complex, words. Coltheart's N is often found to be facilitative in visual lexical decision, even though it is a measure of the presence of words related to the target only in form (see, e.g., Balota et al., 2004). However, Baayen et al. (2006) have recently shown that the effect seems to be inhibitory when several other variables are controlled for. If correct, this inhibition for the N-count measure ties in nicely with the inhibition for our new entropy measures.

**Frequency variables.** The traditional interpretation of the **surface frequency** effect, combined with the results reported by Alegre and Gordon (1999; Gordon & Alegre, 1999), predicts that surface frequency should not be a significant predictor for our data, especially not for our regular inflected words. However, all four tasks revealed a facilitative effect of surface frequency. In three of the four tasks, moreover, the effect of surface frequency was statistically equivalent across the three affix types. Only in auditory naming did the affix type by word frequency interaction reach significance, and follow-up tests showed that the frequency effect was restricted there to the derived prefixed words.

The affix type by surface frequency interaction could be due to the unique characteristics of auditory naming, the task that least engages the mental lexicon because the input already provides all information required for naming the word. Given that this task has the least sensitivity, the question remains why it is that only the prefixed words show a surface frequency effect. A possible explanation focuses on the ordering of the more informative vs. the less informative morpheme (i.e., the root vs. the affix). In the context of a low-information prefix followed by a high-information root, the combinatorial likelihood of the two elements (estimated by the surface frequency measure) matters. For the other ordering, in which the high-information root is followed by a low-information suffix, the combinatorial likelihood of the pairing is much less relevant. We characterize the prefix as less informative because the likelihood of the root given the prefix is smaller than the likelihood of the suffix given the root: Prefix families tend to be much larger than root families.

Equally surprising in the light of the traditional interpretation of surface and root frequency effects was the finding that **root frequency** was a significant predictor only in the visual naming task. Our difficulty in finding a root frequency effect is reminiscent of Bertram et al.'s (2000) work on low-frequency words in Dutch: Using factorial designs, they, too, had difficulty demonstrating root frequency effects. As will become apparent below, it is unwise to try to attribute this effect to the properties of the task. Instead, what we see here is probably nothing more than an effect of chance. Given the power provided by the number of items in our experiment, one out of four experiments may be expected to reveal a significant effect of root frequency.

We note here that the overall pattern of results obtained for surface and root frequency runs counter to the predictions that we derived on the basis of the traditional interpretation of these effects in the literature. The present results therefore provide support for the alternative interpretation of surface and base frequency that we derived on the basis of the experiments reported by Bowers et al. (2005): Surface frequency reflects the combinatorial knowledge based on previous experience that is crucial for accepting the constituents as genuine meaning-bearing units rather than as strings providing accidentally partial matches. Root frequency, by contrast, reflects a contextually less appropriate probability, the likelihood of observing the root as an independent word.

**Semantic variables.** As expected, the lexical decision tasks showed significant effects of semantic transparency, whereas the naming tasks did not. In lexical decision, the more transparent a combination of root and affix was judged to be (by a separate sample of participants), the faster RTs were.

**Morphological variables. Morphological family size** was predictive only for the visual lexical decision latencies. Family size was facilitatory, as expected, but as Table 5 and Figure 2 show, the effect was not linear, and the benefit did not extend to the smallest family sizes. **Inflectional entropy** was predictive in visual lexical decision (because inflectional entropy had a marked bimodal distribution, we dichotomized this variable in order to be conservative, but the non-dichotomized variable yielded essentially the same results). Words with nonzero inflectional entropies had shorter visual lexical decision latencies than those with zero entropies ($B = -23.1172$, $SE\ B = 8.4838$, $p < .01$). This is consistent with previous research (Moscoso et al., 2004b; Baayen et al., 2006; Tabak et al., 2005), and suggests that the entropy associated with an inflectional paradigm provides overall support for the lexicality of an item.

As expected, the **cohort entropy** measure showed a significant facilitatory effect in auditory lexical decision, replicating recent work in Dutch (Wurm et al., 2006). The words that enter into the cohort entropy calculation are morphological continuation forms of the target word, and they provide further support for the target's lexicality. What we see, then, is that the early part of the auditory disambiguation process is made more difficult by the presence of similar but morphologically unrelated forms, as witnessed by the inhibitory effect of UP discussed above. However, the later part of the process is made easier thanks to morphological connectivity, as witnessed by the facilitative cohort entropy effect. This, too, is reminiscent of recent findings in spoken Dutch (Wurm et al., 2006). In that study the authors found that *late entropy* (the term they used for cohort entropy) has a facilitative effect on auditory lexical decision times; but the same entropy calculated at positions prior to the words' UPs has an inhibitory effect.

**Affix type** was a significant predictor of RTs in all four tasks. The pattern that emerged from our analyses was consistent: RTs for the two suffixed types were not statistically distinguishable, and were shorter than those to the prefixed words. The sole exception to this generalization was that derived suffixed words had intermediate RTs in auditory lexical decision. The clearest differentiation across all tasks then seems to be between prefixing and suffixing. We believe this is due to the fact that in the case of prefixed words, the initial set of word candidates that is activated is based on input that is highly ambiguous: Large numbers of words can begin with most English prefixes. It is remarkable that the split between inflection and derivation, which one would expect to be all-important in the light of the claims of dual route theory of Pinker (1999) and the results reported by Alegre and Gordon (1999), was not supported at all.

Finally, we added log P to the statistical models shown in Tables 3-6 for the subset of derived words, to determine whether **affix productivity** had any incremental predictive value. Log P emerged as a significant predictor of RT only for the visual lexical decision data, with a negative coefficient ($B = -16.0422$, $SE\ B = 7.2668$, $t(4329) = -2.2076$, $p < .05$), as expected. In reading, the processing of low-frequency derived words benefits from greater affix productivity. To our knowledge, this is the first time an effect of affix productivity has been documented with a chronometric experimental paradigm.

The facilitation observed for morphological measures (family size, inflectional entropy, cohort entropy and affix productivity) contrasts markedly with the inhibition observed for our measures of non-morphological similarity (embedded entropy, matrix entropy, and UP). This pattern of results shows that morphological and non-morphological similarity have qualitatively different consequences for the dynamics of lexical processing.

## Experiment 2: A Comparison with the English Lexicon Project

Inspection of the item lists in the Appendix reveals a number of very rare words, which of course is not surprising given that we wanted to test the idea of a frequency threshold. The possibility remains, though, that the results we have obtained so far are due to idiosyncratic aspects of our items (many of which were taken directly from Alegre & Gordon, 1999). In addition, a number of our items carried roots not used in their more typical senses (e.g., "dismember"). This stimulus characteristic was captured by our semantic transparency ratings, and we did find that transparency had large significant facilitative effects in visual and auditory lexical decision in Experiment 1. Nevertheless, inclusion of these items may have changed the nature of the tasks for subjects. In what follows, we show that the frequency effects found in Experiment 1 are in fact typical of what can be observed in an analysis of RTs from nearly 8500 words across the frequency spectrum, using the English Lexicon Project database.

### *Method*

*Materials.* We extracted all monomorphemic words from the CELEX lexical database. On the basis of the resulting list, we next extracted all inflected and derived words obtained from these base words by addition of a single affix. For the subset of words in this list for which chronometric measures were available in the English Lexicon Project (Balota et al., 2002), we obtained visual lexical decision and word naming latencies. From the resulting list of 9490 words we removed homographic duplicates. For homographs, the cumulated lemma frequencies of their base words were used as root frequency measure. In this way we obtained a list of 8486 morphologically complex words, comprising 1358 adjectives, 2467 nouns, 4602 verbs, and a small number of words belonging to minor word classes such as adverbs and quantifiers. Because of the exhaustive nature of this item selection, these words represent the full range of both surface frequency and root frequency.

*Data Analysis.* The data base was enriched with a wide range of predictors, of which the following reached significance in our analyses: word length (in letters, Length), number of syllables (NSyll), the frequency of the complex word (Surface Frequency), the frequency of the complex words in the spoken (demographic) subcorpus of the British National Corpus (BNCd), the frequency of the base word (Root Frequency), the family size of the base (VfB), the number of synsets for the base word (NSyn), whether the word was inflected or derived (MorphType), whether the affix is a prefix or a suffix (AffixType), inflectional entropy (Hi), affix family size, i.e., the number of types with the affix in CELEX (V), and the log frequency of the bigram spanning the boundary of stem and affix (LogBigFreq).

We analyzed the visual lexical decision times with a linear mixed effects model with affix as random effect. The results of this analysis are shown in Table 7. Word length revealed a U-shaped effect, as in Experiment 1. Length evaluated by means of the number of syllables (NSyll) revealed an additional linear inhibitory effect.

A greater frequency in spoken English (BNCd) allowed faster response latencies, especially for derived words, but also for inflected words. Since this variable is highly correlated with age of acquisition (see, e.g., Baayen et al., 2006), it is unlikely that our results are distorted by a lack of control for this variable, which is not available for a large majority of the words studied here.

Derived words with more synonyms (as gauged by their number of synsets in WordNet, SynWord) had shorter response latencies, as did words with larger morphological families (VfB). In addition, words with higher inflectional entropies (Hi) likewise allowed for shorter response times.

Of central importance in the current study, both root frequency and surface frequency had facilitative relationships with RT. These main effects are shown in Figure 3. Black lines represent derived words, gray lines inflected words. Note that the effect of surface frequency (left panel) is slightly non-linear, and much larger than the effect of root frequency (right panel) for both derived words and inflected words, even though the effect of surface frequency for inflected words is slightly attenuated. In our own visual lexical decision experiment, we apparently did not have sufficient power to detect this nonlinearity for surface frequency or the small effect of root frequency.

The significant interaction of these two measures is visualized in Figure 4. Note that a weak facilitatory effect of root frequency for the lowest surface frequencies reverses into a slightly inhibitory effect for the highest surface frequencies. This interaction suggests that the processes that these two frequency measures are tapping into may not be independent.

The analysis also revealed affix-specific effects. Affixes with a greater affixal family size (V) elicited shorter response latencies. Furthermore, derivational affixes (but not inflectional affixes) with a higher bigram frequency spanning the boundary between stem and affix (LogBigFreq) emerged with longer reaction times.

Figure 5 visualizes the random intercepts (horizontal axis) and random slopes for surface frequency (vertical axis) for the affixes in our study.[4] Less productive or unproductive affixes such as -th or -ment tend to be more to the right in this scatterplot, and more productive affixes such as -ness or -less somewhat more to the left. The vertical dimension separates the inflectional affixes from the derivational affixes. It is for the derivational affixes that we find the largest facilitatory effects. It is noteworthy, however, that for all inflectional affixes we find unambiguous facilitation: None of the surface frequency effects (i.e. random slopes) for the inflectional affixes are even close to zero and all have a negative sign.

Of course, it is the large number of data points that allows us to obtain detailed insight in the individual processing properties of the different affixes and of the small role of root frequency and the prominent role of surface frequency. Interestingly, our large number of data points also allows us to gauge the likelihood of observing significant effects for smaller sample sizes. Figure 6 summarizes the results obtained for a sequence of sample sizes (graphed on the horizontal axis). The vertical axis graphs the proportion (out of 100 random samples) in which an effect was significant at the 5% level; in other words, it shows the statistical power of the experiment for each effect.

The upper left panel shows that the effect of surface frequency asymptotes to 1 already in samples containing 200 items. The effect of root frequency is hardly detectable at all at that sample size: Only 1 out of 5 experiments can be expected to show it. (It is interesting to note that we observed this effect for this sample size in 1 out of 4 opportunities.) To observe a root frequency effect across more than 50% of one's experiments, at least 1000 items are required. The interaction of the two frequency measures is observed with a somewhat higher likelihood.

The upper right panel restricts the resampling to inflected words. Again it is easy to see that the surface frequency effect asymptotes more quickly than the root frequency effect. Some 1000 items are required to have a 50% chance of detecting a root frequency effect here, as in the previous panel.

The lower left panel restricts the data to inflected words with a frequency of less than 6 per million, the Alegre and Gordon threshold. For these words, there is no significant interaction between surface and root frequency. Therefore, this interaction was not considered by the models fitted to the resampled data sets. It is clear that even for these low-frequency inflected forms, the surface frequency effect emerges much more often than the root frequency effect. Even with the power of 1600 items, the likelihood of observing a significant root frequency effect does not reach 0.8. In the light of this power analysis, the fact that Alegre and Gordon (1999) did not observe any frequency effects at all for the 50 items of their Experiment 3 does not come as a surprise: A sample size of 50 has totally insufficient power. The fact that they did observe a frequency effect for matched monomorphemic adjectives is irrelevant, and only shows that it is easier to observe frequency effects for simple words.

In summary, our analysis of bimorphemic complex words in the English Lexicon Project suggest that the results of our Experiment 1 are not atypical. The fact that a root frequency effect is detectable in only one of our four opportunities in Experiment 1 is to be expected given the small sample sizes that we used (just over 200 items). Our power analysis indicates that studies using much larger samples may observe significant effects of root frequency more often. However, when it is observed, we expect the effect of root frequency to be very much smaller than the effect of surface frequency, and to be more evident for words with low surface frequency values.

We do not take the space to describe the analogous results for visual naming times from the English Lexicon Project, but the results were very similar to those for lexical decision.

The analysis of the data provided by the English Lexicon Project caution against concluding from Experiment 1 that base frequency is irrelevant as predictor. At the same time it is clear that the effect of base frequency is much weaker, with a very much reduced effect size, that will often not reach statistical significance in small samples with only a few hundreds of items.

## General Discussion

In language comprehension, target words have to be selected from sets of lexical candidates that in addition to the target itself contain two kinds of words: Words that bear no semantic relation whatsoever to the target (e.g., *hat* in *chat*) and morphologically related words (e.g., *hat* in *hatless*). We have argued that two kinds of information are important for trimming down the set of lexical candidates (in addition to bottom-up information), namely syntagmatic information (the joint probability of morphological formatives) and paradigmatic information (the extent to which the formatives are embedded in morphological networks).

A comparison of the word naming tasks with the lexical decision tasks in Experiment 1 shows that word naming is insensitive to any of the measures with which we probed the influence of lexical candidates other than those that received strict bottom-up support: Only the two measures involving the UP were predictive. Measures calculated over competitors that diverged from the form that had to be articulated consistently failed to reach significance, irrespective of whether these competitors were morphologically related. This finding ties in with the general insensitivity of word naming to semantics (cf. Balota et al., 2004; Baayen et al., 2006). It is in the lexical decision task, which is highly sensitive to word meaning, that facilitatory morphological effects (family size, inflectional entropy, productivity, and transparency in visual lexical decision; transparency and cohort entropy in auditory lexical decision), and inhibitory non-morphological effects (matrix and embedded entropy in visual lexical decision, UP in auditory lexical decision) emerge. We conclude that lexical decision provides ample evidence for the relevance of morphological structure through our paradigmatic measures, and ample evidence for the detection of the root in morphologically unrelated words in visual lexical decision through the matrix and embedded entropy measures.

The evidence in lexical decision for the relevance of morphological structure is at odds with the presence of surface frequency effects and the absence of root frequency effects in the very same lexical decision experiments, given the traditional interpretation of these frequency measures as diagnostics for storage and computation respectively. We did not replicate the results of Alegre and Gorden (1999; Gordon & Alegre, 1999), but observed effects of surface frequency below the putative threshold of 6 per million across all tasks. For regularly inflected words, quintessentially compositional, only the most insensitive task (auditory naming) failed to show significant facilitation. Not only did we observe a robust effect of surface frequency, we also did not find an effect of root frequency in visual and auditory lexical decision and in auditory naming. Experiment 2 replicated this general pattern of results for the lexical decision latencies for some 8000 English bimorphemic affixed words, which shows that the results of Experiment 1 are unlikely to be contingent on the selection of our stimuli, by-item error rates, strategic task effects, or other potentially invalidating experimental flaws that that one might appeal to in order to dismiss our findings. Given sufficient power, a small effect of root frequency can be observed, in interaction with the surface frequency effect.

These results allows us to conclude that English is not that different from Dutch, a language in which frequency effects for inflections below the supposed threshold of six occurrence per million have been observed repeatedly (Baayen et al., 1997, 2002). More generally, these results dovetail well with the frequency effects reported for regular inflections by Sereno and Jongman (1997) for English, and by New, Brysbaert, Segui, Ferrand, and Rastle (2004) for French and English. However, the traditional interpretations of surface and root frequency leave us with several paradoxes.

If surface frequency is a measure of the availability of a full-form access representation and a diagnostic of non-decompositional processing, the measures of morphological paradigmatics should not have been predictive, contrary to fact. If root frequency is a measure of decompositional processing, it should have been predictive for the lexical decision tasks where we see stem-based measures of lexical competition, both morphological and non-morphological, at work, but it is not. Moreover, in the low frequency range, where there is less entrenchment in memory, stem-driven processing should be prevalent, instead of full-form driven processing.

How can we understand the combined presence of (i) a strong and robust surface frequency effect, of (ii) a weak and much more fragile effect of root frequency, and of (iii) a subtle interaction between these two effects? The explanation we explore here takes as its point of departure a time-dynamic lexicon defined as a tuple ($L$, $Pr_t$). In this tuple, $L$ is a set with three subsets, the set of free morphemes $F$, the set of bound morphemes $B$ and the set of complex words $C$: $L = F \cup B \cup C$. Complex words are defined as ordered n-tuples of formatives from $F$ and $B$, with $n \geq 2$:

$$C = \{ (m_1, \ldots, m_n) \}, m_i \_ \{F \cup B\}. \tag{1}$$

The function $Pr_t$ associates each element $m \in L$ with a probability at time $t$. At $t_0$, the moment in time immediately preceding stimulus presentation, $Pr_0$ associates each element $m \in L$ with its long-term probability, estimated by its relative frequency in a large corpus with $N$ tokens: $Pr(m) \approx f(m)/N$, where $f(m)$ is the frequency of occurrence of $m$ in this corpus. At successive timesteps $t_1$, $t_2$, … following stimulus presentation, $Pr_t$ associates each element $m$ with its probability given the reduced set of lexical competitors active given the depth of processing of the input at that timestep.

Now consider the temporal dynamics of retrieving a bimorphemic derived word such as (good, ness) from this lexicon. General pattern matching processes yield a host of lexical candidates, and it is only by inspecting the contexts in which the formatives appear that the true can be separated from the false. The opposite effects of our measures of morphological and non-morphological lexical competition bears witness to the importance of this distinction. We therefore assume that the surface frequency effect does not tap into the availability of some holistic representation, but rather reflects long-term knowledge of the contextual appropriateness of adjacent detected formatives. For ease of exposition, we abstract away from all non-morphological lexical competition in what follows.

Suppose that a word such as *goodness*, in our model represented by the ordered pair (good,ness), is read with two fixations, one for the base and one for the suffix, and that parafoveal preview indicates that the first constituent is followed by a second, as yet unknown, constituent.

We consider four points in time, the moment before stimulus onset $t_0$, the point $t_1$ at which the eye has fixated on the first constituent, $t_2$, the moment at which the identity of the first constituent is established and the eye moves on to the second constituent, and $t_3$, the time at which the information provided by the second constituent has been processed. At $t_0$, no input has been processed, so the probability of (good,ness) is its long-term probability. For most complex words, this probability is close to zero. At $t_3$, the only word in the lexicon matching the input is (good, ness); under the simplifying assumption that there is no residual probability for mismatching lexical canditates, $Pr_3$(good, ness) = 1.

The crucial timesteps for the present purposes are $t_1$ and $t_2$. During the processing of the first constituent, the likelihood of correctly anticipating (or predicting or guessing) the target word is given by the probability of the stem (good) given that the stem is followed by at least one other morpheme m, $Pr_1(good|m)$,

$$Pr_1 \quad (good|m) = \Sigma_i \quad Pr_1 \quad (good, m_i) = Pr_1 \quad (good+). \tag{2}$$

where the condition implements our assumption that parafoveal information rules out that *good* is seen in isolation. (In (2), $Pr_1(good,m_i)$ denotes the (joint) probability of (good) followed by $m_i$.) In other words, during the uptake of information provide by (good) in the visual input, the search space is narrowed down to those morphological family members of (good) that have (good) as initial constituent. The greater the joint probability of these family members, the greater the evidence for lexicality. In what follows, the term 'morphological family' is to be understood as referring to this subset of morphological family members that have the base in word-initial position.

A second probability that comes into play at $t_2$ is the probability of (ness) given that (good) is the first constituent. We estimate this probability is estimated by the relative frequency of (good,ness) in the morphological family of (good):

$$\begin{aligned} Pr_2(ness|good) &= \\ &= Pr(good,ness)/Pr(good+) \\ &= (f(good,ness)/N)/(f(good+)/N) \\ &= f(good,ness)/f(good+) \end{aligned} \tag{3}$$

In other words, the lexical search space is now reduced to the morphological family members of *good*, and the probability of the target word is defined on this reduced search space.

Our hypothesis is that the amount of (lexical) information on which a lexical decision is made is given by the sum of all information available at the different points in time. We define amount of information as negative log probability: $I_m = -\log_2(p_m)$. Table 8 lists the probabilities and corresponding amounts of information available at the different timesteps. Initially (at $t_0$), the search space is the whole lexicon, the probability of the lexicon equals one, so the amount of information is zero. At $t_1$, the optimal search space is the morphological family, and the corresponding amount of information is $-\log_2(f(good+))+\log_2(N)$. At $t_2$, the morphological family is known with certainty, so it no longer contributes information. The amount of information in long-term memory allowing correct anticipation (guessing, predicting) of the next morpheme is now at issue. Finally, at $t_3$, the target word is know with certainty, and therefore the information contributed at that timestep is zero. Assuming that a lexicality decision is based on the amount of information accumulated over time, $RT \propto I_{tot}$, we have that

$$\begin{aligned} I_{tot} &= -\log_2(f(good+))+\log_2(N) -\log_2(f(good,ness))+\log_2(f(good+)) \\ &= -\log_2(f(good,ness)) + \log_2(N). \end{aligned} \tag{4}$$

and hence

$$RT \propto \log_2(N) -\log_2(f(good,ness)). \tag{5}$$

In other words, response latencies are determined by the joint weight of all possible lexical tokens N and by the frequency of the complex word. According to the linear model specified in (5), the frequency of the base does not contribute to the information for lexicality accumulated over time, and hence does not co-determine response latencies. It should be noted that although we have estimated Pr(good|m) by excluding the unigram probability of (good), the derivation of equation (5) would not be affected if it were included to model a cumulative root frequency effect.

Although (5) captures the main trend in our data, the evidence for a weak base frequency effect and the presence of an interaction of surface frequency and base frequency suggests that it is too simplistic to assign equal weight to the two non-trivial probabilities and their corresponding amounts of information as listed in Table 8. We therefore introduce separate weights to the two sources of information ($w_1$, $w_2 > 0$). Hence,

$$\begin{aligned} I_{tot} &= w_1[-\log_2(f(good+))+\log_2(N)] + w_2[-\log_2(good,ness)+\log_2(f(good+))] \\ &= -(w_1-w_2)\log_2(f(good+)) - w_2 \log_2(good,ness) + w_1 \log_2(N) \qquad (6) \end{aligned}$$

and

$$RT \propto w_1\log_2(N) - w_2 \log_2(good,ness) + (w_2-w_1)[\log_2(f(good+))]. \qquad (7)$$

When $w_1 = w_2$, base frequency plays no role, as before. When $w_2 > w_1$, the information provided by the surface form receives greater weight. In this condition, the coefficient for base frequency in the linear model is positive ($w_2 - w_1 > 0$), predicting an inverse base frequency effect. When $w_1 > w_2$, the information provided by the base receives greater weight. The coefficient of base frequency in the linear model is now negative ($w_2 - w_1 < 0$), and we now have facilitation from the base. By shifting these two weights, we obtain precisely the interaction of root frequency and surface frequency observed for the data from the English Lexicon Project in Experiment 2 (see Figure 4).

Different weights may be required for a wide variety of reasons. One potential mechanism might be part and parcel of the process of lexical competition itself. If the information from the full form is slow in coming available due to its low frequency, then this might be compensated for by assigning greater weight to the information originating in the morphological family ($w_1 > w_2$). Conversely, if the probability of the target word is very high, the evidence for lexicality provided by the morphological family might become less important, and therefore receive less weight ($w_2 > w_1$).

Instead of differential weighting of distinct sources of information, recency in time might be at issue. If information is subject to decay over time, information that became available more recently will have a greater weight than information that became available at earlier timesteps. Time would then favor the information based on Pr(ness|good), and would therefore lead to a reverse base frequency effect.

Yet another factor might pertain to the differences in information gain, the reduction in uncertainty when going from one timestep to the next. The reduction in uncertainty going from $t_0$ to $t_1$ is much greater than when going from $t_1$ to $t_2$. To see this, we define uncertainty with the help of Shannon's entropy. The initial uncertainty is the entropy of the lexicon as a whole,

$$H_L = - \Sigma_i \; p_i \log_2(p_i), \qquad (8)$$

of which a rough approximation is obtained by assuming all words are equiprobable:

$$H_L \approx - \Sigma_i^V \; 1/V \log_2(1/V) = \log_2(V). \qquad (9)$$

Once the set of lexical competitors has been narrowed down to the morphological family, the entropy reduces to

$$H_L \approx - \Sigma_i^V \; 1/F \log_2(1/F) = \log_2(F), \qquad (10)$$

where F is the morphological family size. Once the suffix has been processed, the whole word is recognized and the entropy becomes zero. Hence the reduction in uncertainty achieved by processing the base equals

$$\Delta_{H0,H1} = \log_2(V)\text{-}\log_2(F) = \log_2(V/F), \qquad (11)$$

the subsequent reduction achieved by subsequently processing the suffix is

$$\Delta_{H1,H2} = \log_2(F) - 0 = \log_2(F). \tag{12}$$

The list of English lemmas in the CELEX lexical database comprises some 50,000 entries and the mean base family size equals 3, hence $\Delta_{H0,H1}$ is on the order of magnitude of 50000/3 = 16666, whereas $\Delta_{H1,H2}$ is only 3. The greater reduction in uncertainty achieved by narrowing the search space down to the family of the base may favor a greater weight for Pr(good+), and therefore might help explain a facilitatory base frequency effect.

It is also conceivable that list manipulation might shift the weights. For instance, if the pseudowords in a lexical decision task consist of existing morphemes in illegal combinations (e.g, *good+ee*), the weight for correct decisions is shifted towards the full forms ($w_2 > w_1$), with as immediate consequence the reverse base frequency effect observed by Taft (2004) under these extreme conditions.

Several aspects of the present probabilistic approach are noteworthy. First, equations (4) and (5) embody the hypothesis that *in the lexical decision task different sources of information are considered jointly* - hence the summation of probabilities and the accumulation of amounts of information. We do not claim that the probabilities of the constituents (if necessary properly conditioned on their context) are irrelevant for lexical processing. To the contrary - in our model the probability of the base does play a role, but lexical decision is not the appropriate task for detecting its role. As shown by Kuperman et al. (2007ab), base frequencies and family sizes do play a role during the fixations on first and second constituents.  Our model is also consistent with results coming from eye-movement studies (Bertram & Hyona, 2003; Kuperman et al. 2007ab) that report effects of surface frequency in early measures of visual processing.  Their findings are consistent with a construal of surface frequency as part of an anticipatory conditional probability Pr(ness|good), but inconsistent with accounts holding that full-form frequency effects would only arise after the identification of all a word's constituents.

Second, the present explanation is framed totally within a lexicon that if so desired can be construed as a declarative memory in the sense of Ullman (2004). Our explanation does not depend on a separate parsing route in a distinct procedural memory system. Hence, the present theory provides a high-level probabilistic characterization of a memory system that could be both symbolic or subsymbolic in nature.

Third, the present approach generalizes to the auditory modality. We know that there is sufficient fine phonetic detail in the speech signal to establish that the base is the first constituent of a longer word well before base offset (Davis et al. 2002, Kemps et al. 2005ab, Pluymaekers, Ernestus, & Baayen, 2005; Ernestus, Lahey, Verhees, & Baayen, 2006). Hence, the conditional probability $\Pr_1$(good|m) applies to the auditory modality just as it does when the eye moves through the word.

Fourth, under the simplest possible assumptions, our results for reading generalize to words read with a single fixation only. We begin by noting that when a word such as (good,ness) is read with a single fixation, the only lexical candidate matching the visual input is the full form, so here we expect the full form frequency to be the dominant factor a priori. However, if the constituents are detected as well, several sources of information might come into play simultaneously:

$$\begin{aligned}
&\Pr(\text{ness}|\text{m}) = \Pr(\text{+ness})\\
&\Pr(\text{good}|\text{m}) = \Pr(\text{good+})\\
&\Pr(\text{ness}|\text{good}) = \Pr(\text{good,ness})/\Pr(\text{good+})\\
&\Pr(\text{good}|\text{ness}) = \Pr(\text{good,ness})/\Pr(\text{+ness}) \qquad (13)
\end{aligned}$$

The corresponding amounts of information, given the simplest possible model with equal weights, add up to

$$I = 2\log_2(N) - 2\log_2(f(\text{good,ness})), \tag{14}$$

and again the base frequency drops out of the equation.

We note that the probability Pr(ness|m) might also play a role when the word is read with two fixations, if we assume that the uptake of information from the affix in the visual input during the second fixation initially takes place relatively independently of the information gathered during the first fixation. The corresponding amount of information with its own weight would then have to be added to the model. This would amount to including the affix family as a predictor.

Finally, we need to clarify two important ways in which our approach differs from traditional models. Recall that one popular general architecture postulates morphemes (and in some models also words) at one (access) layer, and words (and in some models also morphemes) at a second (central) layer (e.g., Taft 1994, Schreuder & Baayen, 1995, Giraudo and Grainger, 2001). Morphemes at the first layer have excitatory links to the words at the second layer. Morphemes with higher activation levels would then allow faster access to the whole words with which they are linked. In models that embody this kind of architecture, a higher base frequency should always co-determine response latencies. As we have seen, this prediction receives surprisingly weak support from our data. In the model that we have proposed, all competition takes place within the set of complex words: The morphemes themselves play no active role, their only functionality being to define morphological families. Hence, our model is not faced with the enigma of having to explain why a morpheme that is crucial for providing access to a full form has a marginal effect at best: In our model, morphemes do not mediate access.

A second important difference concerns how competition at the level of full forms is dealt with. In the two-layer architecture, a word's constituent morphemes will activate a range of complex forms. The surface frequency of a complex form is then assumed to reflect the speed with which it can suppress its full-form competitors. What is not taken into account here is that the number of lexical competitors, and their frequencies, will co-determine how quickly the target complex word will reach threshold activation. Our information-theoretical model shows that in an interactive activation model such as proposed in Taft (1994), the costs of resolving the competition between full forms *must* balance the benefits of the prior activation of these full forms. In other words, most of the work done by the base has to be undone. The net outcome in the lexical decision latencies is that the effect of the base is marginal at best.

Another issue that we have addressed is the possibility that lexical decision is more sensitive to meaning than word naming. Two sources of evidence in our data provide further support for this claim. First, additional cross-task comparisons showed that the regression coefficients for the surface frequency effects in lexical decision were more than two times those in naming (see Tables 3-6) in models that controlled for transparency. This is in line with the conclusion of Baayen et al. (2006) that the word frequency effect in lexical decision can largely be attributed to semantics.

Second, the amount of variance that is attributable to the items is larger in lexical decision than in naming. Lexical decision is exquisitely sensitive to the words' semantics and naming is not, so lexical decision taps into a broader spectrum of lexical properties, and thus more extensive item variability is expected. To see this, we proceeded as follows. We first calculated the proportion of explained variance for each mixed-effect model in Tables 3 - 6, which had both subject and item as random effects. We then computed the explained variance for the same models, excluding item but retaining subject as random effect, while maintaining all fixed effect predictors. The drop in explained variance is a measure of the importance of those item-specific properties that are not already accounted for by the fixed effects. For visual naming, $R^2$ dropped by 3.5%, whereas for visual lexical decision it dropped by 8.2%. For auditory processing the pattern was even more dramatic: $R^2$ dropped by 2.1% for auditory naming, but by 9.8% for auditory lexical decision. This shows that item variability is much greater in lexical decision. We interpret this as further evidence for the importance of semantics for lexical decision. If correct, this also indicates that further research is required to develop measures that better capture a word's semantics.

The current study also makes some further methodological contributions to the literature. One contribution has to do with how we accounted for longitudinal effects in the data. It is interesting to note that the previous-trial principal components were among the strongest variables in our statistical models, and that at least two of the three components were significant in all four tasks (see Figures 1 and 2). Our findings converge with Taylor and Lupker's (2001) study on sequential RT dependencies, and we extend this line of research by showing that such effects go back at least to the four previous trials. The unexpectedly large magnitude of these effects suggests that they need to be taken seriously in future studies, if only to dramatically reduce error variance. Similarly, the simple effect of trial number also explains variance in most subexperiments, capturing either fatigue or learning that goes on during the course of the experiment. Such temporal correlational structure, ignored in traditional analyses based on $F_1$ and $F_2$ averaging procedures, can be brought under statistical control in mixed-effects regression analyses. This, too, reduces error variance and improves statistical power. In addition, we replicated several effects in this growing literature on inflectional entropy and cohort entropy (Moscoso et al., 2004b; Tabak et al., 2005; Wurm et al., 2006). Finally, our embedded and matrix entropy measures allowed us to replicate the main findings of Bowers et al. (2005) with lexical decision rather than semantic categorization.

For research on morphological processing, the methodological contribution of this study is the partitioning of the set of lexical candidates, combined with assessing the weight of the resulting subsets by means of dedicated distributional measures that take into account the different dynamics of lexical processing in visual and auditory word recognition. For auditory processing, we contrasted cohort competitors through the UP location, and morphological competitors through cohort entropy. For visual processing, we assessed derivational competitors by means of the family size measure and our measure of affix productivity, inflectional competitors through the inflectional entropy measure, and non-morphological competitors through matrix entropy and embedded entropy. In both modalities, contextual dependencies were assessed with the surface frequency measure as an estimate of constituents' joint probability. The results obtained show that our partitioning of the set of lexical candidates contributes to our understanding the dynamics of lexical processing. Most importantly, we have clarified that it is crucial to distinguish for lexical decision between the facilitation from morphologically related lexical candidates and the inhibition from non-morphologically-related candidates.

References


Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40,* 41-61.

Baayen, R.H. (1991). A stochastic process for word frequency distributions. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 271-278.

Baayen, R.H. (1994). Productivity in language production. *Language and Cognitive Processes, 9,* 447-469.

Baayen, R.H. (2001). *Word frequency distributions.* Dordrecht, The Netherlands: Kluwer.

Baayen, R.H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69-83). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baayen, R.H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37,* 94-117.

Baayen, R.H., Feldman, L.B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55,* 290-313.

Baayen, R.H., McQueen, J.M., Dijkstra, T., & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In R.H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 355-390). Berlin: Mouton de Gruyter.

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

Baayen, H., & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language, 72,* 69-96.

Baayen, R.H., Schreuder, R., de Jong, N.H., & Krott, A. (2002). Dutch inflection: The rules that prove the exception. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and computation in the language faculty* (pp. 61-92). Dordrecht, The Netherlands: Kluwer.

Baayen, R.H., Tweedie, F.J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language, 84,* 55-65.

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., Treiman, R. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. http://elexicon.wustl.edu/, Washington University.

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., & Yap, M.J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283-316.

Bates, D.M. (2005). Fitting linear mixed models in R. *R News, 5,* 27-30.

Bates, D.M., & Sarkar, D. (2005). The lme4 library. <http://lib.stat.cmu.edu/R/CRAN/>.

Bates, E., Devescovi, A., Pizzamiglio, L., & D'Amico, S. (1995). Gender and lexical access in Italian. *Perception and Psychophysics, 57,* 847-862.

Bates, E., & Liu, H. (1996). Cued shadowing. *Language and Cognitive Processes, 11,* 577-581.

Beckwith, R., Fellbaum, C., Gross, D., & Miller, G. (1991). WordNet: A lexical database organized on psycholinguistics principles. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 211-232). Hillsdale, NJ: Lawrence Erlbaum Associates.

Belsley, D.A., Kuh, E., & Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Bergman, M.W., Hudson, P.T.W., & Eling, P.A.T.M. (1988). How simple complex words can be: Morphological processing and word representations. *Quarterly Journal of Experimental Psychology, 40A,* 41-72.

Bertram, R., Baayen, R.H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language, 42,* 390-405.

Bertram, R., Schreuder, R., & Baayen, R.H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Congition, 26,* 489-511.

Bertram, R. & Hyona, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language, 48*, 615-634.

Bowers, J.S., Davis, C.J., & Hanley, D.A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language, 52,* 131-143.

Brysbaert, M., Lange, M., & Wijnendaele, I.V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology, 12,* 65-85.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition, 28,* 297-332.

Carroll, J.B., White, M.N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology, 25,* 85-95.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance* (vol. VI, pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis, M.H., Marslen-Wilson, W.D., & Gaskell, M.G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 218-244.

de Jong, Feldman, L.B., Schreuder, R., Pastizzo, M., & Baayen, R.H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language, 81,* 555-567.

de Jong, Schreuder, R., & Baayen, R.H. (2003). Morphological resonance in the mental lexicon. In R.H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 65-88). Berlin: Mouton de Gruyter.

de Saussure, F. (1916/1966). *Course in general linguistics*. New York: McGraw.

de Vaan, L.W.M., Schreuder, R., & Baayen, R.H. (2006). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. Manuscript submitted for publication.

Diependaele, K., Sandra, D., & Grainger, J. (2005). Masked cross-modal morphological priming: Unravelling morpho-orthographic and morpho-semantic influences in early word recognition. *Language and Cognitive Processes, 20,* 75-114.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R., & Baayen, R.H. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes, 20,* 7-41.

Ellis, A.W., & Lambon Ralph, M.A. Source: (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1103-1123.

Ernestus, M., Lahey, F., Verhees, F., & Baayen, R.H. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America, 120*, 1040-1051.

Faraway, J.J. (2006). *Extending linear models with R: Generalized linear, mixed effects, and nonparametric regression models.* Boca Raton, FL: CRC Press.

Feldman, L.B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1431-1444.

Fellbaum, C.E. (1998). *WordNet: An electronic database.* Cambridge, MA: MIT Press.

Ford, M.A., Marslen-Wilson, W.D., & Davis, M.H. (2003). Morphology and frequency: Contrasting methodologies. In R.H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 89-124). Berlin: Mouton de Gruyter.

Gerhand, S. & Barry, C. (1999). Age-of-acquisition and frequency effects is speeded word recognition. *Cognition,* 73, 27-36.

Giraudo, H, & Grainger, J. (2000). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin & Review, 8,* 127-131.

Giraudo, H, & Grainger, J. (2003). On the role of derivational affixes in recognizing complex words: Evidence from masked priming. In R.H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 209-232). Berlin: Mouton de Gruyter.

Gonnerman, L. (1999). *Morphology and the lexicon: Exploring the semantics-phonology interface*. PhD thesis, University of Southern California.

Gonnerman, L., & Andersen, E. (2002). Graded semantic and phonological similarity effects in morphologically complex words. In S. Bendjaballah, W.U. Dressler, O. Pfeiffer, & M.D.

Voeikova (Eds.), *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000* (pp.137-148). John Benjamins.

Gordon, P., & Alegre, M. (1999). Is there a dual system for regular inflections? *Brain and Language, 68,* 212-217.

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., & Farioli, F. (2003). Masked repetition and phonological priming within and across modalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1256-69.

Harm, M.W., & Seidenberg, M.S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111,* 662-720.

Harrell, F.E. (2001). *Regression modeling strategies.* New York: Springer.

Hay, J.B., & Baayen, R.H. (2002). Parsing and productivity. In G.E. Booij & J. Van Marle (Eds.), *Yearbook of morphology 2001* (pp. 203-235). Dordrecht, The Netherlands: Kluwer.

Hay, J.B., & Baayen, R.H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9,* 342-348.

Joanisse, M., & Seidenberg, M.S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, 96,* 7592-7597.

Johnson, N.F., & Pugh, K.R. (1994). A cohort model of visual word recognition. *Cognitive Psychology, 26,* 240-346.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. (2005a). Prosodic cues for morphological complexity: The case of Dutch noun plurals. *Memory and Cognition*, 33:430-446.

Kemps, R., Wurm, L.H., Ernestus, M., Schreuder, R., & Baayen, R.H. (2005). Prosodic cues for morphological complexity: Comparatives and agent nouns in Dutch and English. *Language and Cognitive Processes, 20,* 43-73.

Kuperman, V., Bertram, R., and Baayen, R. H. (2007a). Morphemes as opportunities for the processing of trimorphemic Finnish compounds. *manuscript submitted for publication*, Radboud University Nijmegen:1-37.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2007b). Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *manuscript submitted for publication*, Radboud University Nijmegen:1-26.

Laudanna, A., Cermele, A., & Caramazza, A. (1997). Morpho-lexical representations in naming. *Language and Cognitive Processes, 12,* 49-66.

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

Luce, P.A., & Cluff, M.S. (1998). Delayed commitment in spoken word recognition: Evidence from cross-modal priming. *Perception & Psychophysics, 60,* 484-490.

Luce, P.A., & Lyons, E.A. (1999). Processing lexically embedded spoken words. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 174-183.

Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. In H. Bouma & D.G. Bouwhuis (Eds.), *Attention and performance: Control of language processes* (vol. X, pp. 125-150). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marslen-Wilson, W. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *Lexical representation and processing* (pp. 3-24). Cambridge, MA: MIT Press.

Marslen-Wilson, W.D., Tyler, L.K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101,* 3-33.

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

McQueen, J.M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 621-638.

Miellet, S., & Sparrow, L. (2004). Phonological codes are assembled before word fixation: evidence from boundary paradigm in sentence reading. *Brain and Language, 90,* 299-310.

Miller, G.A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-312.

Morrison, C.M., & Ellis, A.W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91,* 167-180.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R.H. (2004a). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1271-1278.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R.H. (2004b). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94,* 1-18.

New, B., Brysbaert, M., Segui, F.L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language, 51*, 568-585.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M., (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13,* 45-52.

Newman, R.L., & Connolly, J.F. (2004). Determining the role of phonology in silent reading using event-related brain potentials. *Cognitive Brain Research, 21,* 94-105.

Norris, D., McQueen, J.M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1209-1228.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS.* New York: Springer.

Pinker, S. (1999). *Words and rules: The ingredients of language*. London: Weidenfeld and Nicolson.

Pitt, M. (1994). *Lexical competition: The case of embedded words.* Paper presented at the 85th Annual Meeting of the Psychonomic Society, Washington, DC.

Pluymaekers, M., Ernestus, M. & Baayen, R.H. (2005). Frequency and acoustic length: The case of derivational affixes in Dutch. Manuscript submitted for publication.

Prather, P., & Swinney, D. (1977). *Some effects of syntactic context upon lexical access.* Paper presented at the 85th Annual Convention of the American Psychological Association, San Francisco, CA.

Radeau, M., Mousty, P., & Bertelson, P. (1989). The effect of the uniqueness point in spoken-word recognition. *Psychological Research, 51,* 123-128.

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,* 1090-1098.

Salverda, A., Dahan, D., & McQueen, J. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90,* 51-89.

Sanders, A.F. (1998). *Elements of human performance: Reaction processes and attention in human skill.* Mahwah, NJ: Lawrence Erlbaum Associates..

Schreuder, R., & Baayen, R.H. (1995). Modeling morphological processing. In L.B. Feldman, (Ed.), *Morphological aspects of language processing* (pp. 131-154). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schreuder, R., & Baayen, R.H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37,* 118-139.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 24-49). Cambridge, MA: MIT Press.

Seidenberg, M.S., & Gonnerman, L. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences, 4,* 353-361.

Sereno, J., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory and Cognition 25*, 425-437.

Stanners, R.F., Neiser, J.J., Hernon, W.P., & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior, 18,* 399-412.

Tabak, W., Schreuder, R., & Baayen, R.H. (2005). Lexical statistics and lexical processing: Semantic density, information complexity, sex, and irregularity in Dutch. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives.* Berlin: Mouton de Gruyter, pp. 529-555.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition, 7*, 263-272.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9,* 271-294.

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

Taylor, T.E., & Lupker, S.J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 117-138.

Ullman, M. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition*, 92:231-270.

Vroomen, J. & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 710-720.

Wurm, L.H. (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language, 37,* 438-461.

Wurm, L.H. (2000). Auditory processing of polymorphemic pseudowords. *Journal of Memory and Language, 42,* 255-271.

Wurm, L.H., Ernestus, M., Schreuder, R., & Baayen, R.H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and Conditional Root Uniqueness Points. *The Mental Lexicon, 1,* 125-146.

Wurm, L.H., & Ross, S.E. (2001). Conditional root uniqueness points: Psychological validity and perceptual consequence. *Journal of Memory and Language, 45,* 39-57.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32,* 25-64.

Author Note

We would like to thank Chris Westbury for his generous assistance in generating pseudowords for Experiment 1. We are also indebted to Sean Seaman and Heather Ulstad for help with data collection. Please address all correspondence to Harald Baayen, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands [email: baayen@mpi.nl]

Table 1

*Summary Statistics for Critical Items*

| Variable | Prefixed derivations | Suffixed derivations | Suffixed inflections |
|---|---|---|---|
| Mean length (letters) | 7 | 7 | 8 |
| Mean UP location (msec) | 403 | 331 | 389 |
| Mean UP-to-offset duration (msec) | 247 | 263 | 326 |
| Median surface frequency | 1 | 1 | 3 |
| Median root frequency | 197 | 70 | 52 |
| Mean semantic transparency | 4 | 5 | 6 |
| Mean productivity | .005 | .003 | --- |

Table 2

*Summary of Overall Regression Analysis for Variables Predicting Response Latency: Interactions With Task. (The model has random intercepts for subject (s = 75.4) and item (s = 29.0). $s_\varepsilon$ = 126.1.)*

| Interaction | $df_N$ | *SS* | *MS* | $df_D$ | *F* | *p* |
|---|---|---|---|---|---|---|
| Task x: | | | | | | |
| Trial number | 3 | 19082341 | 6360780 | 34415 | 400.2549 | < .0001 |
| PC1 | 3 | 3598046 | 1199349 | 34415 | 75.4695 | < .0001 |
| PC2 | 3 | 132276 | 44092 | 34415 | 2.7745 | .0398 |
| PC3 | 3 | 136822 | 45607 | 34415 | 2.8699 | .0350 |
| Place of articulation | 6 | 1507090 | 251182 | 34415 | 15.8057 | < .0001 |
| Voiced onset | 3 | 7520760 | 2506920 | 34415 | 157.7491 | < .0001 |
| Affix type | 6 | 4614010 | 769002 | 34415 | 48.3898 | < .0001 |
| Surface frequency | 3 | 1023991 | 341330 | 34415 | 21.4784 | < .0001 |
| UP location | 3 | 13436881 | 4478960 | 34415 | 281.8406 | < .0001 |
| UP-to-offset duration | 3 | 2865386 | 955129 | 34415 | 60.1019 | < .0001 |
| Cohort entropy | 3 | 1136457 | 378819 | 34415 | 23.8374 | < .0001 |
| Semant. transparency | 3 | 1234557 | 411519 | 34415 | 25.8950 | < .0001 |
| Family size (linear) | 3 | 208467 | 69489 | 34415 | 4.3726 | .0044 |
| Family size (quad.) | 3 | 314923 | 104974 | 34415 | 6.6056 | .0002 |
| Embedded entropy | 3 | 241857 | 80619 | 34415 | 5.0730 | .0016 |

(table continues)

| | | | | | | |
|---|---|---|---|---|---|---|
| Matrix entropy | 3 | 250681 | 83560 | 34415 | 5.2581 | .0013 |
| Inflectional entropy[a] | 3 | 137800 | 45933 | 34415 | 2.8904 | .0340 |
| Word length (linear) | 3 | 1298266 | 432755 | 34415 | 27.2313 | < .0001 |
| Word length (quad.) | 3 | 23086 | 7695 | 34415 | 0.4842 | .6932 |
| Root frequency | 3 | 303829 | 101276 | 34415 | 6.3729 | .0003 |

[a]Inflectional entropy was factorized (zero vs. nonzero entropy) because the distribution of continuous values was extremely skewed.

Table 3

*Summary of the Multilevel Analysis of Covariance for Variables Predicting Visual Naming Latency. (The model has random intercepts for subject (s = 54.6) and item (s = 31.6).* $s_\varepsilon$ *= 76.0.) The by-item* $R^2$ *= .459 and the bootstrap-adjusted by-item* $R^2$ *= .419.*

| Variable | $df_N$ | *SS* | *MS* | $df_D$ | *F* | *p* |
|---|---|---|---|---|---|---|
| Factors | | | | | | |
| Voiced onset | 1 | 605142 | 605142 | 8848 | 104.8524 | < .0001 |
| Affix type | 2 | 313178 | 156589 | 8848 | 27.1320 | < .0001 |

| Variable | Regression coefficient (*B*) | Standard error of *B* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Continuous variables | | | | | |
| Trial number | 0.3318 | 0.0098 | 8848 | 33.9085 | < .0001 |
| PC1 | 27.3349 | 1.4971 | 8848 | 18.2590 | < .0001 |
| PC2 | 5.4077 | 1.6650 | 8848 | 3.2478 | .0012 |
| PC3 | -8.6576 | 1.7735 | 8848 | -4.8817 | < .0001 |
| Word length (linear) | -13.2084 | 9.4308 | 8848 | -1.4006 | .1614 |
| Word length (quadratic) | 1.3925 | 0.6247 | 8848 | 2.2291 | .0258 |
| Surface frequency | -4.4463 | 1.8452 | 8848 | -2.4097 | .0160 |
| Root morpheme frequency | -6.8978 | 3.2731 | 8848 | -2.1074 | .0351 |

Table 4

*Summary of the Multilevel Analysis of Covariance for Variables Predicting Auditory Naming Latency. (The model has random intercepts for subject (s = 100.3) and item (s = 26.8).* $s_\varepsilon$ *= 124.3.) The by-item* $R^2$ *= .614 and the bootstrap-adjusted by-item* $R^2$ *= .589.*

| Variable | $df_N$ | *SS* | *MS* | $df_D$ | *F* | *p* |
|---|---|---|---|---|---|---|
| Factors and interactions with factors | | | | | | |
| Affix type | 2 | 599538 | 299769 | 10284 | 19.4018 | < .0001 |
| Affix type x surface frequency | 2 | 96036 | 48018 | 10284 | 3.1079 | .0447 |

| Variable | Regression coefficient (*B*) | Standard error of *B* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Continuous variables | | | | | |
| Trial number | 0.3257 | 0.0133 | 10284 | 24.4824 | < .0001 |
| PC1 | 46.3460 | 1.5896 | 10284 | 29.1563 | < .0001 |
| PC2 | 10.5500 | 2.0491 | 10284 | 5.1487 | < .0001 |
| UP location | 0.3991 | 0.0263 | 10284 | 15.1986 | < .0001 |
| UP-to-offset duration | 0.3625 | 0.0243 | 10284 | 14.9442 | < .0001 |
| Surface frequency | -5.8346 | 2.3905 | 10284 | -2.4407 | .0147 |

Table 5

*Summary of the Multilevel Analysis of Covariance for Variables Predicting Visual Lexical Decision Time. (The model has random intercepts for subject (s = 82.0) and item (s = 45.7), as well as a by-subject random slope for surface frequency (s = 7.0, correlation with intercept r = -.66), and a by-subject random slope for semantic transparency (s = 5.9, correlation with intercept = -.35; correlation between the two random slopes = -.03).* $s_\varepsilon$ *= 115.7.) The by-item* $R^2$ *= .503 and the bootstrap-adjusted by-item* $R^2$ *= .423.*

| Variable | $df_N$ | *SS* | *MS* | $df_D$ | *F* | *p* |
|---|---|---|---|---|---|---|
| Factors | | | | | | |
| Affix type | 2 | 814581 | 407291 | 7715 | 30.4066 | < .0001 |
| Inflectional entropy[a] | 1 | 138713 | 138713 | 7715 | 10.3558 | .0013 |

| Variable | Regression coefficient (*B*) | Standard error of *B* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Continuous variables | | | | | |
| PC1 | 21.6612 | 1.16390 | 7715 | 18.6109 | < .0001 |
| PC2 | 4.7801 | 1.34922 | 7715 | 3.5429 | .0004 |
| PC3 | -7.1040 | 1.36689 | 7715 | -5.1972 | < .0001 |
| Word length (linear) | -29.3769 | 14.65413 | 7715 | -2.0047 | .0450 |

(table continues)

| | | | | | |
|---|---|---|---|---|---|
| Word length (quadratic) | 1.9927 | 0.95426 | 7715 | 2.0882 | .0368 |
| Surface frequency | -12.1086 | 2.92646 | 7715 | -4.1376 | < .0001 |
| Family size (linear) | 16.8795 | 8.73954 | 7715 | 1.9314 | .0534 |
| Family size (quadratic) | -6.2264 | 1.90808 | 7715 | -3.2632 | .0011 |
| Embedded entropy | 18.5922 | 6.91707 | 7715 | 2.6879 | .0072 |
| Matrix entropy | 31.1245 | 14.39447 | 7715 | 2.1623 | .0306 |
| Semantic transparency | -14.0723 | 4.26060 | 7715 | -3.3029 | .0010 |
| UP location | -0.0941 | 0.03145 | 7715 | -2.9930 | .0028 |

[a]Inflectional entropy was factorized (zero vs. nonzero entropy) because the distribution of continuous values was extremely skewed.

Table 6

*Summary of the Multilevel Analysis of Covariance for Variables Predicting Auditory Lexical Decision Time. (The model has random intercepts for subject (s = 80.7) and item (s = 56.5), as well as a by-subject random slope for surface frequency (s = 8.7, correlation with intercept r = -.57).* $s_\varepsilon$ *= 142.4.) The by-item* $R^2$ *= .522 and the bootstrap-adjusted by-item* $R^2$ *= .473.*

| Variable | $df_N$ | *SS* | *MS* | $df_D$ | *F* | *p* |
|---|---|---|---|---|---|---|
| Factors | | | | | | |
| Voiced onset | 1 | 338620 | 338620 | 7126 | 16.6985 | < .0001 |
| Affix type | 2 | 1095712 | 547856 | 7126 | 27.0166 | < .0001 |
| Place of articulation | 2 | 402086 | 201043 | 7126 | 9.9141 | < .0001 |

| Variable | Regression coefficient (*B*) | Standard error of *B* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Continuous variables | | | | | |
| Trial number | -0.0297 | 0.0092 | 7126 | -3.2188 | .0013 |
| PC1 | 21.8600 | 1.6977 | 7126 | 12.8765 | < .0001 |
| PC2 | 4.2023 | 1.9164 | 7126 | 2.1928 | .0284 |
| PC3 | -4.0378 | 1.9899 | 7126 | -2.0291 | .0425 |
| UP location | 0.5217 | 0.0548 | 7126 | 9.5273 | < .0001 |

(table continues)

| | | | | | |
|---|---|---|---|---|---|
| UP-to-offset duration | 0.2694 | 0.0490 | 7126 | 5.4982 | < .0001 |
| Surface frequency | -10.9720 | 3.6338 | 7126 | -3.0193 | .0025 |
| Cohort entropy | -21.3200 | 8.0707 | 7126 | -2.6416 | .0083 |
| Semantic transparency | -25.8320 | 5.4314 | 7126 | -4.7560 | < .0001 |

Table 7

*Coefficients and associated statistics for the visual lexical decision latencies for 8486 bimorphemic words in the English Lexicon Project. MC mean: mean value of the coefficient across 5000 Markov chain Monte Carlo samples of the posterior distribution of the parameters. HPD lower/HPD upper: Highest Posterior Density interval for 95% of the probability density, pMC: the corresponding probability, p(t): probability based on the t-distribution with 8466 degrees of freedom. Contrast coding was used for factors.*

| Variable | Estimate | MC mean | HPD lower | HPD upper | pMC | p(t) |
|---|---|---|---|---|---|---|
| Intercept | 6.6401 | 6.6405 | 6.5959 | 6.6774 | 0.0002 | 0.0000 |
| Length (linear) | 0.0119 | 0.0119 | 0.0085 | 0.0155 | 0.0002 | 0.0000 |
| Length (quadratic) | 0.0028 | 0.0027 | 0.0022 | 0.0033 | 0.0002 | 0.0000 |
| Root frequency | -0.0048 | -0.0048 | -0.0062 | -0.0031 | 0.0002 | 0.0000 |
| Surface frequency (linear) | -0.0371 | -0.0371 | -0.0400 | -0.0338 | 0.0002 | 0.0000 |
| Surface frequency (quadratic) | 0.0011 | 0.0011 | 0.0003 | 0.0020 | 0.0068 | 0.0068 |
| Affix type: suffix | -0.0277 | -0.0279 | -0.0454 | -0.0099 | 0.0020 | 0.0015 |
| NSyll | 0.0153 | 0.0153 | 0.0101 | 0.0205 | 0.0002 | 0.0000 |
| V | -0.0095 | -0.0095 | -0.0161 | -0.0029 | 0.0056 | 0.0058 |
| Hi | -0.0070 | -0.0070 | -0.0110 | -0.0029 | 0.0004 | 0.0006 |
| LogBigFreq | 0.0062 | 0.0062 | 0.0044 | 0.0080 | 0.0002 | 0.0000 |
| BNCd | -0.0237 | -0.0237 | -0.0281 | -0.0191 | 0.0002 | 0.0000 |
| SynWord | -0.0287 | -0.0286 | -0.0357 | -0.0217 | 0.0002 | 0.0000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| MorphType: infl | 0.0460 | 0.0460 | 0.0049 | 0.0906 | 0.0356 | 0.0334 |
| VfB | -0.0077 | -0.0077 | -0.0104 | -0.0046 | 0.0002 | 0.0000 |
| Root freq x surface freq | 0.0021 | 0.0021 | 0.0013 | 0.0030 | 0.0002 | 0.0000 |
| BNCd x MorphType:infl | 0.0091 | 0.0091 | 0.0037 | 0.0141 | 0.0012 | 0.0006 |
| SynWord x MorphType:infl | 0.0179 | 0.0179 | 0.0101 | 0.0257 | 0.0002 | 0.0000 |
| Surface freq x MorphType:infl | 0.0166 | 0.0166 | 0.0085 | 0.0259 | 0.0002 | 0.0001 |
| LogBigFreq x MorphType:infl | -0.0050 | -0.0050 | -0.0072 | -0.0030 | 0.0002 | 0.0000 |

Table 8

*Probability and information in the lexicon at different timesteps.*

| *time* | *probability* | *information* |
|---|---|---|
| $t_0$ | $\Sigma_i Pr(m_i)$ | 0 |
| $t_1$ | f(good+)/N | $-log_2$(f(good+))+$log_2$(N) |
| $t_2$ | f(good+)/f(good+) | 0 |
| | f(good, ness)/f(good+) | $-log_2$(good,ness)+$log_2$(f(good+)) |
| $t_3$ | f(good, ness)/f(good,ness) | 0 |

Appendix -- Critical Items

*Prefixed derivations:* aback, abed, abreast, ado, afar, afield, afoot, afresh, aground, alight, anew, aright, atop, befall, befit, befriend, beget, belay, beset, bespeak, betake, bethink, debone, decamp, decontrol, deflower, deform, descale, detrain, disable, disaffect, disbelieve, discourse, dismember, disorder, disrespect, disunion, disused, embitter, emboss, empanel, foresight, imbalance, impound, impractical, inexperience, inhuman, instate, insure, outgrow, outright, outside, overshoe, overstate, overtone, rearrange, recharge, recreate, redress, reforest, rejoin, repress, reprove, resettle, subdivide, sublet, subserve, underarm, underdog, understudy

*Suffixed derivations:* boarder, boiler, brainy, branchy, brighten, camper, challenger, cheapen, cloudy, creditor, dreamer, fitful, flashy, flighty, freezer, freshen, fruitful, gainful, grandly, hellish, hooker, hurtful, islander, joker, jumper, knocker, leafy, madly, marcher, meaty, milky, mixer, moony, narrowly, packer, pinkish, quicken, rower, sander, searcher, seeker, selector, shapely, sicken, sinker, snowy, solidly, soulful, spreader, squarely, strainer, stretcher, suitor, sweeper, sweeten, tallish, thankful, thinly, throaty, tiller, toothy, tracker, warmish, washer, wasteful, weakly, weighty, wetly, woody, yellowish

*Suffixed inflections:* absences, adjusting, advertises, affords, announces, appointing, authorizing, bays, behaviors, bibles, captains, chests, cleans, commits, compares, composing, concludes, confronts, constituting, constructs, convinces, councils, creations, defends, denies, destroys, diameters, dictionaries, discovers, doctrines, dramas, eats, faiths, finishes, forts, ignores, incomes, installs, interiors, isolates, jumps, justices, leans, listens, majorities, naming, operas, possessing, pouring, preferring, promotes, proposing, regarding, rejecting, reminding, republics, responding, retires, roots, saints, saves, settles, sevens, solves, suffers, suns, surrounds, televisions, tendencies, tuesdays, tying, uncles

*Note.* Items were presented in ALL CAPS in the visual conditions.

*Figure Captions*

*Figure 1.* Results of analyses predicting naming times. Significant effects for visually-presented stimuli are shown by black lines; those for spoken stimuli are shown in gray. For the panel showing the word frequency effect, the interaction with affix type (auditory naming) is indicated by multiple gray lines: the solid line represents prefixed derivations, the dashed line represents suffixed derivations, and the dotted line represents suffixed inflections.

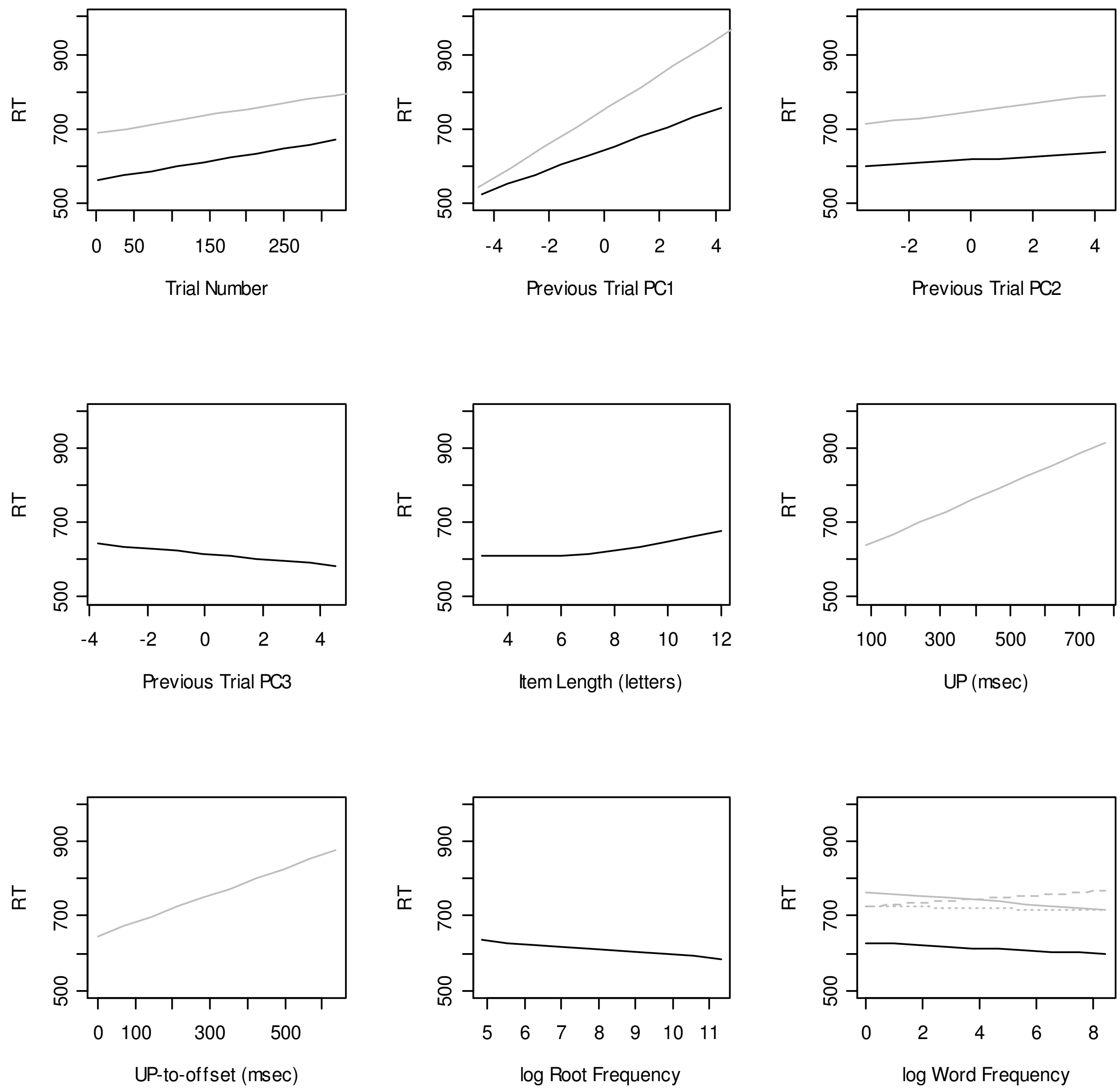
*Figure 2.* Results of analyses predicting lexical decision times. Significant effects for visually-presented stimuli are shown by black lines; those for spoken stimuli are shown in gray.

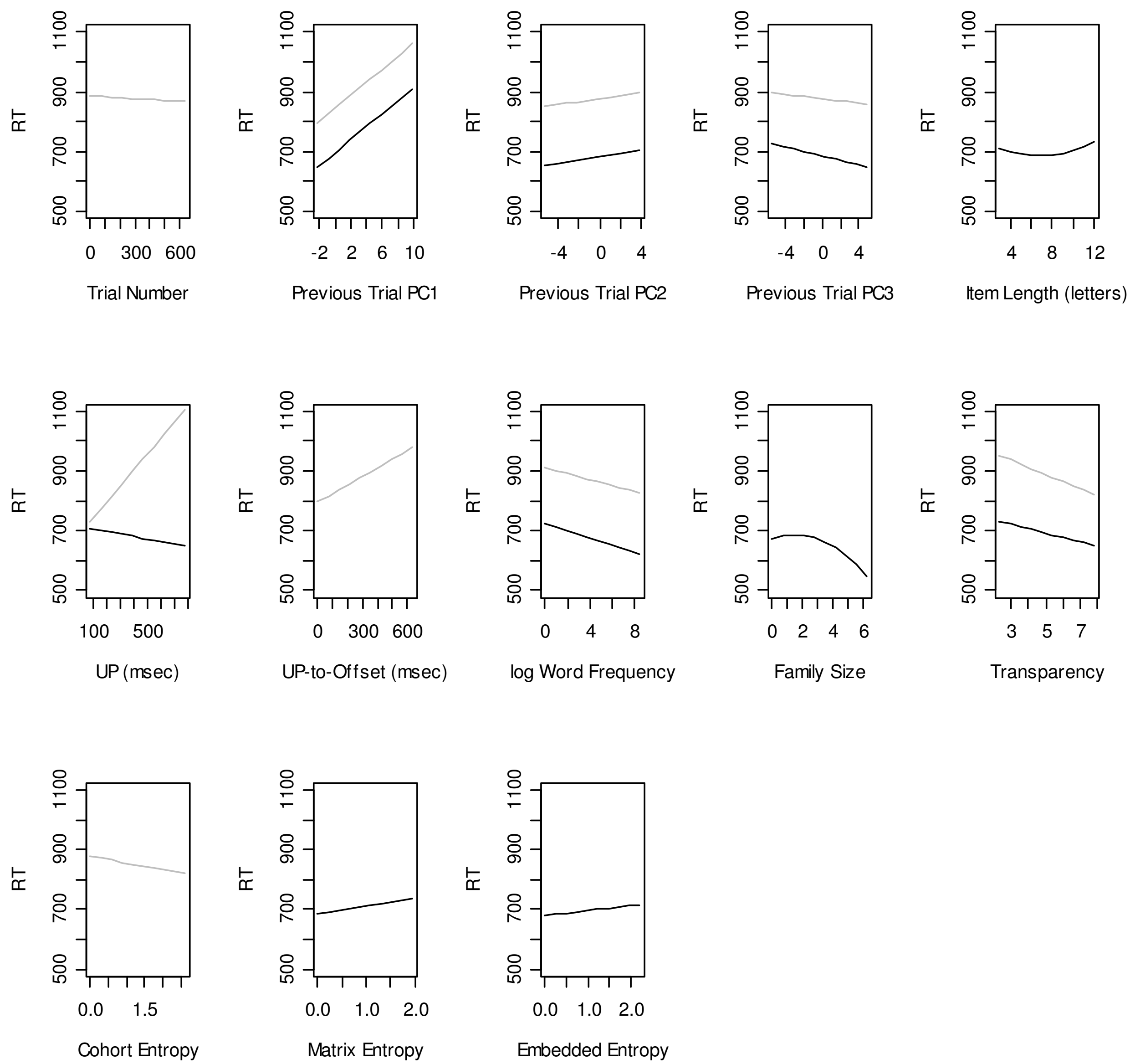
*Figure 3.* The effects of surface frequency (left) and root frequency (right) in visual lexical decision. In these panels, the graphs are adjusted for the median of the other frequency measure in order to take their interaction into account. Black lines represent derived words, gray lines represent inflected words.

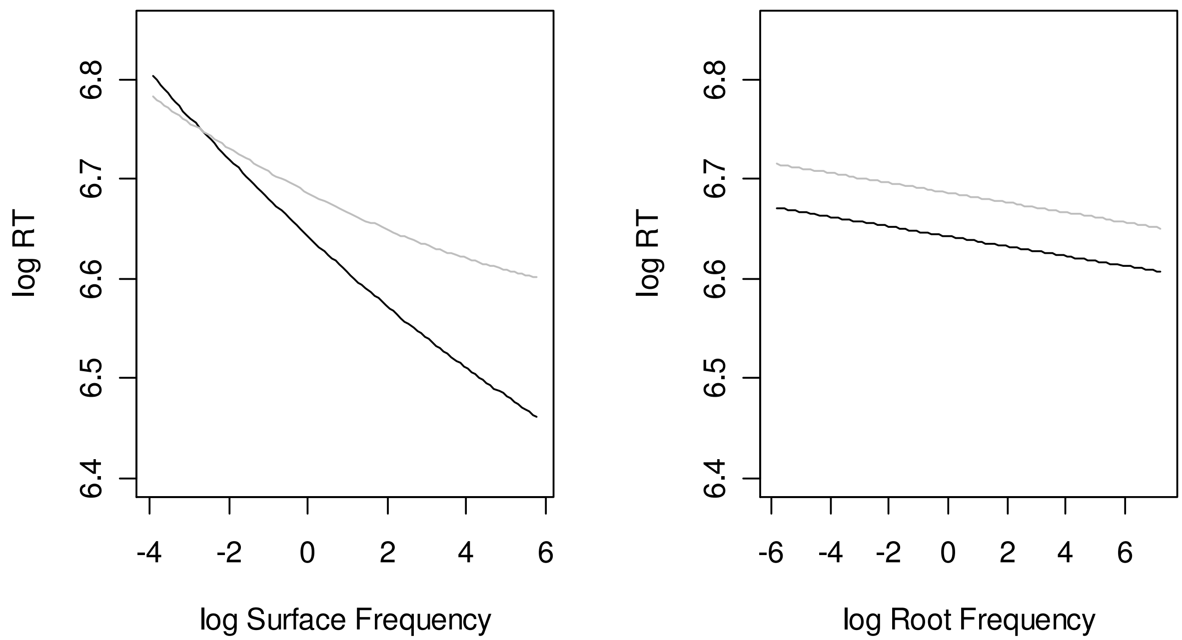
*Figure 4.* The interaction of root frequency and surface frequency in visual lexical decision. Lines in the gray plane represent deciles of the empirical distributions.

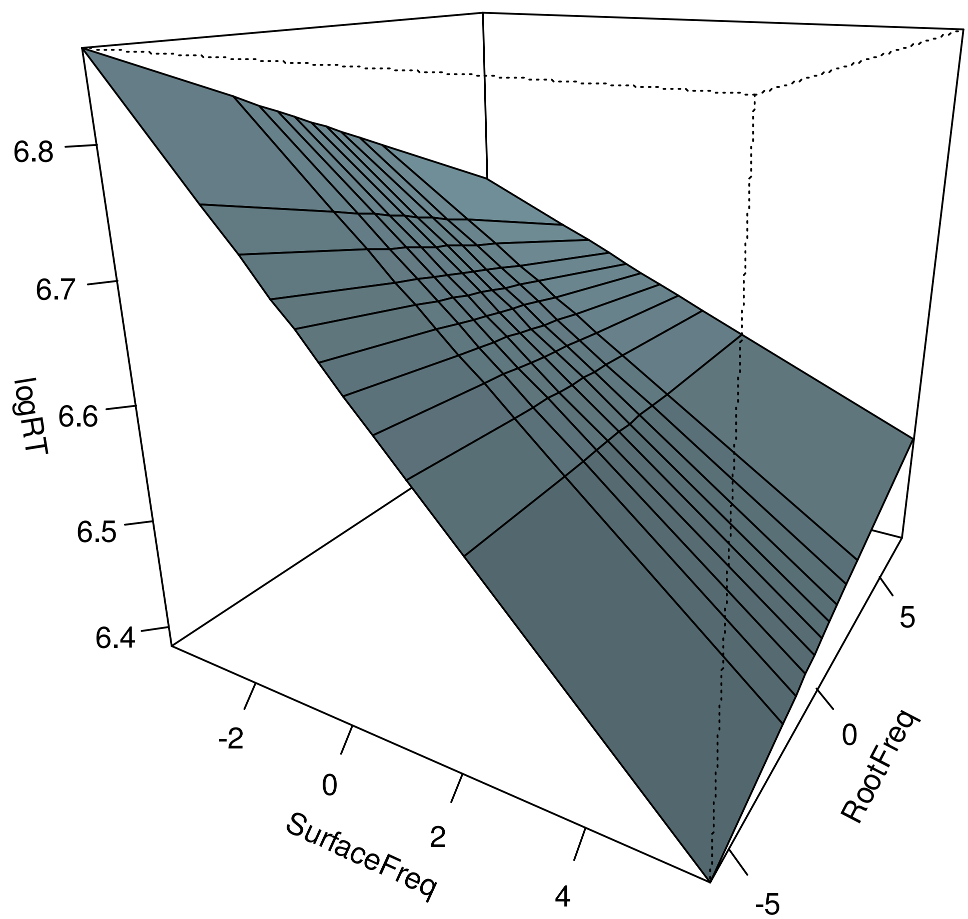
*Figure 5.* Random slopes and random intercepts for surface frequency for the inflectional (large font) and derivational (small font) affixes in the English Lexicon Project. Suffixes are shown in black, prefixes in gray.

*Figure 6.* Proportion of samples (without replacement) with significant effect as a function of sample size for derived and inflected words (upper left), for inflected word (upper right), and for inflected words below the supposed threshold of 6 per million (lower left).

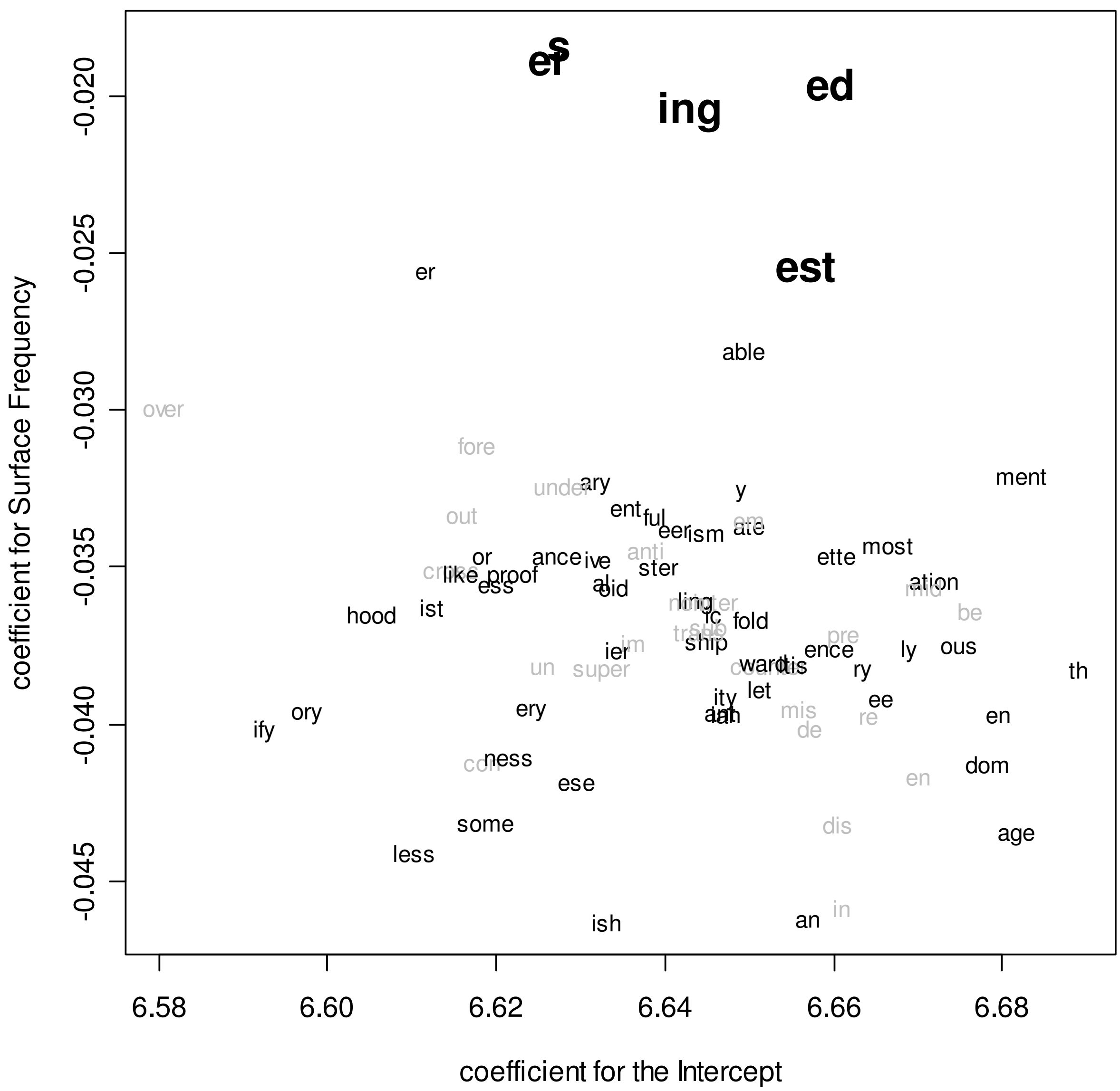

RT
500
700
900
0
50
150
250
Trial Number
-4
-2
0
2
4
Previous Trial PC1
Previous Trial PC2
Previous Trial PC3
4
6
8
10
12
Item Length (letters)
100
300
500
700
UP (msec)
UP-to-offset (msec)
5
6
7
8
9
10
11
log Root Frequency
log Word Frequency

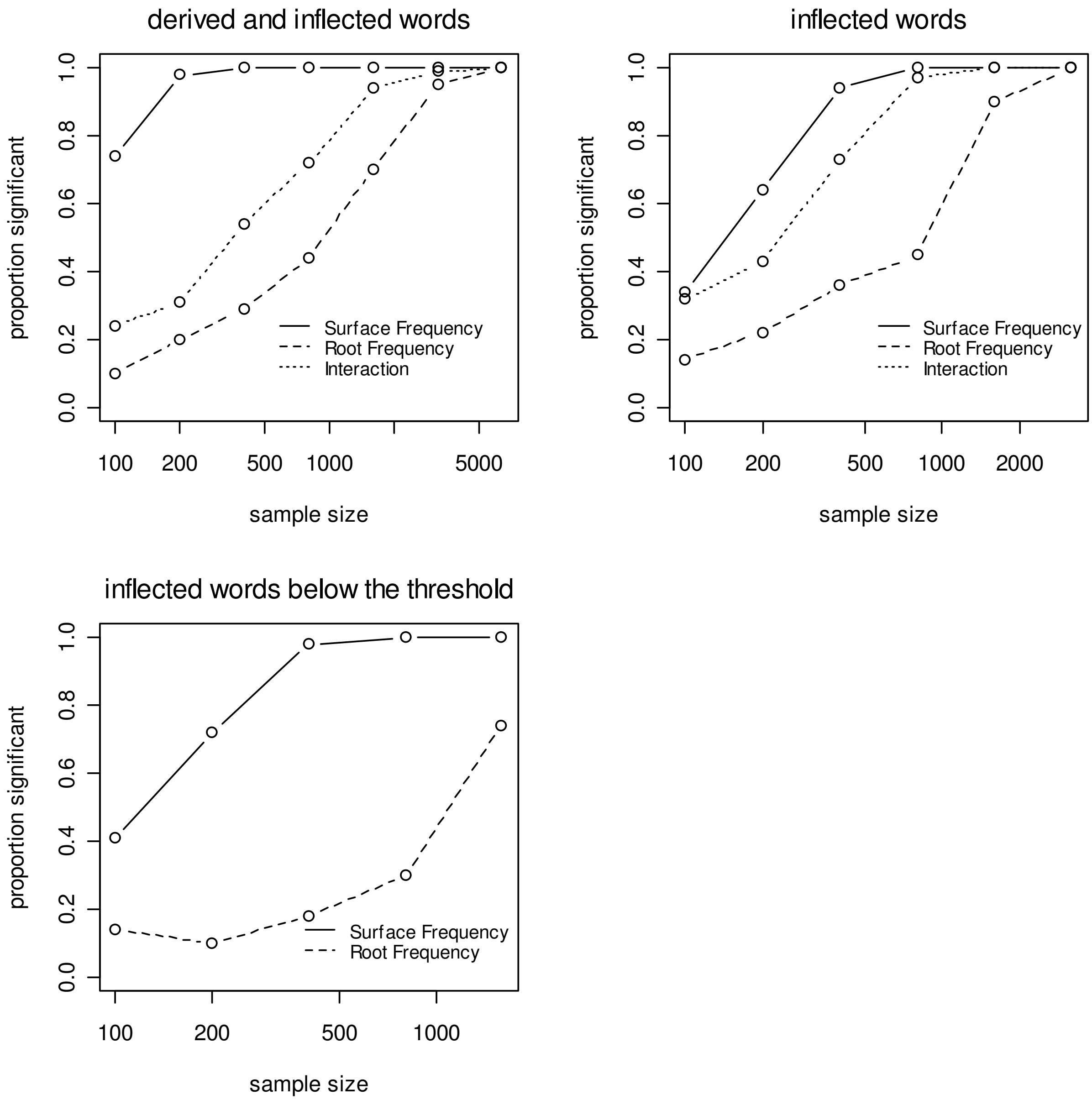

RT
1100
900
700
500
0 300 600
Trial Number
-2 2 6 10
Previous Trial PC1
-4 0 4
Previous Trial PC2
Previous Trial PC3
4 8 12
Item Length (letters)
100 500
UP (msec)
0 300 600
UP-to-Offset (msec)
0 4 8
log Word Frequency
0 2 4 6
Family Size
3 5 7
Transparency
0.0 1.5
Cohort Entropy
0.0 1.0 2.0
Matrix Entropy
Embedded Entropy


log RT
6.4
6.5
6.6
6.7
6.8
-4
-2
0
2
4
6
log Surface Frequency
log RT
6.4
6.5
6.6
6.7
6.8
-6
-4
-2
0
2
4
6
log Root Frequency



6.8
6.7
6.6
6.5
6.4
logRT
-2
0
2
4
SurfaceFreq
5
0
-5
RootFreq


coefficient for Surface Frequency
coefficient for the Intercept
-0.020
-0.025
-0.030
-0.035
-0.040
-0.045
6.58
6.60
6.62
6.64
6.66
6.68
s
er
ing
ed
est
er
able
over
fore
ment
ary
under
y
out
ent
ful
eer
ism
ate
most
ette
or
ance
ive
anti
ster
like
proof
ess
al
oid
ation
hood
ist
ling
fold
be
ship
pre
ier
ence
ly
ous
un
super
ward
ry
th
ity
let
ee
ory
ify
mis
de
re
en
ery
ness
ese
dom
en
some
dis
age
less
ish
an
in


derived and inflected words
proportion significant
0.0
0.2
0.4
0.6
0.8
1.0
100
200
500
1000
5000
sample size
Surface Frequency
Root Frequency
Interaction
inflected words
proportion significant
0.0
0.2
0.4
0.6
0.8
1.0
100
200
500
1000
2000
sample size
Surface Frequency
Root Frequency
Interaction
inflected words below the threshold
proportion significant
0.0
0.2
0.4
0.6
0.8
1.0
100
200
500
1000
sample size
Surface Frequency
Root Frequency

Footnotes

[1]For all of our statistical models we carefully assessed whether we had a collinearity problem because of the inclusion of such a large number of regressor variables. The condition numbers (assessed following Belsley, Kuh, and Welsch, 1980) of the individual models are approximately eight, which is an acceptable value. In the one model in which root frequency was significant the condition number is approximately 25, which would normally be cause for concern (Belsley et al., 1980). Removing root frequency from that model reduces the condition number to approximately eight, and has only negligible effects on the other coefficients. Our conclusion is therefore that collinearity is not a problem in the current analyses.

[2]In a multilevel model, it is not straightforward to assess the amount of variance explained because there are three sources of random variation: subjects, items, and the residual error. As our primary interest is in the predictivity of by-item properties for lexical processing, we therefore assessed the proportion of explained variance by means of a standard regression on the by-item means. This also allowed us to validate our models with the bootstrap. We note that in our experience, multilevel regression models with both subject and item as random effects and by-item bootstrap-validated regression models yield very similar results. By contrast, multilevel models with only subject as a random effect, and similarly random regression models (see for instance Lorch and Myers, 1990) may not validate well in the bootstrap and run a considerable risk of overfitting the data.

[3]The x-axes of the word frequency panels in our figures show the natural log of the raw frequency based on the CELEX database, which has 17.9 million tokens. Thirteen of the inflected suffixed words, which we took from Alegre and Gordon (1999), exceeded the desired cut-off of six occurrences per million. Alegre and Gordon used a smaller, older database of frequency values than that used in the current study.

[4]With respect to the random effects structure of this model, we note that in addition to random intercepts ($\sigma = 0.028$) the model incorporated random slopes for surface frequency ($\sigma = 0.007$) and word length ($\sigma = 0.009$; all $p$-values were $< .0001$). The standard deviation for the residual error was 0.089.