Laboratory Phonology
Journal of the Association for Laboratory Phonology

Stein, S. D., & Plag, I. 2022. How relative frequency and prosodic structure affect the acoustic duration of English derivatives. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 13(1): 6, pp. 1–46. DOI: https://doi.org/10.16995/labphon.6445

Open Library of Humanities

How relative frequency and prosodic structure affect the acoustic duration of English derivatives

Simon David Stein, Department of English and American Studies, Heinrich Heine University Düsseldorf, Germany, simon.stein@uni-duesseldorf.de

Ingo Plag, Department of English and American Studies, Heinrich Heine University Düsseldorf, Germany, ingo.plag@uni-duesseldorf.de

Morphological segmentability, i.e., the degree to which complex words can be decomposed into their morphological constituents, has been considered an important factor in research on morphological processing and is expected to affect acoustic duration (e.g., Hay, 2001, 2003). One way of operationalizing segmentability is through the relative frequency of a complex word to its base word. However, relative frequency has failed to affect duration for different affix categories in many previous studies. One potential reason is the fact that complex words vary in their prosodic structure, depending on the prosodic integration of the affix (Plag & Ben Hedia, 2018).

In a large corpus study with three different corpora and eight affixes each, we investigate how prosodic word structure and relative frequency influence duration, and how these two factors interact. We find that prosodic structure does not significantly interact with relative frequency. Second, we show that relative frequency effects on duration do not emerge consistently across a large number of affixes. Third, not only does prosodic word structure not explain the absence of relative frequency effects, it also often cannot account for durational differences as such. We discuss these findings in light of phonological theory and speech production models.

1. Introduction

Measures of lexical frequency have played an important role in research on morphological processing. For example, word frequency and base frequency effects on acoustic duration have been used to draw inferences about how complex words are stored in the mental lexicon (e.g., N. K. Caselli, M. K. Caselli, & Cohen-Goldberg, 2016). These frequency measures seem to be accepted in the literature also as important predictors of acoustic duration, as a number of studies have demonstrated higher frequency to be associated with more phonetic reduction (see, e.g., A. Bell et al., 2003; J. L. Bybee, 2000; Gahl, 2008; Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, 2003; Losiewicz, 1995; Pluymaekers, Ernestus, & Baayen, 2005a, 2005b). In accordance with the idea that "[w]hat you do often, you do better and faster" (Divjak, 2019, p. 1), more frequent words are more quickly produced and thus more likely to be shortened.[1]

However, a third possible frequency measure has been suggested which so far has produced rather mixed results: the *relative frequency* of a complex word compared to its base word (Hay, 2001, 2003). Relative frequency supposedly taps into morphological segmentability. If speakers encounter a base word often compared to its derivative, the derivative is more easily decomposed into its morphological constituents and is more likely to be processed compositionally. It is expected that words which are more easily segmentable, i.e., words with a stronger morphological boundary, are less likely to be phonetically reduced (Hay, 2001, 2003). However, relative frequency effects on acoustic duration have proven to be notoriously inconsistent. On the one hand, relative frequency effects on duration emerge in some studies under some conditions for some affixes (Hay, 2003, 2007; Plag & Ben Hedia, 2018); on the other, research has also yielded many null effects (Ben Hedia & Plag, 2017; Plag & Ben Hedia, 2018; Pluymaekers et al., 2005b; Schuppler, van Dommelen, Koreman, & Ernestus, 2012; Zimmerer, Scharinger, & Reetz, 2014; see also Hanique & Ernestus, 2012 for a critical review of the findings of earlier studies).

One potential reason for this seemingly inexplicable inconsistency is the prosodic structure of complex words. Plag and Ben Hedia (2018) speculate that it might be more difficult for a relative frequency effect to emerge in affixes that are phonologically integrated into the prosodic word of the derivative, compared to those that form independent prosodic words (Raffelsiefen, 1999, 2007). This is because pre-boundary lengthening caused by a stronger word-internal prosodic boundary might counteract potential reduction effects. While some previous evidence suggests that "morphological effects on fine phonetic detail cannot always be accounted for by prosodic structure" (Pluymaekers, Ernestus, Baayen, & Booij, 2010, p. 523; also see Plag, Homann, & Kunter, 2017, p. 210), there is also evidence in favor of an effect of prosodic word boundaries in complex words on duration. For example, Sproat and Fujimura (1993) show that English /l/ is longer and more likely to be realized as [ɬ] before compound-internal boundaries,

[1] Note that while this rationale is intuitive, it is at present unclear which exact combination of linguistic, cognitive, and social factors lead to the reduction of higher-frequency words (for discussions, see, e.g., Arnold & Watson, 2015; Clopper & Turnbull, 2018).

which are comparatively strong, than before affix boundaries, which are comparatively weak, or than within simplex words. Auer (2002) shows that final devoicing of /b/ in German suffixed words is characterized by a plosive release which is longer in derivatives with a word-internal prosodic word boundary than in derivatives with no such boundary. Sugahara and Turk (2009) find that segments in base-final rhymes of English affixed words preceding a stronger prosodic boundary are lengthened. Bergmann (2018) demonstrates that segments straddling a boundary in infrequent German derivatives are lengthened when these derivatives feature prosodic word-forming suffixes, compared to when they feature integrating suffixes. These findings suggest that the prosodic word structure of complex words impacts on their durational patterning and needs to be considered when investigating segmentability effects.

In addition to the suspicion that reduction effects due to segmentability might be counteracted by effects of prosodic word structure, there are also two further problems of previous research which need to be addressed. One problem is that earlier studies often only looked at few affix categories individually. This makes it difficult to compare studies (with often different methodologies) across affixes. The second problem is that different studies investigated different domains of durational variation. For example, some studies investigated the duration of the affix, while others looked at the deletion of individual segments. This renders a comparison of studies even more difficult.

The present study addresses these issues. In a broad empirical study, we test how prosodic word structure and relative frequency influence duration, and how the two factors might interact. We do this by making use of three different speech corpora of English. We model the durations of words featuring eight different affixes each, looking at both affix durations and base durations. We show that, counter to what was hypothesized by Plag and Ben Hedia (2018), prosodic word structure is not a gatekeeper for relative frequency effects. Second, we demonstrate that across a large number of affixes, relative frequency effects still do not emerge consistently. Third, we show that in addition to the fact that prosodic word structure cannot explain the absence of relative frequency effects, prosodic structure as such also often fails to explain durational differences.

The paper is structured as follows. Section 2 discusses the role of morphological segmentability, relative frequency, and prosodic structure with regard to the duration of complex words, and develops the specific hypotheses tested in this study. This is followed by a description of our methodology. In Section 4 we present our results. We close with a discussion of the implications of our findings.

2. Segmentability and relative frequency

Morphological segmentability[2] refers to the degree to which speakers can decompose a complex word into its morphological constituents. Instead of categorizing words as being either simplex

[2] In the literature, morphological *segmentability* is also known as morphological *decomposability*. For consistency, we will use the term *segmentability* throughout the paper.

(monomorphemic) or complex (multimorphemic), it is widely assumed that there are different degrees of morphological complexity on a gradient scale: Words can be more complex or less complex; morphological boundaries may be stronger or weaker. The strength of a morphological boundary in a given word can be gauged, for example, on the basis of semantic transparency, type of base, various frequential measures, and the degree of phonetic-phonological integration across that boundary.

Consider the verbs *dislike* and *discard*. Both words are considered morphologically complex, as we can identify the bases *like* and *card*, which are prefixed with *dis-*. However, it seems that one of them is more complex than the other: *Dislike* appears to be more easily decomposable than *discard*. One reason for this is because *dislike* is semantically more transparent than *discard*. Its meaning 'to not like' is more straightforwardly compositional, while *discard* 'to cast aside' is semantically more opaque and idiosyncratic.

One way of operationalizing morphological segmentability is through lexical frequency (Hay, 2001, 2003). Derivatives which occur frequently compared to their base are assumed to be less segmentable than derivatives whose base is more frequent than the derived form. If the derivative is more frequent than its base, it is more difficult for speakers to decompose the word into its constituents because they encounter the derivative more often than its individual parts. The derivative will be perceived as a more simplex word. If, on the other hand, the base is more frequent than the derivative, speakers more likely or quickly recognize the base as a constituent and therefore more likely perceive the derivative to be complex. The examples *dislike* and *discard* above have shown how morphological segmentability is closely related to the idea of semantic transparency. Hay (2001) demonstrated that unlike the absolute frequency of the derivative, the relative frequency of base and derived form is a good predictor for semantic transparency. Relative frequency has therefore been considered an important operationalization of segmentability.

However, semantic and frequential properties are not the only correlates of morphological segmentability. It has been proposed that semantic opacity, and hence morphological segmentability, correlates with phonological opacity. Hay (2001, 2003), in what is now known as the *segmentability hypothesis*, suggests that morphological segmentability plays a crucial role in phonetic reduction: More segmentable words are less easily reduced (also see M. J. Bell, Ben Hedia, & Plag, 2020; Ben Hedia, 2019; Ben Hedia & Plag, 2017; Bergmann, 2018). This hypothesis relies on the idea that in derivatives which are infrequent compared to their base, phonetic segments must be more fully realized to facilitate morpheme recognition for lexical access, since the derivative's meaning needs to be computed from its constituents. The rationale is that in words with strong boundaries, it might be important to fully produce individual segments to make it easier for the listener to recognize the constituents. In simplex words, in contrast, individual segments are less important for recognition. This idea relies on findings that non-morphemic phonemes are more likely to be deleted than morphemic ones (see Guy, 1980, 1991;

Labov, 1989; MacKenzie & Tamminga, 2021). Thus, segments in words which are perceived as more simplex (and therefore likely accessed as wholes) should be more vulnerable to reduction than segments in words which are more complex or decomposable, and hence produced as multi-morphemic. In short, the more meaningful a unit, the less easily can it be reduced. Note, however, that this premise is not always supported by the data (see our discussion in Section 5).

The segmentability hypothesis thus relies on assumptions that belong to two classes of explanations for reduction. The first of these classes is what Jaeger and Buz (2017) call a *communicative account*, i.e., a listener-oriented explanation for reduction. Here we can fit the idea that speakers want to make constituent recognition easy for the listener. The second of these classes is the *production ease account*, which is speaker-oriented and relies on processing and memory demands (cf. Jaeger & Buz, 2017). It is this approach that fits best with the idea that words are accessed in different ways, or different stages of difficulty, and that this affects acoustic reduction. Both assumptions of the segmentability hypothesis are, however, dependent on online reduction effects. This makes the segmentability hypothesis contrast with the third class of explanations, the *representational account*, which argues that reduction is encoded offline in linguistic representations (Jaeger & Buz, 2017). It is at present unclear which one of these classes of explanations comes closest to the truth, but several discussions suggest a complex interaction of the three dynamics (Arnold & Watson, 2015; Clopper & Turnbull, 2018; Jaeger & Buz, 2017).

Based on the rationale outlined above, we can formulate the hypothesis that the higher the morphological segmentability (i.e., the higher the relative frequency as calculated by dividing base frequency by derivative frequency), the less reduction is to be expected, i.e., the longer will be the duration of the word. Crucially, such protection against reduction is expected to occur in all domains, i.e., both in the affix and in the base, since both constituents need to be recognized in more complex words. Both the affix and the base (and consequently the whole word) are expected to vary in duration due to relative frequency. The following hypothesis follows from these considerations:

H_{seg1}: The higher the relative frequency of a derivative (i.e., the more segmentable a derivative is), the longer will be its affix duration and base duration.

2.1. Relative frequency effects on duration: Previous studies

While many studies have investigated the effects of other segmentability-related measures on various response variables, only a handful of studies have focused on the effect of relative frequency on duration specifically. Taken together, they produce a very incongruent picture.

Let us start with the studies that find, as predicted by H_{seg1}, a positive effect of relative frequency on duration (Hay, 2003, 2007; Plag & Ben Hedia, 2018; Zuraw, Lin, Yang, & Peperkamp, 2020). The flagship study on durational effects of segmentability is Hay (2003).

She finds relative frequency effects on the deletion of segments investigating the suffix *-ly* in English. In an experimental study, she had six undergraduate students read out *-ly*-affixed words containing a base-final /t/ (e.g., *swiftly*) embedded in carrier sentences. Relative frequencies were categorized into bins by arranging 20 words into four frequency-matched paradigms (consisting of five words each). She finds that base-final /t/ in highly segmentable words (like *abstractly*) is more likely to be fully realized than in low-segmentability words (like *exactly*). In a follow-up study, Hay (2007) finds a relative frequency effect on affix duration for the prefix *un-* in English. She conducted a corpus study on the ONZE corpus, analyzing 359 affixed and 310 monomorphemic forms containing the string *un-*. She finds that the *un-* string is more likely to be reduced in monomorphemic words than in more complex words. She also finds that *un-* is more reduced in words containing a 'legal' phonotactic transition from the affix-final nasal to the base-initial onset (i.e., a transition that also occurs in monomorphemic words), as opposed to those containing an 'illegal' transition (i.e., a transition that never occurs in monomorphemic words). Phonotactic transition probabilities can be seen as another correlate of morphological segmentability and will be controlled for in our study.

More recently, Plag and Ben Hedia (2018) find relative frequency effects for two of their four investigated affixes. In a study of the Switchboard corpus, relative frequency affects the affix duration of *un-* and *dis-* in the expected direction. The more segmentable a word derived with these prefixes, the longer the prefix becomes. And finally, Zuraw et al. (2020) find in a production experiment with 16 speakers of American English varieties that increasing word frequency and base frequency variables do reduce and promote aspiration, respectively. While they fail to find some effects for different ways of calculating relative frequency, they also observe that voice onset time becomes longer for infrequent words when their base frequency increases. In general, they interpret the results as overall providing support for the segmentability hypothesis.

Let us now move to the studies that find the opposite of what H_{seg1} predicts, i.e., a negative effect of relative frequency on duration (Pluymaekers et al., 2005b; Schuppler et al., 2012). Pluymaekers et al. (2005b) investigated frequency effects for four Dutch affixes in the Corpus of Spoken Dutch, also including relative frequency in their models. For *ge-*, they observe an effect counter to what Hay (2001, 2003) predicts. In their data, less segmentable words are associated with longer durations for *ge-*. The authors hypothesize that this could be because speakers are more likely to place stress on the first syllable in words they perceive as monomorphemic, given that Dutch monomorphemic words are usually stressed on the first syllable.

Similarly, Schuppler et al. (2012) find a relative frequency effect in the opposite direction (i.e., unexpected following Hay, 2001, 2003) on the presence or absence of the suffix *-t* in Dutch. In a study on the ECSD corpus, they analyze 2110 tokens ending in the Dutch suffix *-t* and find relative frequency to be significantly correlated with the likelihood of [t] presence. However, contrary to the segmentability hypothesis, [t] is more likely to be deleted in words which are

more segmentable. They hypothesize that this may have been due to differences in both setup and language. First, they measured reduction on the affix instead of on the base-final segment, like Hay did. Second, the uncertainty in choosing from the morphological paradigm was greater in their study because speakers had to decide between different suffixes. This supposedly gives the /t/ a greater information load and therefore a strengthened realization.

Finally, let us briefly review the studies that found null effects of relative frequency on duration (Ben Hedia & Plag, 2017; Plag & Ben Hedia, 2018; Pluymaekers et al., 2005b; Zimmerer et al., 2014; Zuraw et al., 2020). Apart from the one unexpected effect mentioned above, Pluymaekers et al. (2005b) largely fail to find relative frequency effects on duration. The affixes *ont-*, *ver-*, and *-lijk* all do not yield any significant effect of relative frequency. Likewise, Zimmerer et al. (2014) do not find a relative frequency effect on /t/-deletion in German. They constructed a new corpus from recordings of ten German native speakers who were instructed to produce paradigms for specific verbs given a set of pronouns. In this data set, relative frequency does not reach significance as a predictor for the deletion of morphemic /t/ in second- and third-person singular verbs. Further, Ben Hedia and Plag (2017) fail to find a relative frequency effect for the English prefixes *un-* and *in-*. In a study of the Switchboard corpus, they do find that double and singleton nasal durations at affix boundaries in words with *un-* are longer than in words with negative *in-*, which in turn are longer than in words with locative *in-*. This mirrors a descending hierarchy of segmentability. However, relative frequency as a gradient segmentability measure of individual words does not reach significance in any of their regression models. Similarly, Plag and Ben Hedia (2018) fail to observe relative frequency effects on affix duration for *in-* and *-ly*.

Finally, Zuraw et al. (2020) fail to find a categorical relative frequency effect on the aspiration of base-initial /t/ after the English prefixes *dis-* and *mis-*. They categorically coded for each word whether the word or its stem has a higher frequency. This variable did not yield a significant effect on /t/-aspiration. This may be related to their operationalization of relative frequency, as they binned the relative frequency distinction into two distinct categories instead of four as in Hay (2003), or instead of including it as a continuous variable. Moreover, an interaction of frequency (operationalized as the number of movies a word appears in) and the word's base frequency, which is another possible way to measure relative frequency, does not reach significance when predicting the likelihood of /t/ aspiration.

In sum, it is apparent that relative frequency does not always affect acoustic duration. The question remains why relative frequency produces such an incoherent picture, and to what extent it is a reliable measure as a morphological predictor of acoustic duration. This gap has been noted before, and there have been calls for research investigating when the phonetics of a derivative are influenced by its relative frequency and when they are not (Arndt-Lappe & Ernestus, 2020, p. 199). One idea is that the difference in the emergence of relative frequency effects between affixes might arise because the affixes differ in their prosodic structure. It has been speculated

(Plag & Ben Hedia, 2018) that prosodic word integration might inhibit segmentability effects. This will be discussed in the following section.

2.2. The prosodic structure of complex words

As they found relative frequency effects for *un-* and *dis-*, but not for *in-* and *-ly*, Plag and Ben Hedia (2018, p. 112) hypothesize that these mixed results may arise because of prosodic integration. The affixes *un-* and *dis-* are considered to form prosodic words and thus to not integrate into the prosodic word of their hosts, while prosodic word status is less clear for *in-* and *-ly*. The question arises if their different prosodic word status might provide an explanation for why the affix categories differ with regard to where segmentability effects can emerge.

In the prosodic hierarchy (**Figure 1**), the prosodic word (represented by lowercase omega ω) is considered to be a constituent above the foot level and below the phonological phrase level (Hall, 1999; Hildebrandt, 2015; Nespor & Vogel, 2007). There are several diagnostics that can be used to justify postulating such a constituent, such as violations of syllabic constraints (onset or coda conditions, violation of the law of initials, ambisyllabicity), stress, vowel changes, or semantic criteria (Hildebrandt, 2015; Raffelsiefen, 1999, 2007). The prosodic word is assumed to be the lowest constituent in the hierarchy that reflects morphological information, making it a key player in the transfer of morphological structure to phonetic output.

U	Phonological utterance
|	
IP	Intonation phrase
|	
Φ	Phonological phrase
|	
C	Clitic group
|	
ω	Prosodic word
|	
Σ	Foot
|	
σ	Syllable

Figure 1: The prosodic hierarchy, adapted from Hall (1999, p. 9) and Nespor and Vogel (2007, p. 11).

Within the hierarchy, researchers assume a number of constraints, one of which is especially important for the relationship between affix structure and prosodic word structure. Raffelsiefen (2007) suggests that the boundaries of grammatical categories must align with prosodic word boundaries, a constraint referred to as GP-ALIGNMENT. Crucially for our purposes, it follows from this constraint that whenever there are prosodic boundaries, they must coincide with morphological boundaries at the exact same place. This constraint, however, can be dominated by other constraints (such as, e.g., a constraint requiring syllables to have onsets). Because of this, if there are no prosodic boundaries, this does not necessarily mean that there are no morphological boundaries (Raffelsiefen, 2007, p. 212). In other words, a morpheme cannot include multiple prosodic words, but a prosodic word can include multiple morphemes. This means that in our study, there might be cases where an affix is a prosodic word on its own, as well as cases where the affix is just part of a larger prosodic word.

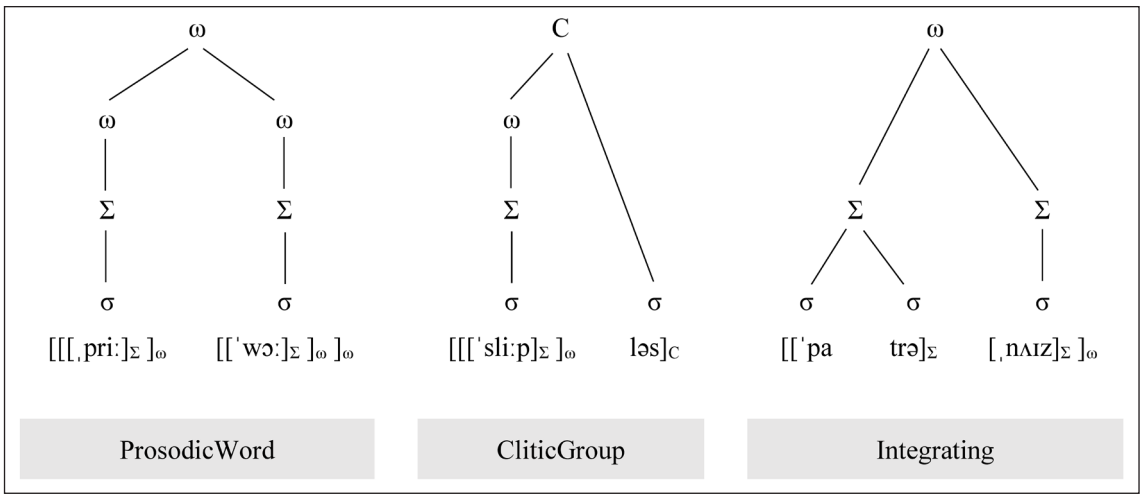

Figure 2: Three types of prosodic word integration for affixes (following Raffelsiefen, 1999), given as trees as well as in bracket notation. ProsodicWord: In this category, affixes form prosodic words on their own. CliticGroup: Here, affixes form a clitic group with their base. Integrating: In this case, affixes integrate into the prosodic word of their host base. Note that we use C to refer to the phonological domain of the clitic group, following the standard notation from the literature, whereas we use CliticGroup to refer to the category of derivatives in which such clitic group structure can be found.

From these constraints, Raffelsiefen (1999) deduces three possible scenarios that describe how an affix can integrate into prosodic word structure. The three structure trees with example words are given in **Figure 2**. First, an affix can form a prosodic word on its own (the ProsodicWord case). This is, for example, the case with the affix *pre-*. In the word *pre-war*, both syllables are stressed and heavy, hence both form a foot as well as a prosodic word each. Second, an affix might form a clitic group with the prosodic word of the stem (the CliticGroup case). An example

of a clitic group affix is *-less*. In the word *sleepless*, the suffix is unstressed, hence it cannot form a prosodic word. Yet, the suffix does not fully integrate into the host prosodic word either, as the base-final plosive is not resyllabified into the affix syllable, despite /pl/ being a possible onset in English. Third, there are affixes that fully integrate into the prosodic word of their host (the Integrating case). This is the case with *-ize*. In a word like *patronize*, the base-final coda nasal resyllabifies into the onset of the derivative-final syllable. The whole derivative is therefore considered to be one prosodic word. In sum, we can see that from left to right, there is an increasing degree of prosodic word integration. Standalone prosodic word affixes are least integrated, while integrating affixes are most integrated. The question remains which potential consequences these structural differences have for acoustic duration.

One of the most important and well-known effects of (prosodic) constituent structure on acoustic duration is *pre-boundary lengthening*. Before a structural boundary, such as a phrase boundary or word boundary, phonetic material is lengthened, i.e., increases in duration. This phenomenon has long since been described in many different studies (see, e.g., Beckman & Pierrehumbert, 1986; Campbell, 1990; Edwards & Beckman, 1988; Klatt, 1975; Vaissière, 1983; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992) and is now considered to be a robust and reliable predictor of duration. Specifically, it has also been demonstrated to occur in complex words with different prosodic structure (Auer, 2002; Bergmann, 2018; Sproat & Fujimura, 1993; Sugahara & Turk, 2009). In the literature, the amount of lengthening is said to be dependent on two factors: the boundary strength and the position of the boundary type in the prosodic hierarchy.

First, how much a string of segments is lengthened depends on the amount, or strength, of the boundaries. The more constituents begin or end between two segments, the stronger the boundary, and the stronger the pre-boundary lengthening. This means that boundaries add up to form stronger boundaries. For example, the internal foot boundary in *pre-war* is stronger than the internal foot boundary in *sleepless* (**Figure 2**), because *pre-war* $[[[_{\iota}\text{priː}]_{\Sigma}]_{\omega}[[\text{ˈwɔː}]_{\Sigma}]_{\omega}]_{\omega}$ features a double foot boundary (one foot ends after the affix and provides a closing boundary, a second foot begins and provides an opening boundary), whereas *sleepless* $[[[\text{ˈsliːp}]_{\Sigma}]_{\omega}\text{ləs}]_{C}$ just features a closing foot boundary after the base and no second foot. How strong a given boundary is, and consequently how much lengthening occurs before it, thus depends on the number of boundaries that co-occur together.

Second, the amount of lengthening depends on the position of the constituent providing the boundary in the prosodic hierarchy. Higher-level boundaries should be associated with stronger lengthening effects than lower-level boundaries. For example, segments before a prosodic word boundary are expected to be more lengthened than segments before a foot boundary, but less lengthened than segments before, e.g., a clitic group boundary. If the two boundaries in question

belong to the same level in the prosodic hierarchy (like a prosodic word contained in another prosodic word), we would still expect the higher-level boundary in the tree to have the stronger lengthening effect than the lower-level boundary contained in the higher-level constituent. Based on these two factors governing prosodically induced lengthening, we can formulate predictions about how duration should vary between our affixes affiliated with different prosodic word categories, as well as predictions about how prosodic word structure could interact with segmentability.

Coming back to the question of how the prosodic structure outlined above might interfere with morphological segmentability, we can hypothesize that morphological segmentability effects on duration might be counteracted by a strong prosodic boundary: According to the segmentability hypothesis, barely segmentable words should be more reduced than highly segmentable words, but if a barely segmentable word has a stronger internal prosodic word boundary, the preceding domain might be protected against reduction. In other words, pre-boundary lengthening effects introduced by prosody might cancel out reduction effects which low segmentability would have allowed for. We can complement our segmentability hypothesis (repeated here for convenience) with this idea as follows:

H_{seg1}: The higher the relative frequency of a derivative (i.e., the more segmentable a derivative is), the longer will be its affix duration and base duration.

H_{seg2}: The more prosodically integrated an affix is (the weaker the prosodic boundary), the less likely can higher relative frequency protect the derivative against reduction.

Second, prosodic word structure should have an effect on duration in general, independent of segmentability. It is important to think carefully about which domains should be more lengthened or less lengthened in specific prosodic affix categories since these are nested with prefix or suffix status. For example, many English prefixes are considered to form prosodic words, whereas English suffixes never form prosodic words of their own (Raffelsiefen, 1999). Since the phenomenon of pre-boundary lengthening assumes lengthening only for the domain *preceding* the boundary in question, which prosodic category we look at will determine which domain (base or affix) we expect to be affected, depending on whether the affix is a prefix or a suffix.[3] We can formulate the following predictions for the contrasts between the categories, depending

[3] One reviewer suggests that since affix position is such an important criterion in formulating the predictions for duration, one could also hypothesize about and test effects due to affix position directly (i.e., simply test for a durational influence of an affix's status as prefix or suffix). However, as explained above, the following hypotheses are not just built on affix position but on a variety of criteria relating to prosodic category (syllable and foot structure, stress, resyllabification). So, for example, *-ness* versus *-ation*, or *-ment* versus *-able* are predicted to behave differently, even though they are all suffixes. In English, affix position and prosodic category are not separable. The results show that we would lose information by just testing for affix position, as there are effects of prosodic category that cannot be explained by position alone.

on whether we compare the domains affix and base within a prosodic category, or the prosodic categories within a domain:

Comparing bases and affixes

H_{PW}: ProsodicWord bases should be more lengthened than ProsodicWord affixes. This is because ProsodicWord affixes are prefixes and therefore the base will be subject to word-final lengthening, with the second base-final pword boundary in a word like $[[[_{ˌ}pri:]_{\Sigma}]_{\omega}[[^{ˈ}wɔ:]_{\Sigma}]_{\omega}]_{\omega}$ ranking higher than the two internal pword boundaries.

H_{INT}: Integrating affixes should be more lengthened than Integrating bases. Integrating affixes are mostly suffixes, so the suffix will be subject most strongly to word-final lengthening, with the final pword boundary outranking the internal foot boundaries, as illustrated in $[[^{ˈ}patrə]_{\Sigma}[_{ˌ}nʌɪz]_{\Sigma}]_{\omega}$.

H_{CG}: CliticGroup affixes should be more lengthened than CliticGroup bases. CliticGroup affixes are generally suffixes and lack a foot boundary and prosodic word boundary compared to the bases, as seen in $[[[^{ˈ}sli:p]_{\Sigma}]_{\omega}ləs]_{C}$. However, the word-final C boundary ranks higher than the internal pword and foot boundaries, so word-final lengthening should be dominant.

Comparing prosodic categories

H_{pros1}: ProsodicWord bases should be more lengthened than Integrating bases. This is because ProsodicWord bases end with one foot boundary and two pword boundaries and occur word-finally ($[[[_{ˌ}pri:]_{\Sigma}]_{\omega}[[^{ˈ}wɔ:]_{\Sigma}]_{\omega}]_{\omega}$), whereas Integrating bases end with just two foot boundaries and mostly occur word-internally ($[[^{ˈ}patrə]_{\Sigma}[_{ˌ}nʌɪz]_{\Sigma}]_{\omega}$).

H_{pros2}: Integrating affixes (generally suffixes) should be more lengthened than ProsodicWord affixes (prefixes) due to word-final lengthening. The ProsodicWord prefix has two more boundaries provided by the derivative alone (compare $[[[_{ˌ}pri:]_{\Sigma}]_{\omega}[[^{ˈ}wɔ:]_{\Sigma}]_{\omega}]_{\omega}$ and $[[^{ˈ}patrə]_{\Sigma}[_{ˌ}nʌɪz]_{\Sigma}]_{\omega}$), but since the Integrating affix occurs word-finally, it will be followed by at least the same amount of potentially stronger boundaries belonging to whatever word or phrase comes next.

H_{pros3}: CliticGroup affixes should be more lengthened than Integrating affixes. This is because the word-final C boundary of CliticGroup affixes, as seen in $[[[^{ˈ}sli:p]_{\Sigma}]_{\omega}ləs]_{C}$, outranks the pword and foot boundaries of Integrating affixes, as seen in $[[^{ˈ}patrə]_{\Sigma}[_{ˌ}nʌɪz]_{\Sigma}]_{\omega}$.

H_{pros4}: CliticGroup bases should be more lengthened than Integrating bases. The CliticGroup internal morphological boundary coincides with just one single foot boundary in words like $[[[^{ˈ}sli:p]_{\Sigma}]_{\omega}ləs]_{C}$ instead of a double foot boundary like for the Integrating bases in words like $[[^{ˈ}patrə]_{\Sigma}[_{ˌ}nʌɪz]_{\Sigma}]_{\omega}$; however, the CliticGroup bases feature a higher-ranking internal pword boundary, which the Integrating bases do not.

H_{pros5}: CliticGroup affixes (generally suffixes) should be more lengthened than ProsodicWord affixes (prefixes). CliticGroup affixes lack foot boundaries and pword boundaries compared to the ProsodicWord affixes (compare $[[[ˈsliːp]_{Σ}]_{ω}ləs]_{C}$ and $[[[ˌpriː]_{Σ}]_{ω}[[ˈwɔː]_{Σ}]_{ω}]_{ω}$), but CliticGroup affixes are mostly suffixes and are therefore subject to the higher-ranking word-final C boundary lengthening compared to the ProsodicWord prefixes.

H_{pros6}: ProsodicWord bases should be more lengthened than CliticGroup bases due to word-final lengthening and because they feature an additional pword boundary compared to the CliticGroup bases (compare $[[[ˌpriː]_{Σ}]_{ω}[[ˈwɔː]_{Σ}]_{ω}]_{ω}$ and $[[[ˈsliːp]_{Σ}]_{ω}ləs]_{C}$).

We can summarize these predictions as follows (**Table 1**). We use the > sign to mean 'should be more lengthened than':

Comparing bases and affixes		
Integrating	CliticGroup	ProsodicWord
affix > base	affix > base	base > affix
Comparing prosodic categories		
bases	affixes	
ProsodicWord > CliticGroup > Integrating	CliticGroup > Integrating > ProsodicWord	

Table 1: Summary of predictions for prosodic category effects on duration. The right arrow sign between categories means that the left-hand element is expected to be more lengthened than the right-hand element.

It has become clear that an empirical investigation is called for that investigates segmentability and prosodic word effects on duration simultaneously. It is further necessary to conduct a study which examines more affixes at the same time, so we can better compare when effects do and do not emerge in the same conditions. We need to assess whether relative frequency is a useful predictor for duration, and under which prosodic circumstances it is not. The aim of the present study is to account for this gap by testing the hypotheses outlined above, as well as by aiming for a larger-scale investigation than previous studies.

3. Methodology

We used three different corpora from two varieties of English, each corpus study including eight affixes each, and measuring both affix durations and base durations as responses. There are at least three advantages to such a wide corpus approach. First, we are able to analyze a lot of data, without having to carry out large-scale experiments in different places. Second, the type of data is conversational speech produced outside of a lab context. Note that 'conversational speech' does

not necessarily exclude lab speech—conversational speech can also be elicited, and many alleged drawbacks of experimental speech can be alleviated with careful design of experimental setup and stimuli (Xu, 2010). However, there is hardly a way to come closer to a reflection of everyday use of language than by using spontaneous speech outside of the experimental context, without any input by the researcher. It has been argued (e.g., B. V. Tucker & Ernestus, 2016) that research on speech production in particular needs to shift its focus to spontaneous speech to be able to draw valid conclusions about language processing. And third, much psycholinguistic research has been conducted on experimental data that has often not been elicited to guarantee spontaneousness or casualness. However, it is generally important to use a diversity of methods in linguistic research: Spontaneous and non-spontaneous speech both come with their advantages and disadvantages and might yield different results in the investigation of linguistic phenomena. This paper thus contributes to the diversification of psycholinguistic research by using a corpus approach.

The affixes investigated in the present study are listed in **Table 2**, together with their corpus study affiliation and their assumed prosodic category. The prosodic classification is based on Raffelsiefen (1999, 2007) and, whenever she does not explicitly mention the affix in question, on our application of her criteria for determining prosodic word status (i.e., syllable and foot structure, stress, resyllabification). Note that the selected affix categories are the same across the QuakeBox and ONZE corpora, but partly different for the Audio BNC. This is because after initial sampling from the Audio BNC, some of the affix categories did not yield enough tokens in the New Zealand corpora and were therefore replaced by others.

	ProsodicWord	**CliticGroup**	**Integrating**
Audio BNC	*dis-*, *in-*, *pre-*, *un-*	*-ness*, *-less*	*-ation*, *-ize*
QuakeBox corpus	*dis-*, *un-*, *re-*	*-ness*, *-ment*	*-ation*, *-able*, *-ity*
ONZE corpus	*dis-*, *un-*, *re-*	*-ness*, *-ment*	*-ation*, *-able*, *-ity*

Table 2: Investigated affixes in the three corpus studies, sorted by prosodic word category.

Let us briefly have a look at some of the characteristics of these affixes (following Bauer, Lieber, & Plag, 2013; Plag, 2018). The ProsodicWord affixes are *dis-*, *in-*, *pre-*, *un-,* and *re-*. These affixes have in common that they are all prefixes and are characterized by relatively clear semantics: *dis-*, *in-,* and *un-* have a clear negative meaning, while *re-* is clearly iterative and *pre-* can carry spatial or temporal meaning. They mostly take free morphemes as bases and their derivatives tend to be rather transparent. They vary in their productivity, with *pre-* and *un-* being considered very productive, and *dis-*, *in-* and *re-* somewhat productive. They are generally assumed to be secondarily stressed, although there is some variability, with unstressed or main-stressed attestations. Most of the prefixes are not obligatorily subject to phonological alternation

except *in-*, which assimilates to the following base-initial onset (*irregular*, *illegal*, *impossible*). None of the affixes cause resyllabification. It has to be noted that the prefixes *dis-* and *in-* are considered to be prosodic words only most of the time, i.e., in semantically transparent derivatives with words as bases. In opaque derivatives with bound roots, they are expected to integrate into the host prosodic word. However, this is irrelevant for our purposes, as we excluded all derivatives whose base is not a word in order to be able to properly calculate the base frequency, which is needed for relative frequency (see dataset description in Section 3.1). All our *dis-* and *in-*affixed words are therefore associated with the ProsodicWord structure.

The CliticGroup affixes are the suffixes *-ness*, *-less*, and *-ment*. Their semantics are generally a little less straightforward than those of the ProsodicWord affixes. While *-less* has a clear privative meaning, the meanings of *-ness*, which can denote abstract states, traits, or properties, and *-ment*, which can denote processes, actions, or results, are not always predictable. Like the ProsodicWord affixes, however, the CliticGroup affixes mostly take free bases and produce mostly transparent derivatives. Both *-ness* and *-less* are considered to be highly productive, whereas *-ment* used to be productive but has lost some of that capability in contemporary English varieties. The CliticGroup affixes are never stressed, not subject to phonological alternations, and not involved in resyllabification processes.

Lastly, the Integrating affixes are the suffixes *-ation*, *-ize*, *-able* and *-ity*. Compared to the other affixes, their semantics are rather multifaceted, covering a wide range of meanings each. For example, *-ize* can have locative, ornative, causative, resultative, inchoative, performative or similative meaning, *-ation* denotes events, states, locations, products, or means, *-able* is used to express capability, liability, or quality, and *-ity* denotes properties, states, qualities, sometimes with a 'nuance of pomposity' compared to *-ness* (Bauer et al., 2013, p. 248), and has various idiosyncratic derivatives, too. Their bases are mostly free morphemes, but all of them can attach to bound roots as well. They are mostly transparent, but overall less so compared to the affixes in the other two prosodic categories. The most productive of them by far is *-able*, whereas *-ity* is more restricted in its application. Integrating affixes can cause differences in the stress patterns of the derivatives, either being mostly secondarily (*-ize*) or primarily (*-ation*) stressed themselves, or capable of causing stress shifts and other phonological alternations within their bases (*-able, -ation,* and *-ity*). Resyllabification is commonplace among all of them, making them clear cases of the Integrating category. Having categorized the affixes in this way, let us now turn to the data we used to represent these categories.

3.1. Corpora and datasets

The three corpora we used are the Audio BNC (Coleman, Baghai-Ravary, Pybus, & Grau, 2012), the QuakeBox corpus (Walsh et al., 2013), and the ONZE corpus (Gordon, Maclagan, & Hay, 2007). Finding enough tokens for each affix category in spoken corpora is often a problem, but these three corpora are large enough to yield a sufficient number of observations per category.

The Audio BNC (Coleman et al., 2012) is the largest of the three corpora. It consists of both monologues and dialogues from different speech genres of a number of British English varieties, and contains about 7.5 million words. We extracted the data via its web interface (Hoffmann & Arndt-Lappe, 2021; Hoffmann & Evert, 2018). The QuakeBox corpus (Walsh et al., 2013) consists of mainly monologues spoken by inhabitants of Christchurch, New Zealand, who tell the interviewer about their experiences surrounding the 2010–2011 Canterbury earthquakes. The data was extracted via the LaBB-CAT interface (Fromont, 2003–2020; Fromont & Hay, 2012). At the time of extraction, the corpus contained about 800,000 tokens, or 86 hours of speech spoken by 758 participants. Lastly, the ONZE corpus (Gordon et al., 2007) consists of three collections of recordings from different New Zealand English varieties, the historical Mobile Unit (with speakers born between 1851–1910), the later Intermediate Archive (with speakers born between 1890–1930), and the contemporary Canterbury Corpus (with speakers born between 1930–1984). As with the QuakeBox corpus, LaBB-CAT was used for data extraction. At the time of extraction, all subcorpora contained about 3.3 million tokens, or 392 hours of speech spoken by 1,589 participants.

Wordlists, recordings, and textgrids were obtained by entering query strings into the corpora. These query strings searched for all word tokens that begin (for prefixes) or end (for suffixes) in the orthographic and phonological representation of each of the investigated affix categories. The wordlists were cleaned manually, excluding words which were monomorphemic (e.g., *bless, pregnant, station*), whose semantics or base were unclear (e.g., *harness*, *predator*, *dissertation*), or which were proper names or titles (e.g., *Guinness*, *Stenness*, *Stromness*). We only included derivatives with words as bases, not bound roots, meaning that the bases had to be attested as independent words with a related sense to the derivative. This is important because we need the frequency of the base word outside of the derivative in order to calculate relative frequency.[4] The existence of such bases was determined by consulting pertinent dictionaries such as the Oxford English Dictionary Online (2020), as well as web attestations. Following Pluymaekers et al. (2005b), we included all word forms of a given type (e.g., *discover*, *discovered, discovering* etc.).

The three corpora come phonetically aligned by an automatic forced aligner. We extracted the recordings and textgrids twice, (1) including just the affixed word, and (2) including the affixed word plus one additional second of speech before and one additional second after it. The larger interval was used to calculate the covariate SPEECH RATE (see Section 3.2.2.) and will be referred to in the following as *utterance*. In addition to the audio and text files, we extracted a spreadsheet of the tokens including corresponding meta-information and variables which had already been coded in the corpora.

[4] Including bound roots is also possible using workarounds such as categorically assigning a low frequency (of, e.g., 1) to them, or counting their occurrence in other derivatives. The empirical and theoretical consequences of such workarounds are not clear.

Before starting the acoustic analysis, manual inspection of all items was necessary to exclude items that were not suitable for further analysis. This was done by visually and acoustically inspecting the items in the speech analysis software Praat (Boersma & Weenik, 2001). Items were excluded that fulfilled one or more of the following criteria: The textgrid was a duplicate or corrupted for technical reasons, the target word was not spoken or inaudible due to background noise, the target word was interrupted by other acoustic material, laughing, or pauses, the target word was sung instead of spoken, the target word was not properly segmented or incorrectly aligned to the recording. In cases where the alignment did not seem satisfactory, we specifically examined three boundaries in order to decide whether to exclude the item: the word-initial boundary, the word-final boundary, and the boundary between base and affix. We considered an observation to be correctly aligned if none of these boundaries would have to be shifted to the left or right under application of the segmentation criteria in the pertinent phonetic literature (cf. Ladefoged & Johnson, 2011; Machač & Skarnitzl, 2009). Following Machač and Skarnitzl (2009, pp. 25–26), we considered the shape of the sound wave to be the most important cue, followed by the spectrogram, followed by listening. We also excluded derivatives with geminates (e.g., *openness*, *unnecessary*), since in these cases it is usually impossible to determine the acoustic boundary between base and affix (also see Ben Hedia, 2019).

We further reduced the datasets to only those word types and speakers for which there are more than three observations (see note on random effects for WORD and SPEAKER in Section 3.3.). An overview of the resulting datasets is given in **Table 3**.

	Audio BNC		**QuakeBox**		**ONZE**	
	Tokens	**Types**	**Tokens**	**Types**	**Tokens**	**Types**
-ness	300	34	106	13	62	7
-less	144	21				
pre-	40	10				
-ize	365	27				
-ation	3370	193	375	36	772	76
dis-	503	61	112	19	159	25
un-	594	73	203	26	224	23
in-	248	26				
-able			150	19	207	19
-ity			532	21	323	24
-ment			287	24	541	35
re-			279	24	250	36

Table 3: Overview of types and tokens sorted by corpus and affix category.

3.2. Variables

Acoustic duration may be affected by a number of variables other than relative frequency and prosodic boundaries, and it is vital to control for as much variation as possible. In order to be sure to remove potential durational differences that arise simply because of the number and type of segments in a given word, we not only introduce important covariates, but also modify the response variable by calculating the duration difference to the expected item duration instead of the absolute duration. The following sections outline all of our variables in detail.

3.2.1. Response variable

DURATION DIFFERENCE

For each word token in our data set we calculated as dependent variable two durational measurements, one for the affix, one for the base. This gives us two observations per word token. Each measurement was paired with the value `affix` or `base` in an additional variable (called TYPE OF MORPHEME) that coded whether the measurement was for the affix or for the base. This coding had the advantage that one can look at base durations and affix durations in a single statistical model.

One important problem in analyzing spontaneous speech is that which words are spoken is uncontrolled for phonological and segmental makeup. This problem is particularly pertinent for the present study, as our datasets feature different affixes whose derivatives vary in word length. To mitigate potential durational differences that arise simply because of the number and type of segments in each word, we refrained from using absolute observed duration as our response variable. Instead, we derived our duration measurement in the following way.

First, we measured the absolute acoustic duration of the word in milliseconds from the textgrid files with the help of scripts written in Python. Second, we calculated the mean duration of each segment in a large corpus (Walsh et al., 2013) and computed for each word the sum of the mean durations of its segments.[5] This sum of the mean segment durations is also known as 'baseline duration,' a measure which has been successfully used as a covariate in other corpus-based studies (e.g., Caselli et al., 2016; Engemann & Plag, 2021; Gahl, Yao, & Johnson, 2012; Sóskuthy & Hay, 2017). It would now be possible to subtract this baseline duration from the observed duration, giving us a new variable that represents only the difference in duration to what is expected based on segmental makeup. However, we found that this difference is not constant across longer and shorter words. Instead, the longer the word is on average, the smaller the difference between the baseline duration and the observed duration. In a third and final step, we therefore fitted a simple linear regression model predicting observed duration as a function of baseline duration. The residuals of this model represent our response variable. Using this method, we factor in the non-constant relationship between baseline duration and observed duration. We named this response variable DURATION DIFFERENCE, as it encodes the difference between the observed duration and a duration that is expected on the basis of the segmental makeup.

[5] We used the QuakeBox corpus for this task because it provides this information in an accessible and reliable form.

3.2.2. Predictor variables

RELATIVE FREQUENCY

We extracted the WORD FREQUENCY, i.e., the frequency of the derivative, and the BASE FREQUENCY, i.e., the frequency of the derivative's base, from the *Corpus of Contemporary American English* (COCA, Davies, 2008), with the help of the corpus tool Coquery (Kunter, 2016).

Derived words are often rare words (see, e.g., Plag, Dalton-Puffer, & Baayen, 1999). For this reason, very large corpora are necessary to obtain frequency values for derived words. We chose COCA because this corpus is much larger than the BNC or the New Zealand corpora themselves, and therefore had a much higher chance of the words and their bases being sufficiently attested. We prioritized covering a bigger frequency range with more tokens. Moreover, we consider it important to use the same frequencies consistently across the three studies. As corpora in different varieties have different sizes, it is difficult to compare effects of frequency for sets of words of which many have very low frequencies, or do not occur at all in one or two of the three corpora. Note that while COCA covers a different group of English varieties than the three audio corpora, these three corpora already differ in the varieties they cover. We therefore cannot avoid using frequency values from varieties which are not covered in at least one of the three corpora we analyze.

We calculated RELATIVE FREQUENCY by dividing BASE FREQUENCY by WORD FREQUENCY. Calculated this way, the higher the relative frequency, the more segmentable the item is assumed to be. Following standard procedures, we log-transformed RELATIVE FREQUENCY before it entered the models to improve its skewed distribution. We added a constant of $+1$ to the variables in order to be able to take the log of the zero frequency of non-attested derivatives and bases (cf. Baayen, 2008).

PROSODIC CATEGORY

This predictor variable was categorically coded for each affix by assigning the value `PW` (ProsodicWord), `CG` (CliticGroup), or `INT` (Integrating) to each observation containing the pertinent affix (see Section 2.2. and Raffelsiefen, 1999, 2007 for the classification criteria).

TYPE OF MORPHEME

As described in Section 3.2.1., there are two observations for each word token, one with the pertinent measurements of the affix, the other with the pertinent measurements of the base. The resulting variable, type of morpheme, can thus take two values, `base` or `affix`.

SPEECH RATE

How fast we speak naturally influences the duration any produced linguistic unit will have. SPEECH RATE can be operationalized as the number of syllables a speaker produces in a given time interval (see, e.g., Plag et al., 2017; Pluymaekers et al., 2005b). We divided the number of syllables in the utterance by the duration of the utterance. As explained in Section 3.1., *utterance* was defined as the window containing the target word plus one second before and one second after it. We considered this window to be a good compromise between a maximally local

speech rate which just includes the adjacent segments, but allows the target item to have much influence, and a maximally global speech rate, which includes larger stretches of speech but is vulnerable to changing speech rates during this larger window. The utterance duration and the number of syllables in the utterance window were extracted from the textgrids with a Python script. Pauses (i.e., periods of silence where no syllables were annotated) were not factored into the total duration of the utterance, nor into the total syllable count. We expect that the higher the speech rate (i.e., the more syllables are produced within the utterance window), the shorter the duration of base and affix should become.

NUMBER OF SYLLABLES

Researchers have observed a compression effect where segments are reduced if they are followed by more syllables (Lindblom, 1963; Nooteboom, 1972). If we modeled absolute durations, we would expect a higher number of syllables to lead to longer durations, because more syllables mean more phonetic material. However, we already control for the number of segments and thereby indirectly for the number of syllables in our response variable DURATION DIFFERENCE. We therefore expect NUMBER OF SYLLABLES to be negatively correlated with duration due to the compression effect. The NUMBER OF SYLLABLES in a given derivative was extracted with Python from the textgrids and converted into a categorical variable.

BIGRAM FREQUENCY

BIGRAM FREQUENCY refers to the frequency of the target derivative occurring together with the word following it in COCA (Davies, 2008). It has been found that the degree of acoustic reduction can be influenced by the predictability conditioned on the following context (see, e.g., A. Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pluymaekers et al., 2005a; Torreira & Ernestus, 2009). We thus expect that the higher the bigram frequency, the shorter the duration. Following standard procedures, we log-transformed BIGRAM FREQUENCY before it entered the models to improve its skewed distribution. We added a constant of $+1$ to the variables in order to be able to take the log of non-attested bigrams (cf. Baayen, 2008).

BIPHONE PROBABILITY

The variable BIPHONE PROBABILITY refers to the sum of all biphone probabilities (the likelihood of two phonemes occurring together in English) in a given target derivative. It has been found that segments are more likely to be reduced or deleted when they are highly probable given their context (see, e.g., Edwards, Beckman, & Munson, 2004; Munson, 2001; Turnbull, 2018; also see Hay, 2007 on effects of the legality of the phoneme transition on reduction). There are two possible lines of reasoning behind this. First, the more probable a biphone is, the easier it should be to access it, since with higher probability speakers have a stronger representation of phoneme templates occurring in succession. Note that this idea of reduction based on lexical

access speed is debated in the literature (see, e.g., Arnold & Watson, 2015; Clopper & Turnbull, 2018).[6] Second, the more probable a biphone is, the better our phonotactic motor skills of pronouncing the biphone will be, as our articulator muscles are trained better in pronouncing segment sequences that we pronounce often. Taken together, we expect biphone probability to be negatively correlated with duration: the more probable the biphones, the shorter the durations.

Biphone probabilities were calculated by the Phonotactic Probability Calculator (Vitevitch & Luce, 2004). For this, we first manually translated the target derivatives' ASCII transcriptions of the Audio BNC, as well as the QuakeBox and ONZE transcription systems into the coding referred to as Klattese, as this is the computer-readable transcription convention required by this calculator. We then summed the biphone probabilities for each word and divided the result by the number of this word's segments to obtain the average biphone probability for each word.

3.3. Modeling procedure

When modeling acoustic duration, it is important to control for any potential durational variation that arises simply from word type idiosyncrasies. We therefore decided to fit mixed-effects regression models to the data, using R (R Core Team, 2020), the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015), and `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2016). Mixed-effects regression can deal with unbalanced datasets and is particularly suited to investigate variables of interest while controlling for other potentially relevant variables (Baayen, 2008). We included a random intercept for WORD, i.e., our target derivative type, for SPEAKER, and for AFFIX.

In order to be able to include a random intercept for WORD and SPEAKER, we included only word types and speakers for which we have at least three observations. For WORD, this reduced the data by 953 tokens for the Audio BNC, 377 tokens for the QuakeBox corpus, and 563 tokens for the ONZE corpus (leaving 6299, 2270, and 3111 tokens, respectively). For SPEAKER, we lost 735 tokens for the Audio BNC, 226 tokens for the QuakeBox, and 573 tokens for the ONZE (leaving 5564, 2044, and 2538 tokens, respectively). For AFFIX random intercepts, the token counts remain unchanged. For the final counts, refer again to **Table 3**.

In the course of fitting the regression models, we trimmed the dataset by removing observations from the models whose residuals were more than 2.5 or 2.0 standard deviations away from the mean, which led to a satisfactory distribution of the residuals (cf., e.g., Baayen & Milin, 2010). This resulted in a loss of 114 tokens (2 % of the data) for the Audio BNC, 85 observations (4.2 % of the data) for the QuakeBox corpus, and 112 observations (4.4 % of the data) for the ONZE corpus.

[6] It is also possible to have a competing hypothesis about the direction of such probability effects. For example, we can read high probability as high *certainty*: The more probable a phonological string is, the more certain the speaker is in their articulation. When the speaker is uncertain, they want to invest less energy in maintaining the acoustic signal, which should lead to shorter durations of low-probable strings (not of high-probable ones). Some studies have found such effects (e.g., Stein & Plag, 2021; Tomaschek, Plag, Ernestus, & Baayen, 2019; B. Tucker, Sims, & Baayen, 2019; but see Watson, Buxó-Lugo, & Simmons, 2015).

We used variance inflation factors to test for possible multicollinearity of the variables. Collinearity diagnostics are naturally inflated for models with interaction terms, which is why we created separate models with all our variables, but without the interactions, in order to check for collinearity problems. All VIFs were smaller than 2, i.e., far below the critical value of 10 (Chatterjee & Hadi, 2006).

Models were fitted with interactions between the variables of interest. The models included interactions between RELATIVE FREQUENCY and PROSODIC CATEGORY, PROSODIC CATEGORY and TYPE OF MORPHEME, and RELATIVE FREQUENCY and TYPE OF MORPHEME. The initial models also included all covariates. The models were then simplified according to the standard procedure of removing non-significant terms in a stepwise fashion. An interaction term or a covariate was eligible for elimination when it was non-significant at the .05 alpha level. Non-significant terms with the highest *p*-value were eliminated first, followed by terms with the next-highest *p*-value. This was repeated until only variables remained in the models of which at least one level reached significance at the .05 alpha level. To investigate the differences between different interaction levels we relevelled the dataset for each contrast, i.e., for each reference level constellation of base and affix, and of prosodic categories.

4. Results

We report the *p*-values for the analysis of variance of the fixed effects in our three final models in **Table 4**. These *p*-values were calculated with the `Anova()` function (Type III) from the `car` package (Fox & Weisberg, 2011). We document the full models for one respective reference level each in **Table 6**. The interested reader can access the models in the scripts at https://osf.io/xuh42/.

Before we discuss the effects which turned out significant and remained in the models, let us first examine the null results. We can see from **Table 4** that the interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY was not significant in any of the corpora and was therefore removed from the models. The lack of interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY means that we already must dismiss H_{seg2} at this point. We do not find evidence that the effect of relative frequency depends on prosodic integration or vice versa.

One problem within the null hypothesis significance testing (NHST) framework is that it is strictly speaking not possible to interpret non-significance as the non-existence of an effect. While we do not find support for an interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY and thus no support for H_{seg2}, we cannot claim that the opposite is true, i.e., we cannot 'confirm' the null hypothesis. To be able to claim with more confidence that prosodic word integration does not affect whether relative frequency protects against reduction, it can be useful to quantify the evidence for the null. One way to do this is by using the BIC approximation

	Audio BNC				QuakeBox				ONZE			
	Chi²	**DF**	**Pr**		**Chi²**	**DF**	**Pr**		**Chi²**	**DF**	**Pr**	
Intercept	12.7	1	0.000	***	18.8	1	0.000	***	4.0	1	0.045	*
RELATIVE FREQUENCY	0.3	1	0.564									
TYPE OF MORPHEME	236.7	1	0.000	***	832.4	1	0.000	***	434.9	1	0.000	***
PROSODIC CATEGORY	11.0	2	0.004	**	51.4	2	0.000	***	12.9	2	0.002	**
SPEECH RATE	2467.2	1	0.000	***	742.4	1	0.000	***	929.3	1	0.000	***
BIGRAM FREQUENCY	6.1	1	0.014	*					4.7	1	0.031	*
BIPHONE PROBABILITY	23.2	1	0.000	***					9.2	1	0.002	**
NUMBER OF SYLLABLES	56.5	4	0.000	***	15.4	4	0.004	**	6.6	4	0.160	[7]
RELATIVE FREQUENCY: PROSODIC CATEGORY												
RELATIVE FREQUENCY: TYPE OF MORPHEME	19.4	1	0.000	***								
TYPE OF MORPHEME: PROSODIC CATEGORY	409.3	2	0.000	***	1495.1	2	0.000	***	862.6	2	0.000	***

Table 4: ANOVA *p*-values of fixed effects fitted to DURATION DIFFERENCE in the three corpora (Type III Wald chi-square tests). For empty cells, the predictors were non-significant and thus removed from the model during the fitting procedure.

[7] This covariate is non-significant in the ANOVA summary, but was not eliminated because one contrast between the levels of this covariate is significant in the full model (documented in Table 5).

to the Bayes Factor (Wagenmakers, 2007).[8] With this method, we can approximate the Bayes Factor by using the difference between BIC values of a full model for H_{seg2} (including the interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY) and a model for its null hypothesis (i.e., a model that does not include this interaction). If we assume that it is a priori equally plausible that RELATIVE FREQUENCY and PROSODIC CATEGORY interact and that they do not interact, we can estimate the models' posterior probabilities with the help of the Raftery (1995) classification scheme. **Table 5** compares the BIC and Bayes Factor estimates for the three corpora.

	BIC (H_{seg2} model)	BIC (H_0 model)	BIC difference	BF(H_{seg2})	BF(H_0)
Audio BNC	–29,087.43	–29,123.82	36.39	1.25e–08	79,795,562
QuakeBox	–9,923.55	–9,954.07	30.52	2.36e–07	4,236,810
ONZE	–13,146.06	–13,179.25	33.19	6.20e–08	16,120,645

Table 5: BIC approximation to the Bayes Factor (BF) comparing the models for H_{seg2} (a full model including the interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY) and H_0 (the same model but without this interaction) for the three corpora.

We can see that in all three cases, the null hypothesis model (i.e., the one that assumes prosodic word integration to not have an influence on relative frequency effects) provides the better fit because of its lower BIC, and that it has the higher Bayes Factor value. The Bayes Factor, roughly speaking, tells us how many times 'more' we should believe in the respective hypothesis. According to Raftery (1995), if we start with the belief that the hypothesis and the null hypothesis are equally likely, a Bayes Factor of >150 constitutes 'very strong' evidence for the respective hypothesis (i.e., a posterior probability of $>.99$). We can thus say that for all three corpora, we find very strong evidence for the null. Prosodic word integration does not play any role in whether higher relative frequency can protect against reduction.

In addition, **Table 4** shows that relative frequency was generally removed from all models both as interaction and as main effect due to non-significance, except from the Audio BNC. In the Audio BNC, we find one significant interaction between RELATIVE FREQUENCY and TYPE OF MORPHEME, where RELATIVE FREQUENCY is significant only if we look at the affix. In all three cases (despite the 'significant' effect in the Audio BNC), though, the BIC approximation to the Bayes Factor for the models with versus without RELATIVE FREQUENCY again tells us that these models constitute very strong evidence for the null (BF $>$ 8000, posterior probability $>.99$ for all three corpora). H_{seg1} is thus not well supported either: RELATIVE FREQUENCY generally does not affect duration in our data.

Before turning to the significant variables of interest, let us briefly look at the covariates. As seen in **Table 6**, the covariates generally behave as expected in the three models: SPEECH RATE

[8] We thank an anonymous reviewer for this useful suggestion.

	AudioBNC			QuakeBox			ONZE		
	Estimate	***SE***		**Estimate**	***SE***		**Estimate**	***SE***	
Intercept	–0.0398	0.0112	*	–0.0531	0.0122	***	–0.0247	0.0123	
RELATIVE FREQUENCY	–0.0010	0.0017							
TYPE OF MORPHEME `affix`	0.0223	0.0014	***	0.0804	0.0028	***	0.0477	0.0023	***
PROSODIC CATEGORY `CG`	0.0196	0.0153		0.0204	0.0160		0.0050	0.0172	
PROSODIC CATEGORY `PW`	0.0421	0.0129	*	0.0943	0.0138	***	0.0488	0.0148	*
SPEECH RATE	–0.0338	0.0007	***	–0.0336	0.0012	***	–0.0299	0.0010	***
BIGRAM FREQUENCY	–0.0017	0.0007	*				–0.0021	0.0010	*
BIPHONE PROBABILITY	–0.0063	0.0013	***				–0.0085	0.0028	**
NUMBER OF SYLLABLES `3`	0.0183	0.0040	***	0.0057	0.0073		0.0022	0.0063	
NUMBER OF SYLLABLES `4`	0.0223	0.0050	***	0.0167	0.0085		0.0090	0.0073	
NUMBER OF SYLLABLES `5`	0.0362	0.0054	***	0.0398	0.0109	***	0.0200	0.0091	*
NUMBER OF SYLLABLES `6`	0.0593	0.0112	***	0.0024	0.0216		0.0237	0.0181	
RELATIVE FREQUENCY: TYPE OF MORPHEME `affix`	0.0056	0.0013	***						
TYPE OF MORPHEME `affix`: PROSODIC CATEGORY `CG`	–0.0374	0.0047	***	–0.0929	0.0053	***	–0.0796	0.0040	***
TYPE OF MORPHEME `affix`: PROSODIC CATEGORY `PW`	–0.0571	0.0028	***	–0.1790	0.0047	***	–0.1092	0.0041	***
N (two observations per token)	10901			3918			4852		
N SPEAKER	595			264			312		
N WORD	441			178			242		
N AFFIX	8			8			8		
R^2 fixed	0.2274			0.3523			0.2932		
R^2 total	0.3870			0.4821			0.4836		

Table 6: Final models fitted to DURATION DIFFERENCE for all three corpora.

is always very highly significant and negatively correlated with duration. BIGRAM FREQUENCY is significant in the Audio BNC and ONZE and negatively correlated with duration. BIPHONE PROBABILITY is significant in the Audio BNC and ONZE models and also negatively correlated with duration. Finally, NUMBER OF SYLLABLES is positively correlated with duration in the three models.

Let us now examine the variables of interest that remained in the models, starting with RELATIVE FREQUENCY. RELATIVE FREQUENCY only showed one significant interaction with TYPE OF MORPHEME, namely in the Audio BNC. The interaction is illustrated in **Figure 3**. The x-axis represents the log of RELATIVE FREQUENCY; the y-axis represents the response variable DURATION DIFFERENCE. We use a blue, solid line to represent the slope of relative frequency when looking at the base, and an orange, dotted line to represent the slope of relative frequency when looking at the affix. Significance is indicated next to a given slope with asterisks, which means that only the slope of the dotted line is significantly different from zero.

In general, and in accordance with hypothesis H_{seg1}, we expect higher relative frequency to lead to increased durations of both base and affix. In the AudioBNC, this is indeed the case for affixes, but not for bases: Affix durations in the Audio BNC increase with increasing relative frequency (i.e., increasing segmentability). In other words, in the one case where RELATIVE FREQUENCY interacts with TYPE OF MORPHEME, a relative frequency effect manifests itself only on the affix, not on the base. All the other cases yield a null result: Base durations in the Audio BNC, and affix and base durations in the two New Zealand corpora, do not increase with higher relative frequency. As mentioned above, thus, our data do not convincingly support the segmentability hypothesis H_{seg1}, even though we find one effect in the expected direction. In addition, as discussed above, there were no interactions of RELATIVE FREQUENCY with PROSODIC CATEGORY.

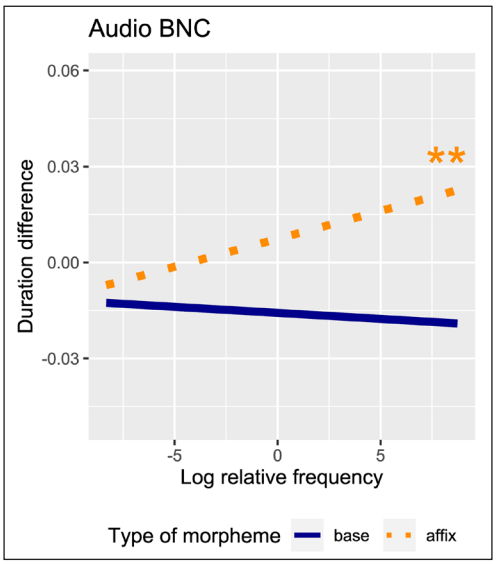

Figure 3: Interaction of TYPE OF MORPHEME with RELATIVE FREQUENCY in the Audio BNC. Levels of slope significance within the conditions are marked by asterisks.

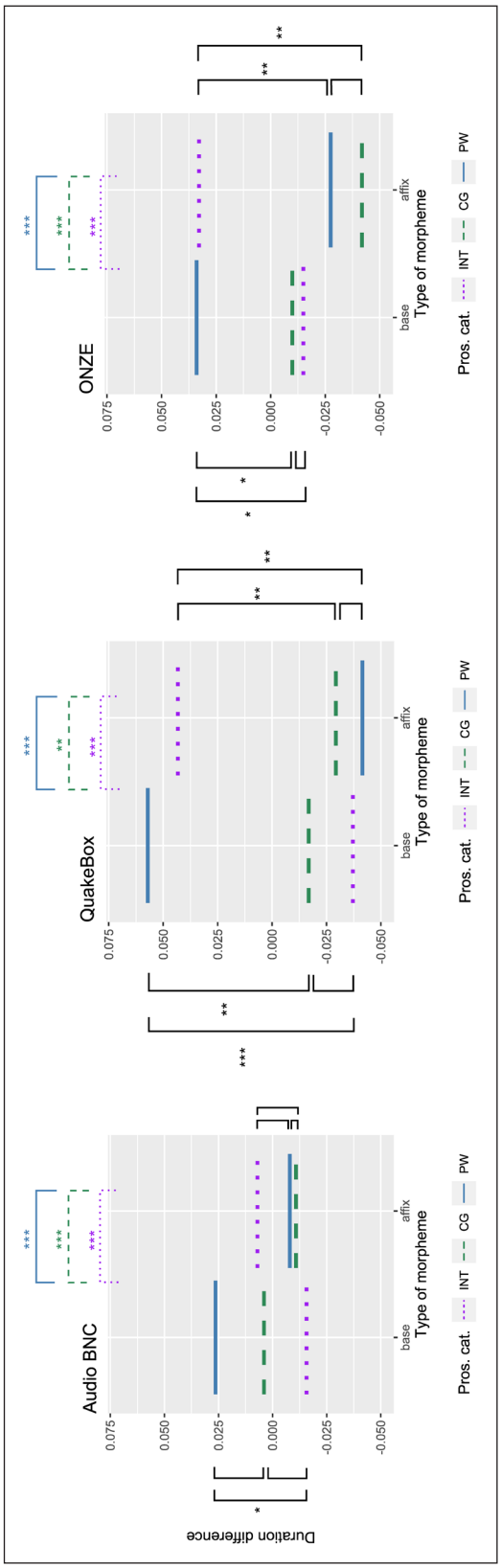

Figure 4: Interaction effects between PROSODIC CATEGORY and TYPE OF MORPHEME in the three corpora. Significance levels of contrasts between conditions are indicated by asterisks on brackets. Abbreviations: `PW` for ProsodicWord, `CG` for CliticGroup, `INT` for Integrating.

Next, we turn to the effects of prosodic category on duration. **Figure 4** shows the interaction effects between prosodic category and type of morpheme on duration difference in the three corpora (Audio BNC, QuakeBox, and ONZE). The x-axis represents the TYPE OF MORPHEME, i.e., the two domains of durational measurement, `base` or `affix`. The y-axis represents DURATION DIFFERENCE. We use purple dotted lines to represent the Integrating (`INT`) condition; green dashed lines to represent the CliticGroup (`CG`) condition; and blue solid lines to represent the ProsodicWord (`PW`) condition. Significance between the contrasts is indicated by asterisks on brackets connecting the respective conditions above and beside each plot. For the contrasts between prosodic categories (Integrating versus CliticGroup versus ProsodicWord) within a type of morpheme condition, significance brackets are given on the left-hand and right-hand side of each plot. For the contrasts between types of morpheme (base versus affix) within a prosodic condition, significance brackets are given above each plot with colored solid, dashed, or dotted lines, corresponding to the prosodic category within which base and affix durations are compared.

First, let us look at the durations of affixes and bases in the different prosodic conditions. **Figure 4** shows that seven of the 18 bars fall below the zero line: This means that in most cases, both affix durations and base durations are shorter than expected from their average segment duration.[9] However, if we compare affixes and bases, we see that there are differences in the amount they deviate from their average segment duration (significance brackets above the plots).

Starting with the Integrating category (`INT`, purple dotted lines), we can see that integrating affixes behave as expected in all three corpora: They are consistently and significantly more lengthened than the bases to which they attach. In the ProsodicWord category (`PW`, blue solid lines), too, affixes behave as expected in all three corpora (they are less lengthened than their bases). In the CliticGroup category (`CG`, green dashed lines), affixes are less lengthened than their bases in all three corpora, while they would have been expected to be more lengthened.

Second, let us examine the contrasts between the prosodic categories within the respective domains of base and affix (significance brackets to the left and right of the plots, respectively). Let us look at the base condition first. We can see that in the Audio BNC, ProsodicWord bases are most lengthened. CliticGroup bases are less lengthened, and Integrating bases are least lengthened. Expressed in terms of the lengthening hierarchy formulated in **Table 1** (Section 2.2.), the Audio BNC thus follows exactly the expected pattern ProsodicWord > CliticGroup > Integrating. However, two of the contrasts are not significant, so we could formulate the

[9] This is expected since derived words are, on average, longer than the average word. Hence, the segments of derived words should be pronounced shorter than they are on average, due to the compression effect discussed in Section 3.2.2. (Lindblom, 1963; Nooteboom, 1972).

more accurate hierarchy ProsodicWord = CliticGroup = Integrating (with ProsodicWord > Integrating). The QuakeBox and ONZE corpora, too, are consistent with the expected lengthening hierarchy. However, the contrast between the clitic-group-forming category and the integrating category in the two New Zealand corpora does not reach significance. This results in the hierarchy ProsodicWord > CliticGroup = Integrating. Bases with prosodic-word-forming affixes are most lengthened, followed by bases with clitic-group-forming affixes and bases with integrating affixes.

Third, let us look at the affix condition. In the Audio BNC, integrating affixes are most lengthened, followed by prosodic-word-forming affixes and clitic-group-forming affixes. This is a different pattern from the expected one (which is CliticGroup > Integrating > ProsodicWord). However, since none of these contrasts are significant, the lengthening hierarchy is more accurately represented as Integrating = ProsodicWord = CliticGroup. In the QuakeBox corpus, Integrating affixes are most lengthened, followed by CliticGroup affixes and ProsodicWord affixes. The contrast between the integrating affixes and the other two affix types is significant, so that we are dealing with this hierarchy: Integrating > CliticGroup = ProsodicWord. Lastly, we also observe deviations from the expected lengthening hierarchy for the ONZE corpus. The ONZE affixes are most lengthened in the Integrating condition, followed by the ProsodicWord condition and the CliticGroup condition. This results in the pattern Integrating > ProsodicWord = CliticGroup.

We can now evaluate all of our contrasts in terms of how the observed lengthening hierarchies overall fit with the expected hierarchies. **Table 7** summarizes the significance levels for all measured contrasts we have discussed in the three corpora, color-coded for the direction of a given effect. For convenience, the last column of this table displays the directional changes in duration we would have expected based on the hypotheses outlined in Section 2.2. We set up the contrasts for this table so that in every cell, we expect more lengthening from the first condition (for example `base` or `CG`) to the second condition (for example `affix` or `PW`). More lengthening is indicated by orange shading. In cases where we observe less lengthening, we use blue shading.

Overall, we can observe that many, but not all of our expectations are confirmed by the data. Out of the 27 expectations (cells), 13 are in line with the predictions, five contradict the predictions, and nine do not reach significance. Prosodic contrasts within the base condition (**Table 7**, rows 4–6) are in line with the expectations, but two of the prosodic contrasts in the affix condition (row 7) go in the opposite direction from what was expected. A similar situation holds for the base-affix contrasts: The Integrating and ProsodicWord conditions (rows 2–3) behave as expected, while the CliticGroup condition (row 1) contradicts our expectations. The nine non-significant contrasts all pertain to the prosodic category contrasts (rows 4–9). Overall, only three out of nine rows are completely in line with the hypotheses.

Significance of ...	Condition	AudioBNC	QuakeBox	ONZE	Expected	Row no.
Base-affix contrasts	CG	base-affix***	base-affix**	base-affix***	base-affix	1
	INT	base-affix***	base-affix***	base-affix***	base-affix	2
	PW	affix-base***	affix-base***	affix-base***	affix-base	3
Prosodic category contrasts	base	CG-PW	CG-PW**	CG-PW*	CG-PW	4
		INT-PW*	INT-PW***	INT-PW*	INT-PW	5
		INT-CG	INT-CG	INT-CG	INT-CG	6
	affix	INT-CG	INT-CG**	INT-CG**	INT-CG	7
		PW-INT	PW-INT**	PW-INT**	PW-INT	8
		PW-CG	PW-CG	PW-CG	PW-CG	9

longer durations from first condition (e.g., base) to second condition (e.g., affix)

shorter durations from first condition (e.g., base) to second condition (e.g., affix)

Table 7: Overview of prosodic category effects in the different conditions. Orange shading indicates a positive change in duration from the first to the second category, blue shading a negative change in duration. Abbreviations: PW for ProsodicWord, CG for CliticGroup, INT for Integrating.

Together with the findings for relative frequency, we can now relate all our findings to the hypotheses as follows. First, we saw earlier that H_{seg1} is largely not supported, despite one effect in the expected direction in the Audio BNC: Higher relative frequency is not associated with longer durations. Second, H_{seg2} is not supported either, as there was not a single significant interaction between RELATIVE FREQUENCY and PROSODIC CATEGORY: How integrated an affix is into prosodic word structure does not modulate whether higher relative frequency can protect against reduction. In both cases, we have very strong evidence for the respective null hypothesis. Third, based on the findings summarized in **Table 7**, quite a few contrasts between bases and affixes within the prosodic categories (H_{PW}, H_{INT}, H_{CG}), as well as contrasts between the prosodic categories within the affix domain ($H_{pros1–6}$), do not behave as expected: Durations often do not pattern according to prosodic boundary strength. The prosodic category approach cannot explain the patterning of the durational differences satisfactorily. We will now proceed to discuss these findings in more detail, returning to the theoretical assumptions of segmentability and prosodic word structure.

5. Discussion

We discuss our findings in the light of the hypotheses given in Section 2, which we repeat here for convenience:

(1) H_{seg1}: The higher the relative frequency of a derivative (i.e., the more segmentable a derivative is), the longer will be its affix duration and base duration.

(2) H_{seg2}: The more prosodically integrated an affix is (the weaker the prosodic boundary), the less likely can higher relative frequency protect the derivative against reduction.

(3) H_{PW}: ProsodicWord bases should be more lengthened than ProsodicWord affixes.
H_{INT}: Integrating affixes should be more lengthened than Integrating bases.
H_{CG}: CliticGroup affixes should be more lengthened than CliticGroup bases.
H_{pros1}: ProsodicWord bases should be more lengthened than Integrating bases.
H_{pros2}: Integrating affixes should be more lengthened than ProsodicWord affixes.
H_{pros3}: CliticGroup affixes should be more lengthened than Integrating affixes.
H_{pros4}: CliticGroup bases should be more lengthened than Integrating bases.
H_{pros5}: CliticGroup affixes should be more lengthened than ProsodicWord affixes.
H_{pros6}: ProsodicWord bases should be more lengthened than CliticGroup bases.

Three main findings emerge from our analysis: (1) H_{seg1} is largely not supported by our data (with one exception), (2) H_{seg2} is not supported by our data, and (3) H_{PW}, H_{INT}, H_{CG} as well as the hypotheses $H_{pros1-6}$ are sometimes contradicted by the data.

First, H_{seg1} is not supported, with one exception: There is one expected, positive effect of relative frequency on duration in the Audio BNC. In this case, we observe longer durations the higher the relative frequency becomes (the more segmentable a derivative is). This speaks in favor of the idea that words are protected against reduction by morphological segmentability (Hay, 2001, 2003). This effect also supports the idea that the more meaningful a stretch of segments, the more important it is for speakers to fully realize these segments so as to facilitate recognition by their interlocutors (Guy, 1980, 1991; Labov, 1989; MacKenzie & Tamminga, 2021), resonating with listener-oriented or *communicative accounts*, as well as with the speaker-oriented accounts of *production ease* (Jaeger & Buz, 2017). If we took this one effect seriously (which we are not convinced we should), this finding would imply that morphological information is somehow still reflected at the subphonemic level. Models which assume an articulator module that has no access to morphological information, and that, therefore, always realizes phonemic representations with pre-programmed gesture templates independently of morphemic status, might not adequately capture the morphology-phonology-phonetics interaction (e.g., Kiparsky, 1982; Levelt, Roelofs, & Meyer, 1999). If a word's morphological structure or semantics cause differences in articulatory gestures, pre-programmed templates are an unlikely architecture.

However, we observe mostly null effects. We are thus able to confirm with more affixes than previous studies that segmentability effects often do not emerge. In a way, our results are in line with previous studies inasmuch as they replicate the mixture of effects and null effects (Hay, 2003, 2007; Plag & Ben Hedia, 2018; Pluymaekers et al., 2005b; Schuppler et al., 2012; Zimmerer et al., 2014; Zuraw et al., 2020). They are also in line with Hanique and Ernestus (2012), one of the earlier attempts to consolidate work on the effects of morphological segmentability on acoustic duration. Reviewing several studies and re-analyzing data by Hay (2003) after identifying methodological issues, these authors conclude that the hypothesis that more easily decomposable words are longer in duration is not convincingly supported.

In addition, research has questioned the very premise on which the segmentability hypothesis is built, namely the rationale that segments which are important for morpheme recognition are enhanced. For instance, Poplack (1980) shows that Puerto Rican Spanish /s/ is frequently deleted even when it is a morphemic plural marker. J. Bybee (2002, 2017) suggests that morphemic strings may be durationally enhanced simply because they occur in contexts which disfavor reduction, compared to contexts of nonmorphemic strings. Finally, Hanique and Ernestus (2012) review several studies investigating the hypothesis that the morphemic status of segments is associated with longer durations and conclude that there is no convincing evidence for such an effect. Segments which are more relevant for the identification of a morpheme (e.g., word-initial segments or morphemic segments, i.e., single-segment morphemes) are not longer than less relevant segments (e.g., word-final segments). Previous studies either have failed to demonstrate an effect of a segment's morphological status on reduction, suffer from methodological issues, or

can be interpreted differently. Several studies, however, support the idea that segments which are more relevant for the identification of a complete *word* are enhanced. This indicates that rather than morphemic structure, it may be word-based information load that affects duration.

The emergence or non-emergence of relative frequency effects might, however, be related to the emergence of word frequency effects in general. We also investigated each of the individual affix categories as a separate dataset and found that we can only observe a relative frequency effect in datasets that also feature a word frequency effect (relative frequency and word frequency were tested in separate regression models due to collinearity). Whenever there is an expected (positive) effect of relative frequency on duration, there is an expected (negative) effect of word frequency in the same affix dataset. It seems that whether relative frequency affects duration is mainly dependent on whether word frequency affects duration. This is not surprising, since these two frequency measures must necessarily be correlated, as one is calculated on the basis of the other. If we assume that relative frequency effects mainly emerge together with word frequency effects, it is not unexpected that there will be some null results, since simple word frequency effects, too, sometimes fail to be observed (e.g., Bowden, Gelfand, Sanz, & Ullman, 2010; Pluymaekers et al., 2005b). The fact that relative frequency effects only emerge in the presence of word frequency effects may, however, also indicate that we should discard relative frequency as a predictor of duration.[10]

Second, H_{seg2} is not supported by our data. Prosodic word structure does not interact with relative frequency in any of the corpora. This means that whether an affix is more or less prosodically integrated does not influence how relative frequency affects duration. A segmentability protection against reduction, then, is not diminished by the lack of a strong prosodic boundary. This conclusion is indirectly also supported by previous studies. Affixes for which effects of relative frequency on duration have been found before are sometimes non-integrating affixes (e.g., *dis-* and *un-* in Plag & Ben Hedia, 2018, *un-* in Hay, 2007) and sometimes integrating affixes (e.g., *-ly* in Hay, 2003). What is more, neither an integrating nor a non-integrating affix guarantees a relative frequency effect, as demonstrated by the null results for both integrating and non-integrating affixes (e.g., *-ly* and *in-* in Plag & Ben Hedia, 2018, *un-* and *in-* in Ben Hedia & Plag, 2017, *ont-*, *ver-*, and *-lijk* in Pluymaekers et al., 2005b). It thus seems that we need to look for further factors that might be responsible for the apparent arbitrariness of the emergence of relative frequency effects. Prosodic word structure is not the culprit.

Third, H_{PW}, H_{INT}, H_{CG}, and the hypotheses $H_{pros1\text{-}6}$ are often not supported. It seems that the prosodic structure of complex words can neither explain the capricious nature of relative

[10] We also tested for an effect of WORD FREQUENCY on DURATION DIFFERENCE for the three complete datasets, using the same model specifications as for the models reported in Section 4, eliminating RELATIVE FREQUENCY and its interactions. An effect of WORD FREQUENCY was present only in the QuakeBox corpus ($t = -2.493$, $p = 0.014$).

frequency effects on duration, nor can it satisfactorily explain durational variation as such. For the base-affix contrasts, the contrasts for the Integrating category and the ProsodicWord category show a consistent and predicted pattern (see again **Table 7**, second and third row). In the CliticGroup category, however, not a single contrast behaves as expected (first row). For the contrasts between prosodic categories, the patterns are more difficult to interpret due to many non-significant effects. As summarized in rows 4 through 6 in **Table 7** and in **Figure 4**, base durations mostly follow the predicted pattern ProsodicWord > CliticGroup > Integrating: Bases with prosodic-word-forming affixes are more lengthened than bases with clitic-group-forming affixes, which in turn are more lengthened than bases with integrating affixes (even though this often does not reach significance). In contrast, affix duration contrasts in two cases contradict the expected pattern CliticGroup > Integrating > ProsodicWord (row 7 in **Table 7**).[11]

It is interesting that predictions for bases are never contradicted by the empirical data, whereas affixes sometimes produce the opposite behavior. The reasons for this difference between bases and affixes are not clear. Both the predictions for bases and the ones for affixes are based on the same underlying prosodic mechanisms assumed in the literature. Affixes are, of course, generally shorter in segments and consequently in duration than their bases, which an analysis of the affix-base ratio confirms for our data. With fewer segments and shorter durations, there might be less potential to reduce or enhance, and thus less potential for effects to play out significantly. While this could explain the fact that there are few significant effects for affixes, this cannot explain why there are effects going in unexpected directions.

One other possibility that has been suggested is that prosodic word effects may underlyingly be related to transitional probability effects (see Côté, 2013). A reviewer points us to the observation that patients who make few anticipatory speech errors often fail to apply cross-boundary phonology (Michel Lange, Cheneval, Python, & Laganaro, 2017), which would support this idea. While Côté (2013) makes this suggestion based on the transitional probability between words, not morphemes, it is conceivable that the transitional probability of an affix might also affect duration. We conducted a separate study with our data (for all three corpora) in which we tested the effects of what we called CONDITIONAL AFFIX PROBABILITY. CONDITIONAL AFFIX PROBABILITY is the probability of the affix given the preceding linguistic element, and thus captures the transitional probability between morphemes. It is calculated by dividing the preceding-element/affix bigram frequency by the frequency of the preceding element itself. Thus, for suffixes, it encodes the probability of the suffix given the probability of the base; for

[11] One reviewer hypothesizes that the fact that we find some significant effects which have no clear theoretical basis is the consequence of analyzing a sample with high variance, in this case natural speech: Plotting standardized effect sizes against their standard errors, we should find a funnel shape where the largest effects (both positive and negative) also have the largest standard errors, and as the standard errors decrease, the estimated effects should shrink to zero. This suspicion was not confirmed by the data; we could not find a clear pattern or strong correlation between standardized coefficients and standard errors (see the scripts at https://osf.io/xuh42/).

prefixes, it encodes the probability of the prefix given the probability of the preceding word. Unfortunately, this measure is very strongly correlated with relative frequency, which makes it impossible to include both measures in the same model. When tested separately from relative frequency, CONDITIONAL AFFIX PROBABILITY very often did not affect duration significantly (similarly to relative frequency). Though an interesting option for further research, transitional probability so far does not provide a better account of our data than prosodic category. The interested reader can access this analysis in the scripts at https://osf.io/xuh42/.

Prosodic structure, if understood in terms of boundary strength, then, is not able to consistently account for the durational differences in English derivatives. This is interesting given the fact that some studies had previously suggested that the prosodic structure of complex words can explain some of their durational variation (Auer, 2002; Bergmann, 2018; Sproat & Fujimura, 1993; Sugahara & Turk, 2009). However, there are some important differences between these studies and the present one.

First, all of these studies investigated specific segments and did not consider other durational domains. Prosodic boundary effects might explain more of the variation in a very small domain, like a single base-final segment, than of the variation in a larger domain, like an affix. Second, some of the studies compared different prosodic categories than we did. For example, Sproat and Fujimura (1993) compared compound boundaries to affix boundaries and to simplex words. The durational difference between a prosodic boundary in an affixed word and a non-affixed word might be greater than the difference between prosodic boundaries in different types of affixed words. Third, some of the studies report that prosodic boundary effects only emerge in very specific conditions. For example, Sugahara and Turk (2009) find that stem-final rhymes in prosodic word-forming suffixes are only lengthened at a slow speech rate, and Bergmann's (2018) segments in derivatives with prosodic word-forming suffixes are only lengthened if they are infrequent. Fourth, most of the studies investigated elicited data, while we investigated conversational data. Fifth, some of the studies suffer from methodological issues. For example, Sugahara and Turk (2009) only used data from five speakers and did not fit any statistical models that included potentially influential covariates. Their durational effects of prosodic structure might disappear if the data were analyzed with more controlled statistical modeling.

Taken together with these caveats, our results show that effects of the prosodic structure of complex words on duration are to be taken with much more caution than assumed. We thus join Pluymaekers et al. (2010, p. 523) in their conclusion that "contrary to received wisdom, morphological effects on fine phonetic detail cannot always be accounted for by prosodic structure."

With regard to the previous work that found evidence for pre-boundary lengthening in general (e.g., Beckman & Pierrehumbert, 1986; Campbell, 1990; Edwards & Beckman, 1988; Klatt, 1975; Vaissière, 1983; Wightman et al., 1992), it must be noted that much of this evidence comes

from boundaries that are higher in the prosodic hierarchy than the ones we investigated in this study. As explained in Section 2.2, prosodic boundaries, and thus lengthening, are expected to be stronger the higher these boundaries are ranked in the prosodic hierarchy. Hence, it is possible that lengthening effects before higher-level boundaries (for example phrase-final lengthening or utterance-final lengthening) show more consistent patterns than lengthening before low-level boundaries (like foot boundaries or pword boundaries). This means that our results do not necessarily have to be interpreted as refuting the prosodic account in general, but perhaps merely as demonstrating that effects become more unstable the lower we zoom in on the hierarchy. It may well be that in order to meaningfully observe prosodic lengthening effects, we need to look at higher-level constituents. To detect effects of very small magnitude, even though we would lose some advantages of spontaneous speech data, future studies could also use an experimental setup. This way, one could manipulate only relative frequency and prosodic category without the variance introduced by real conversations.

Overall, our results show that neither relative frequency nor prosodic word structure are able to explain our data perfectly for all corpora. At least for speech articulation, the predictions of a prosodic account thus are not satisfactory. Of course, the fault for the contradictory contrasts may not necessarily lie in the prosodic predictions themselves, but might also be caused by other, external factors. For example, potential prosodic differences between the varieties represented in the corpora, differences in the forced alignment, or differences in dataset size and statistical power might all be factors which introduce unwanted variation. It is interesting, for instance, that the Audio BNC seems to differ in its behavior somewhat from the two New Zealand corpora, while the results from the two New Zealand corpora are virtually the same (see again **Table 7**).[12] However, since not all affixes are present in all corpora (cf. again the affix distribution across the corpora in **Table 2**), we were not able to test whether an interaction of PROSODIC CATEGORY with CORPUS in a single dataset would turn out significant or not. The differences between corpora might as well be a coincidence and should, in our opinion, be taken with a grain of salt.

One potentially important additional factor which needs to be discussed is syntactic position. In the present study, we did not control for part-of-speech factors. Which syntactic category a derivative belongs to, however, is related to prosodic structure. This is because derivatives of a certain affix category occur more often in certain positions in the sentence, and thus occur more often in certain prosodic positions. For example, the suffix *-less* derives adjectives and therefore

[12] A reviewer suggests that this might be related to affix stress. The affixes included in the Audio BNC were different than those included in the two New Zealand corpora, and some of them are stressed differently. However, it is not straightforwardly possible to incorporate stress as an additional variable, as it is partly nested with prosodic category. Which prosodic category an affix belongs to, after all, is defined partly based on affix stress. Therefore, indirectly, prosodic category already codes for stress. In addition, we already control for the affix as such by using random intercepts for AFFIX. We thus consider it both conceptually and mathematically problematic to test the influence of affix stress within the setup of the present study.

often occurs in prenominal position within a noun phrase or after a copula in predicative position. The suffix *-ness* derives nouns and will therefore often be found in noun phrase-final position. This may affect how these suffixes are pronounced in these contexts and in other contexts (J. Bybee, 2002, 2017). In the present study, we investigated the effect of lower-level prosodic boundaries (foot boundaries, pword boundaries, and clitic group boundaries), but we would also expect duration to be affected by higher-level prosodic boundaries (such as phrase boundaries or utterance boundaries).

The present study did not control for effects of prosodic phrasing since the coding of prosodic phrasing with the natural conversation data as found in our corpora is extremely difficult and highly problematic in terms of rater reliability. We believe, however, that potential effects of prosodic phrasing are not overly influential in our study for several reasons.

First, we investigated a large number of different affixes which belong to different part-of-speech categories simultaneously. It is not the case that in our data, for instance, all prosodic word-forming affixes derive nouns, whereas all integrating affixes derive adjectives. If this were the case, syntactic position could be considered an important potential confound. Instead, our affixes come from different syntactic categories even within prosodic categories. Clitic group-forming affixes derive both nouns and adjectives; integrating affixes derive nouns, adjectives, and verbs; and prosodic word-forming affixes derive nouns, adjectives, verbs, and even adverbs.

Second, as mentioned above, we conducted additional analyses, investigating the tokens in one affix category at a time, for all eight affixes separately. These category-internal analyses again indicated that relative frequency effects frequently fail to emerge, without any clear pattern between categories and within categories (comparing base durations, affix durations, and word durations). The results indicate, then, that relative frequency is neither a reliable predictor within nor across affix categories, prosodic categories, or syntactic categories.

Finally, and crucially, phrase boundaries and utterance boundaries—just as other prosodic boundaries—cause the strongest lengthening of segments which are close to those boundaries, compared to those segments that are further away. In other words, lengthening effects weaken with increasing distance from the boundary. This has been shown in a number of acoustic and articulatory studies (e.g., Berkovits, 1993a, 1993b, 1994; Byrd, Krivokapić, & Lee, 2006; Cambier-Langeveld, 1997; Shattuck-Hufnagel & Turk, 1998). Higher-level boundaries, thus, will affect word-final segment lengthening more than word-internal segment lengthening. This in turn means that our predictions for base lengthening versus affix lengthening as well as the interpretation of the results would remain unchanged: We always expected a word-final element (bases in the case of prefixes, or affixes in the case of suffixes, respectively) to be lengthened more than the corresponding word-internal element. The presence of higher-level boundaries would only enhance that effect.

If relative frequency and prosodic categories cannot explain the durational variation we observe, the question remains whether there are alternative approaches that can. One recent approach that seems promising is to model speech production as the result of a dynamic discriminative learning process, where relations between form and meaning are constantly recalibrated based on the speaker's experience (linear discriminative learning, Baayen, Chuang, & Blevins, 2018; Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019; Chuang et al., 2020). Such an approach does not rely on morphemic or prosodic boundaries, but estimates the strength of associations between sound and meaning based on their discriminative potential. Stein and Plag (2021) and Schmitz, Plag, Baer-Henney, and Stein (2021) demonstrate that it is possible to predict durations of words and affixes based on association measures derived from a linear discriminative learning network. These results indicate that we do not necessarily need boundary strength (morphological or prosodic) to model phonetic detail in speech production. We think it would be worthwhile for future studies to consider such alternative approaches in research on the morphology-phonology-phonetics interface.

6. Conclusion

This study set out to investigate the influence of prosodic structure and relative frequency on the durational properties of complex words. In particular, we wanted to test whether prosodic structure can explain the inconsistency in the emergence of relative frequency effects on duration in previous studies. We showed that for our data, we can rule out prosodic word structure as a gatekeeper for relative frequency effects. Second, we demonstrated that relative frequency effects can emerge, but mostly fail to emerge over a large set of affixes and a large number of tokens. Finally, we have presented evidence that word-internal prosodic boundaries fail to account consistently for durational differences. In many cases, with more strength and a higher position in the prosodic hierarchy, prosodic boundaries are indeed associated with more pre-boundary lengthening of morphological constituents. But in some other cases, this pattern does not hold—especially not for affixes.

What does that mean for the status of word-internal prosodic structure in phonological theory, or in models of speech production? Based on the findings of this study, as well as those of some other studies, we think the jury is still out on the question of how exactly the purported word-internal boundaries translate into articulatory gestures or acoustic properties. While there is good evidence for some phonetic correlates of word-internal prosodic structure (for example base-initial aspiration after certain prefixes, e.g., Zuraw et al., 2020), the durational evidence is much less convincing.

Additional files

We provide the data and scripts for this study at https://osf.io/xuh42/.

Acknowledgements

We thank Jennifer Hay and the New Zealand Institute of Language, Brain and Behaviour at the University of Canterbury in Christchurch, NZ, for their generous support of the first author during his stay at NZILBB. We also thank two anonymous reviewers for their useful suggestions.

Funding information

This research was funded by the Deutsche Forschungsgemeinschaft (Research Unit FOR2373 'Spoken Morphology,' project 'Morpho-Phonetic Variation in English,' PL 151/7-2 and PL 151/8-2), which we gratefully acknowledge.

Competing interests

The authors have no competing interests to declare.

Author contributions

Ingo Plag and Simon David Stein contributed to conception and design of the study. Simon David Stein retrieved and curated the data, performed the modeling and statistical analysis, and wrote the first draft of the manuscript. Funding acquisition and supervision: Ingo Plag. Both authors contributed to manuscript revision, read, and approved the submitted version.

References

Arndt-Lappe, S., & Ernestus, M. (2020). Morpho-phonological alternations: The role of lexical storage. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Trends in Linguistics: Studies and Monographs: Vol. 337. Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon* (pp. 191–227). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110440577-006

Arnold, J. E., & Watson, D. G. (2015). Synthesizing meaning and processing approaches to prosody: Performance matters. *Language, Cognition and Neuroscience*, *30*(1–2), 88–102. DOI: https://doi.org/10.1080/01690965.2013.840733

Auer, P. (2002). Die sogenannte Auslautverhärtung in ne[b]lig vs. Lie[p]lich: Ein Phantom der deutschen Phonologie? In M. Bommes, C. Noack, & D. Tophinke (Eds.), *Sprache als Form: Festschrift für Utz Maas zum 60. Geburtstag* (1st ed., pp. 74–86). Westdeutscher Verlag.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230–268. DOI: https://doi.org/10.1075/ml.18010.baa

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, *2019*, 1–39. DOI: https://doi.org/10.1155/2019/4895891

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. DOI: https://doi.org/10.21500/20112084.807

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198747062.001.0001

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309. DOI: https://doi.org/10.1017/S095267570000066X

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. DOI: https://doi.org/10.1016/j.jml.2008.06.003

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024. DOI: https://doi.org/10.1121/1.1534836

Bell, M. J., Ben Hedia, S., & Plag, I. (2020). How morphological structure affects phonetic realisation in English compound nouns. *Morphology*, 1–34. DOI: https://doi.org/10.1007/s11525-020-09346-6

Ben Hedia, S. (2019). *Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics*. *Studies in Laboratory Phonology: Vol. 8*. Language Science Press. DOI: https://doi.org/10.5281/zenodo.3232849

Ben Hedia, S., & Plag, I. (2017). Gemination and degemination in English prefixation: Phonetic evidence for morphological organization. *Journal of Phonetics*, *62*, 34–49. DOI: https://doi.org/10.1016/j.wocn.2017.02.002

Bergmann, P. (2018). *Morphologisch komplexe Wörter: Prosodische Struktur und phonetische Realisierung*. *Studies in Laboratory Phonology: Vol. 5*. Language Science Press. DOI: https://doi.org/10.5281/ZENODO.1346245

Berkovits, R. (1993a). Progressive utterance-final lengthening in syllables with final fricatives. *Language and Speech*, *36*(1), 89–98. DOI: https://doi.org/10.1177/002383099303600105

Berkovits, R. (1993b). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, *21*(4), 479–489. DOI: https://doi.org/10.1016/S0095-4470(19)30231-1

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, *37*(3), 237–250. DOI: https://doi.org/10.1177/002383099403700302

Boersma, P., & Weenik, D. J. M. (2001). *Praat* (Version 5.4.04) [Computer software]. http://www.praat.org/

Bowden, H. W., Gelfand, M. P., Sanz, C., & Ullman, M. T. (2010). Verbal inflectional morphology in L1 and L2 Spanish: A frequency effects study examining storage versus composition. *Language Learning*, *60*(1), 44–87. DOI: https://doi.org/10.1111/j.1467-9922.2009.00551.x

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261–290. DOI: https://doi.org/10.1017/S0954394502143018

Bybee, J. (2017). Grammatical and lexical factors in sound change: A usage-based approach. *Language Variation and Change*, *29*(3), 273–300. DOI: https://doi.org/10.1017/S0954394517000199

Bybee, J. L. (2000). The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 65–85). Cambridge University Press.

Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *The Journal of the Acoustical Society of America*, *120*(3), 1589–1599. DOI: https://doi.org/10.1121/1.2217135

Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. *Linguistics in the Netherlands*, *14*, 13–24. DOI: https://doi.org/10.1075/avt.14.04cam

Campbell, W. N. (1990). Evidence for a syllable-based model of speech timing. *Proceedings of the International Conference on Spoken Language Processing*, 9–12. http://www.isca-speech.org/archive

Caselli, N. K., Caselli, M. K., & Cohen-Goldberg, A. M. (2016). Inflected words in production: Evidence for a morphologically rich lexicon. *The Quarterly Journal of Experimental Psychology*, *69*(3), 432–454. DOI: https://doi.org/10.1080/17470218.2015.1054847

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). John Wiley & Sons. DOI: https://doi.org/10.1002/0470055464

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, *53*, 945–976. DOI: https://doi.org/10.3758/s13428-020-01356-w

Clopper, C. G., & Turnbull, R. (2018). Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking reduction* (pp. 25–72). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110524178-002

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). *Audio BNC* [Computer software]. University of Oxford. http://www.phon.ox.ac.uk/AudioBNC. DOI: https://doi.org/10.1007/978-1-4614-3894-6

Côté, M.-H. (2013). Understanding cohesion in French liaison. *Language Sciences*, *39*, 156–166. DOI: https://doi.org/10.1016/j.langsci.2013.02.013

Davies, M. (2008). *The Corpus of Contemporary American English* [Computer software]. http://corpus.byu.edu/coca/

Divjak, D. (2019). *Frequency in language: Memory, attention and learning*. Cambridge University Press. DOI: https://doi.org/10.1017/9781316084410

Edwards, J., & Beckman, M. E. (1988). Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica*, *45*(2–4), 156–174. DOI: https://doi.org/10.1159/000261824

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421–436. DOI: https://doi.org/10.1044/1092-4388(2004/034)

Engemann, M., & Plag, I. (2021). Phonetic reduction and paradigm uniformity effects in spontaneous speech. *The Mental Lexicon*, *16*(1), 165–198. DOI: https://doi.org/10.1075/ml.20023.eng

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). SAGE Publications.

Fromont, R. (2003–2020). *LaBB-CAT* [Computer software]. University of Canterbury. https://labbcat.canterbury.ac.nz/

Fromont, R., & Hay, J. (2012). LaBB-CAT: An annotation store. *Proceedings of the Australasian Language Technology Association Workshop*, *10*, 113–117.

Gahl, S. (2008). *Thyme* and *time* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496. DOI: https://doi.org/10.1353/lan.0.0035

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. DOI: https://doi.org/10.1016/j.jml.2011.11.006

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE corpus. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora, Volume 2: Diachronic corpora* (pp. 82–104). Palgrave Macmillan. DOI: https://doi.org/10.1057/9780230223202_4

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In W. Labov (Ed.), *Quantitative Analyses of Linguistic Structure: Vol. 1. Locating language in time and space* (pp. 1–36). Academic Press.

Guy, G. R. (1991). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change*, *3*(1), 1–22. DOI: https://doi.org/10.1017/S0954394500000429

Hall, T. A. (1999). The phonological word: A review. In T. A. Hall & U. Kleinhenz (Eds.), *Current Issues in Linguistic Theory: Vol. 174. Studies of the phonological word* (pp. 1–22). John Benjamins. DOI: https://doi.org/10.1075/cilt.174.02hal

Hanique, I., & Ernestus, M. (2012). The role of morphology in acoustic reduction. *Lingue E Linguaggio*, *11*, 147–164. DOI: https://doi.org/10.1418/38783

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(6), 1041–1070. DOI: https://doi.org/10.1515/ling.2001.041

Hay, J. (2003). *Causes and consequences of word structure*. Routledge. DOI: https://doi.org/10.4324/9780203495131

Hay, J. (2007). The phonetics of *un*. In J. Munat (Ed.), *Studies in Functional and Structural Linguistics: Vol. 58. Lexical creativity, texts and contexts* (pp. 39–57). John Benjamins. DOI: https://doi.org/10.1075/sfsl.58.09hay

Hildebrandt, K. A. (2015). The prosodic word. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 221–245). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199641604.013.035

Hoffmann, S., & Arndt-Lappe, S. (2021). Better data for more researchers: Using the audio features of BNCweb. *ICAME Journal*, *45*(1), 125–154. DOI: https://doi.org/10.2478/icame-2021-0004

Hoffmann, S., & Evert, S. (2018). *BNCweb: CQP Edition* (Version 4.4) [Computer software]. http://bncweb.lancs.ac.uk/

Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. In E. M. Fernández & H. S. Cairns (Eds.), *Blackwell Handbooks in Linguistics. The handbook of psycholinguistics* (1st ed., pp. 38–81). Wiley-Blackwell. DOI: https://doi.org/10.1002/9781118829516.ch3

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–95). MIT Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. J. Hopper (Eds.), *Typological Studies in Language: Vol. 45. Frequency and the emergence of linguistic structure* (pp. 229–254). John Benjamins. DOI: https://doi.org/10.1075/tsl.45.13jur

Kiparsky, P. (1982). Lexical morphology and phonology. In I.-S. Yang (Ed.), *Linguistics in the morning calm: Selected papers from SICOL* (pp. 3–91). Hanshin.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129–140. DOI: https://doi.org/10.1016/S0095-4470(19)31360-9

Kunter, G. (2016). *Coquery* [Computer software]. www.coquery.org

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest* (Version 3.1.2) [Computer software]. https://cran.r-project.org/web/packages/lmerTest/index.html

Labov, W. (1989). The child as linguistic historian. *Language Variation and Change*, *1*(1), 85–97. DOI: https://doi.org/10.1017/S0954394500000120

Ladefoged, P., & Johnson, K. (2011). *A course in phonetics* (6th ed.). Wadsworth Cengage Learning.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. DOI: https://doi.org/10.1017/S0140525X99001776

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. DOI: https://doi.org/10.1121/1.1918816

Losiewicz, B. L. (1995). Word frequency effects on the acoustic duration of morphemes. *The Journal of the Acoustical Society of America*, *97*(5), 3243. DOI: https://doi.org/10.1121/1.411745

Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation*. Epocha Publishing House.

MacKenzie, L., & Tamminga, M. (2021). New and old puzzles in the morphological conditioning of coronal stop deletion. *Language Variation and Change*, *33*(2), 217–244. DOI: https://doi.org/10.1017/S0954394521000119

Michel Lange, V., Cheneval, P. P., Python, G., & Laganaro, M. (2017). Contextual phonological errors and omission of obligatory liaison as a window into a reduced span of phonological encoding. *Aphasiology*, *31*(2), 201–220. DOI: https://doi.org/10.1080/02687038.2016.1176121

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*, *44*(4), 778–792. DOI: https://doi.org/10.1044/1092-4388(2001/061)

Nespor, M., & Vogel, I. (2007). *Prosodic phonology*. Walter de Gruyter. DOI: https://doi.org/10.1515/9783110977790

Nooteboom, S. G. (1972). *Production and perception of vowel duration: A study of the durational properties of vowels in Dutch*. University of Utrecht.

OED (2020). *Oxford English Dictionary Online*. Oxford University Press. www.oed.com

Plag, I. (2018). *Word-formation in English* (2nd ed.). Cambridge University Press. DOI: https://doi.org/10.1017/9781316771402

Plag, I., & Ben Hedia, S. (2018). The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration. In S. Arndt-Lappe, A. Braun, C. Moulin, & E. Winter-Froemel (Eds.), *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity* (pp. 93–116). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110501933-095

Plag, I., Dalton-Puffer, C., & Baayen, R. H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, *3*(2), 209–228. DOI: https://doi.org/10.1017/S1360674399000222

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*(1), 181–216. DOI: https://doi.org/10.1017/S0022226715000183

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005a). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2–4), 146–159. DOI: https://doi.org/10.1159/000090095

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005b). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569. DOI: https://doi.org/10.1121/1.2011150

Pluymaekers, M., Ernestus, M., Baayen, R. H., & Booij, G. E. (2010). Morphological effects on fine phonetic detail: The case of Dutch *-igheid*. In A. Lahiri, C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Phonology and Phonetics. Laboratory Phonology 10* (pp. 511–531). Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110224917.5.511

Poplack, S. (1980). The notion of the plural in Puerto Rican Spanish: Competing constraints on (s) deletion. In W. Labov (Ed.), *Quantitative Analyses of Linguistic Structure: Vol. 1. Locating language in time and space* (pp. 55–67). Academic Press.

R Core Team. (2020). *R* (Version 4.0.1) [Computer software]. R Foundation for Statistical Computing. Vienna. http://www.R-project.org/

Raffelsiefen, R. (1999). Diagnostics for prosodic words revisited: The case of historically prefixed words in English. In T. A. Hall & U. Kleinhenz (Eds.), *Current Issues in Linguistic Theory: Vol. 174. Studies of the phonological word* (pp. 133–201). John Benjamins. DOI: https://doi.org/10.1075/cilt.174.07raf

Raffelsiefen, R. (2007). Morphological word structure in English and Swedish: The evidence from prosody. In G. E. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli, & S. Scalise (Chairs), *Online Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*, Fréjus.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. DOI: https://doi.org/10.2307/271063

Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. D. (2021). Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, *12*, 680889. DOI: https://doi.org/10.3389/fpsyg.2021.680889

Schuppler, B., van Dommelen, W. A., Koreman, J., & Ernestus, M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*(4), 595–607. DOI: https://doi.org/10.1016/j.wocn.2012.05.004

Shattuck-Hufnagel, S., & Turk, A. (1998). The domain of phrase-final lengthening in English. *The Journal of the Acoustical Society of America*, *103*(5), 2889. DOI: https://doi.org/10.1121/1.421798

Sóskuthy, M., & Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand English. *Cognition*, *166*, 298–313. DOI: https://doi.org/10.1016/j.cognition.2017.05.032

Sproat, R., & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, *21*(3), 291–311. DOI: https://doi.org/10.1016/S0095-4470(19)31340-3

Stein, S. D., & Plag, I. (2021). Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with linear discriminative learning. *Frontiers in Psychology*, *12*, Article 678712. DOI: https://doi.org/10.3389/fpsyg.2021.678712

Sugahara, M., & Turk, A. (2009). Durational correlates of English sublexical constituent structure. *Phonology*, *26*, 477–524. DOI: https://doi.org/10.1017/S0952675709990248

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, *57*(1), 1–39. DOI: https://doi.org/10.1017/S0022226719000203

Torreira, F., & Ernestus, M. (2009). Probabilistic effects on French [t] duration. *Interspeech*, 448–451. DOI: https://doi.org/10.21437/Interspeech.2009

Tucker, B., Sims, M., & Baayen, R. H. (2019). Opposing forces on acoustic duration. *PsyArXiv*, 1–38. (Preprint submitted to Elsevier). DOI: https://doi.org/10.31234/osf.io/jc97w

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375–400. DOI: https://doi.org/10.1075/ml.11.3.03tuc

Turnbull, R. (2018). Patterns of probabilistic segment deletion/reduction in English and Japanese. *Linguistics Vanguard*, *4*(s2). DOI: https://doi.org/10.1515/lingvan-2017-0033

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Springer Series in Language and Communication: Vol. 14. Prosody: Models and measurements* (pp. 53–66). Springer. DOI: https://doi.org/10.1007/978-3-642-69103-4_5

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 481–487. DOI: https://doi.org/10.3758/BF03195594

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. DOI: https://doi.org/10.3758/BF03194105

Walsh, L., Hay, J., Derek, B., Grant, L., King, J., Millar, P., Papp, V., & Watson, K. (2013). The UC QuakeBox Project: Creation of a community-focused research archive. *New Zealand English Journal*, *27*, 20–32. DOI: https://doi.org/10.26021/2

Watson, D. G., Buxó-Lugo, A., & Simmons, D. C. (2015). The effect of phonological encoding on word duration: Selection takes time. In L. Frazier & E. Gibson (Eds.), *Studies in Theoretical Psycholinguistics: Vol. 46. Explicit and implicit prosody in sentence processing* (pp. 85–98). Springer. DOI: https://doi.org/10.1007/978-3-319-12961-7_5

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717. DOI: https://doi.org/10.1121/1.402450

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, *38*(3), 329–336. DOI: https://doi.org/10.1016/j.wocn.2010.04.003

Zimmerer, F., Scharinger, M., & Reetz, H. (2014). Phonological and morphological constraints on German /t/-deletions. *Journal of Phonetics*, *45*, 64–75. DOI: https://doi.org/10.1016/j.wocn.2014.03.006

Zuraw, K., Lin, I., Yang, M., & Peperkamp, S. (2020). Competition between whole-word and decomposed representations of English prefixed words. *Morphology*, *31*, 201–237. DOI: https://doi.org/10.1007/s11525-020-09354-6