Changing word usage predicts changing word durations in New Zealand English

Márton Sóskuthy[a], Jennifer Hay[b]

[a] *University of York, UK*
[b] *University of Canterbury, NZ*

Abstract

This paper investigates the emergence of lexicalized effects of word usage on word duration by looking at parallel changes in usage and duration over 130 years in New Zealand English. Previous research has found that frequent words are shorter, informative words are longer, and words in utterance-final position are also longer. It has also been argued that some of these patterns are not simply online adjustments, but are incorporated into lexical representations. While these studies tend to focus on the synchronic aspects of such patterns, our corpus shows that word-usage patterns and word durations are not static over time. Many words change in duration and also change with respect to frequency, informativity and likelihood of occurring utterance-finally. Analysis of changing word durations over this time period shows substantial patterns of co-adaptation between word usage and word durations. Words that are increasing in frequency are becoming shorter. Words that are increasing/decreasing in informativity show a change in the same direction in duration (e.g. increasing informativity is associated with increasing duration). And words that are increasingly appearing utterance-finally are lengthening. These effects exist independently of the local effects of the predictors. For example, words that are increasing utterance-finally lengthen in all positions, including utterance-medially. We show that these results are compatible with a number of different views about lexical representations, but they cannot be explained without reference to a production-perception loop that allows speakers to update their representations dynamically on the basis of their experience.

Keywords: New Zealand English, word duration, word frequency, informativity, production-perception loop, language change

1. Introduction

It is well-established that a number of usage factors affect word duration – including frequency, the word's predictability in context, and the position of the word in relation to utterance boundaries. In theory, there are two ways in which such effects can be realized (see, e.g. Bybee 2002; Jaeger and Buz to appear). First, they can manifest as context-dependent, local adjustments that apply online during speech production. The existence of such local effects is uncontroversial, and they are the main focus of a large portion of the literature on variation in word duration. But usage-based effects can also manifest as offline lexicalized changes that affect words regardless of their context. Recent research based on synchronic corpus data shows that such lexical effects may exist alongside local effects (Seyfarth, 2014), and suggests that the two are linked: changes to lexical representations arise through repeated exposure to local effects (this idea is already anticipated in Paul 1880, p. 46).

This paper presents an empirical investigation of the emergence of lexical effects on word duration. Such lexical effects likely exist at all points in the history of a language, so it is not possible to look at their 'ultimate' origin. Instead, we focus on a specific question that can be investigated using relatively recent historical data: what happens to word duration when a word's usage patterns are not stable over time? In such situations, it should be possible to directly observe the emergence of lexical effects in the form of co-adaptation between usage and form. Therefore, we ask the following questions: Can patterns of changing word usage predict patterns of changing word production? Is there evidence that lexical representations are directly impacted

Email addresses: marton.soskuthy@york.ac.uk (Márton Sóskuthy), jen.hay@canterbury.ac.nz (Jennifer Hay)

by changing word usage patterns?

The research presented in this paper extends previous work substantially by tracking word duration trajectories and changes to word usage over time in a diachronic corpus. Our data set comes from the spoken Origins of New Zealand English corpus (ONZE, Gordon et al. 2007), which contains speech samples from over 500 speakers born between 1851 and 1987. We track changes to 698 content words represented by more than 270,000 tokens, focusing on word usage, word duration and the extent to which they change together.

Using this unique data set, we are able to obtain a direct view of the accumulation of usage-based effects in lexical representations over time. These show up in the form of robust parallels between changes in word duration and usage. We suggest that these findings are compatible with a range of different views about lexical representations, but are difficult to explain without reference to the so-called *production-perception loop* (Pierrehumbert, 2001; Wedel, 2007).

The paper is structured as follows. In section 2, we summarize observations about patterns of variation in word duration and briefly describe the potential pathways that can lead to lexical patterns, with special emphasis on the production-perception loop. Section 3 sets out our synchronic and diachronic hypotheses relying on the discussion in the preceding section. Section 4 first gives an overview of the spoken diachronic corpus that serves as the basis of the project, and then defines our key variables. Section 5 plays a mainly descriptive role, presenting general patterns of change in word duration and usage factors based on the corpus, and setting the scene for the main statistical analysis presented in section 6. Section 7 concludes the paper with a discussion of the results and an evaluation of the hypotheses, along with some more general conclusions about the nature of language change.

2. Background

2.1. Word duration and usage factors

One of our key variables is word duration, defined as the duration of spoken word forms measured in seconds. Word duration varies substantially as a function of frequency, predictability, repetition, syntactic probability and a range of other variables (Whalen, 1991; Jurafsky et al., 2001; Bell et al., 2003, 2009; Gahl, 2008; Tily et al., 2009; Seyfarth, 2014). This paper uses the term *usage factor* to refer to these variables collectively. We do not assign special theoretical significance to variation in word duration, and simply take it to be one of the many phonetic reflexes of more general processes of hypo- and hyper-articulation (cf. Lindblom et al. 1995) conditioned by usage factors. Other examples of such reflexes include variation in segmental and syllabic duration, the peripherality of vowels and consonant deletion (Jurafsky et al., 2001; Bybee, 2002; Aylett and Turk, 2006; Cohen Priva, 2015).

As noted in the introduction, patterns of variation in word duration can be divided into two types based on the way they are expressed: as differences between tokens of the same word in different local contexts, or as context-independent differences across multiple lexical items. These patterns will be labelled *local* and *lexical*, respectively.

An example of a local pattern is the effect of predictability from the preceding or following context: words tend to be shorter in predictable contexts (Jurafsky et al., 2001; Bell et al., 2009; Seyfarth, 2014). Since words typically appear both in high and low predictability contexts (e.g. the word *hunt* in *witch hunt* vs. *which hunt*; Lieberman 1963), they display within-item local variation based on predictability. There are a variety of proposals about the mechanisms through which contextual predictability comes to be related to reduced forms, some of which relate to speaker-based factors such as ease of access or planning, and some of which are more listener-oriented, relating to appropriately conveying the intended message. A good recent outline of various accounts is provided in Jaeger and Buz (to appear). The topic of interest in this paper is the potential accumulated consequences of these local forces at the lexical level.

As opposed to local patterns, lexical patterns are stable across contexts for each word, but vary across different words. A simple example of a lexical pattern is the effect of unigram word frequency: high-frequency items tend to be shorter than low-frequency items (e.g. Gahl 2008; Bell et al. 2009). Since the unigram frequency of a word is not context-dependent, a given lexical item will always show the same effect of frequency, and the effect of frequency can only be seen by comparing multiple lexical items.

Before we take a closer look at the specific usage factors investigated in this paper, it will be useful to provide a brief overview of the types of correlations we may observe between changes in usage factors and word durations. There is a trivial sense in which shifts in the distribution of local conditioning factors may lead to changes in observed word durations. All things being equal, a word that becomes more predictable in a given context will also undergo more shortening in that context, which also lowers its average duration. Such par-

allel changes between word duration and usage factors are superficial in the sense that they do not affect lexical representations. Although the surface distribution of word durations may change along with the word's predictability in specific contexts, this change simply and directly reflects the online local reductive forces at work. Those tokens of the word that happen to occur in low-predictability contexts will not undergo shortening.

In this paper, we are particularly interested in lexical changes that go beyond local conditioning factors and whose effects are not dependent on the immediately local context – in other words, changes that arguably take place at the level of lexical representations. An example of such an effect is presented by Seyfarth (2014), who demonstrates that words which tend to occur in predictable contexts are shorter *even when their local predictability is low*. He argues that such lexical effects reflect stored patterns of reduction, which come about through repeated exposure to local reductive biases. The crucial step in his analysis is the separation of two different effects: local predictability and a cumulative measure of predictability calculated over all contexts for a given word, called informativity (cf. below). He shows that informativity has an independent contribution to word duration even after local predictability and a range of other control variables have been taken into account. This paper follows Seyfarth (2014) in separating local and lexical measures and looking for an independent contribution of the latter in an attempt to detect changes that affect lexical representations.

We focus on three main groups of usage factors: predictability, position within the utterance and frequency. Predictability can be defined on many different levels and in many different ways. One of the simplest definitions is based on immediately adjacent words: the conditional probability of a word x given a preceding or a following word y, which is usually approximated through the following equation (where $p(x|y)$ stands for the conditional probability of x given y, $c(xy)$ is the number of times x and y occur together in a corpus and $c(y)$ is the frequency of y in the same corpus; see e.g. Jurafsky et al. 2001; Bell et al. 2009; Seyfarth 2014):

$$p(x|y) = \frac{c(xy)}{c(y)} \tag{1}$$

As explained above, predictability is a local measure, with a corresponding lexical measure called informativity. Informativity is closely related (although not identical) to average predictability. Following Piantadosi et al. (2011) and Seyfarth (2014), we define informativity as follows:[1]

$$\text{info}(x) = -\sum_{c=1}^{n} p(y_c|x)\ log\ p(x|y_c) \tag{2}$$

In other words, informativity is the average surprisal (where surprisal is a decreasing function of predictability) for a word x calculated over all contexts y_c that it appears in, weighted by the frequency with which it appears in each context. Informativity is low for words that tend to be predictable from their context, and high for words that tend to be unpredictable. Seyfarth (2014) provides two examples: the word *current* is often predictable from the following word (e.g. *events*, *news*, *president*), so it has low informativity based on the following context, while the word *nowadays* is rarely predictable from the following word, so it has high informativity. As noted above, predictability has been shown to correlate negatively with word durations (Jurafsky et al., 2001; Bell et al., 2009; Seyfarth, 2014), while informativity has a slightly weaker positive effect on word durations (Seyfarth, 2014). In English, these effects are especially robust when predictability and informativity are calculated on the basis of the following context (Bell et al., 2009; Seyfarth, 2014).

Our second set of usage factors relates to the position of the word in relation to the final utterance boundary. Words in utterance-final position are considerably longer than they are utterance-medially and initially (e.g. Klatt 1976; Wightman et al. 1992; Turk and Shattuck-Hufnagel 2007). Assuming that consistent exposure to local biases can lead to lexical effects, we also expect that words which are frequently utterance-final should be longer than words which tend not to occur utterance-finally. An example of a content word that is frequently utterance-final (at least in the ONZE corpus) is *today*, which appears in final position nearly 20% of the time; an example of a content word that rarely appears utterance-finally is *make*, which only occurs in final position about 0.5% of the time. Gahl (2008) reports an effect that seems consistent with this prediction: she finds that words that are frequently prepausal are significantly longer than words that tend not to occur before pauses (the presence of pauses is presumably strongly correlated with utterance-final position). However, her

[1] We base our measure on the natural logarithm of predictability, which means that the basic unit of informativity in this paper is the so-called *nat*. Previous research has quantified informativity using other units such as *bans* (using base-10 logarithm; Seyfarth 2014). Since these values are linearly correlated, the choice of *nats* as opposed to *bans* or *bits* (base-2 logarithm) does not affect the results from our regression models.

model does not control for the local effect of position in the utterance, which makes it difficult to tell whether this is a local or lexical effect. To separate these two effects, we look both at the local effect of utterance-final position and the lexical effect of typical position within the utterance (similarly to the case of predictability and informativity).

The third usage factor that we investigate is unigram word frequency. As noted above, word frequency has a negative effect on word duration (Gahl, 2008; Bell et al., 2009), although recent studies have found that the effect of word frequency is less robust in statistical models that also incorporate informativity (Piantadosi et al., 2011; Seyfarth, 2014). Word frequency (as defined here) differs from our other lexical factors in that it does not have a corresponding local factor. Thus, while the availability of local and lexical factors makes it possible to isolate online versus lexicalized contributions for predictability / informativity and position within the utterance, we cannot do the same for frequency. It is not possible to tell whether the frequency effects we report below are due to low-level distributional shifts or deeper representational changes.

2.2. Pathways to lexical effects

Lexical effects go beyond local effects in that they rely on information that forms part of lexical representations. There are two interconnected but conceptually distinct issues concerning their origins: what is the nature of the lexical information that underlies these effects and how does it make its way into lexical representations?

The lexical effects reported in the literature are small and gradient. This implies that lexical representations must contain at least some additional information beyond a single abstract categorical form. There are a number of different proposals as to the nature of this information (cf. Seyfarth 2014; Jaeger and Buz to appear). Perhaps the most straightforward one is that lexical representations incorporate fine-grained phonetic information. This information could be stored in the form of phonetically detailed exemplar clouds (e.g. Bybee 2001; Pierrehumbert 2002), word-specific detail about the tightness of intergestural timing relations (Lavoie, 2002) or a single phonetically detailed default form for each word (Seyfarth, 2014, p. 151). An alternative view is that words have multiple abstract categorical representations, and gradient differences arise from differences in the relative frequencies of these variants (Bürki et al. 2010; Seyfarth 2014, p. 150). Finally, it is also possible that lexical representations do not contain any *phonetic* detail, but they do contain information about cumulative usage statistics (e.g. average predictability; Seyfarth 2014, p. 151). Under this view, lexical effects are not offline but online, arising during production as a function of information about word usage. We attempt to relate these different views about the nature of lexical representations to our findings in section 7.

In this paper, the focus is on the second issue identified above: how does this information become incorporated into lexical representations? A plausible explanation comes from the so-called production-perception loop (Pierrehumbert, 2001; Oudeyer, 2006; Wedel, 2006; Sóskuthy, 2015). The production-perception loop is a hypothesized evolutionary pathway in speech. This pathway requires two conditions to be met: (i) detailed lexical representations of the type described above and (ii) an ability to update these representations as a function of linguistic experience. If these conditions hold, any production by a member of a given speech community has some probability of influencing future productions within that speech community, thereby creating a loop. If a consistent bias in production or perception enters this loop (e.g. a given word frequently appears in predictable contexts and therefore consistently undergoes a small amount of reduction), the update of speech representations will be overwhelmed by biased variants, and the bias may leave a permanent mark on these representations. In the case of gradient biases, this could result in substantial shifts as the continuous incorporation of biased variants into representations pushes production targets further and further (see e.g. Pierrehumbert 2001).

As noted above, the existence of gradient lexical effects alone is a strong argument for the presence of some type of detail in lexical representations. There is also a line of research indicating that these representations are regularly updated to include information about novel exemplars (Goldinger 2000; Hay and Maclagan 2012; Hay and Foulkes 2016). Moreover, there is a range of results that would be very difficult to interpret without reference to some mechanism akin to the production-perception loop. These include the finding that frequent (and in some cases infrequent) words are not only ahead of other words in sound changes, but also increase their advantage over time (Hay and Foulkes, 2016; Hay et al., 2015); Seyfarth's (2014) finding that low informativity leads to decreased word durations even after we control for local predictability; and the general observation that extremely high-frequency words show a degree of reduction that far surpasses the online reduction effects found in studies of word duration (e.g. Bybee 2001).

In sum, when a word undergoes systematic local biases on its production, the production-perception loop

provides a mechanism through which these biases are predicted to accumulate in the word's lexical representation.

3. Hypotheses

While there are many examples of phenomena that are arguably the result of the production-perception feedback loop, there are few studies that look at the emergence of word-specific patterns in real time, and no studies that connect the emergence of these patterns to changes in usage. This study focuses specifically on these areas.

One prerequisite for this research is to establish that our corpus does, in fact, show lexicalized effects of frequency, informativity and typical position within the utterance, which are independent of local effects. If such synchronic lexical effects did not exist in our data set, that would also mean that we cannot look at their diachronic emergence. Therefore, although the main focus of this paper is on changes to word duration, we will also attempt to replicate previous findings about frequency and informativity effects (Gahl, 2008; Bell et al., 2009; Seyfarth, 2014), and to extend this line of investigation to the effects of typical position within the utterance. As explained in section 2.1, the synchronic hypotheses would predict overall lexical effects, such that frequent and low-informativity words will be shorter, while words that are often utterance-final will be longer.

We now turn to our main diachronic hypotheses, which focus on the idea that the emergence of lexical effects should be directly observable when the usage of a word changes.

(3) Diachronic hypotheses:

   a. *informativity*: words that are decreasing in informativity will also decrease in duration compared to other words

   b. *utterance-final*: words that are becoming increasingly frequent utterance-finally will increase in duration compared to other words

   c. *frequency*: words that are becoming more frequent will decrease in duration compared to other words

We expect to see these dynamic effects after controlling for local predictability and position in the utterance.

These predictions follow straightforwardly from the notion of the production-perception loop. Changes in usage lead to changes in the distribution of local biases. If lexical effects reflect the accumulation of local biases, and the changes in local biases are sufficiently large, we expect to see concurrent changes in lexical representations.

4. Data

4.1. Corpus and measurements

The data analysed in this paper come from the spoken diachronic ONZE corpus (Gordon et al., 2007), which consists of three sub-corpora: the Mobile Unit archive (collected between 1946–1948, speakers born 1851–1900), the Intermediate Archive (collected between 1960 and the 1990s, speakers born 1891–1963 – with most born before 1935) and the Canterbury Corpus (collected after 1994, speakers born 1926–1987). The recordings in these archives are predominantly informal interviews. Together, the three corpora contain recordings from over 500 speakers born over a period of 136 years (1851–1987).

Overall, the corpus contains 2.1 million word tokens. These words were automatically aligned with corresponding orthographic transcriptions (using algorithms from the HTK Speech Recognition toolkit; Young et al. 1997) and stored in a searchable database using the LaBB-CAT software package (Fromont and Hay, 2008). The automatic alignments were then used to generate word duration measurements and other measures (see below) for all the words in the corpus with the help of LaBB-CAT. The fact that we used automatic methods to extract word durations means that the data set inevitably contains some measurement errors. However, forced-alignment tends to be relatively accurate at the word level. Tests of the particular aligner we used show good levels of accuracy for speech samples of over 5 minutes (Fromont and Watson, in press), which is true of the majority of our recordings. Moreover, while the errors introduced by forced-alignment likely decrease the power of our statistical analyses by adding random noise to the measurements, there is no reason to assume that they introduce problematic systematic biases into our study.

This study is based on a smaller subset of the ONZE word duration data set containing measurements for 271,764 word tokens representing 698 word types. We only include content words in our data set. The rationale for this decision is that the durations of content and function words show differential conditioning with respect to usage factors (Bell et al., 2009), and content words are much more diverse in terms of word frequency, informativity and other predictors. The set of content words was further filtered to only include word

types that occurred at least 50 times in the ONZE corpus and were well represented over the entire time period. A range of further filters were applied to the data to remove word duration outliers (likely due to measurement errors) and other problematic data points. The full details of the filters that we applied to the data set are presented in the supplementary materials.[2]

Following previous research (Gahl, 2008; Bell et al., 2009; Seyfarth, 2014), our data set is based on word forms, not lemmas (e.g. both *year* and *years* are included). This decision was motivated by the fact that word forms representing the same lemma often differ substantially with respect to usage factors (e.g. *years* is nearly twice as frequent as *year*). Word forms were defined on the basis of the orthographic transcripts. Since the ONZE corpus did not include any semantic or syntactic labels, no attempt was made to deal with cases where a given spelling could represent more than one lexeme (e.g. *block N* vs. *block V*; see e.g. Gahl 2008 for a similar approach). It is unlikely that this shortcoming of the data set had a substantial impact on our results. While we have no way of estimating the extent of within-word-class homonymy (e.g. *chest N* 'body part' vs. *chest N* 'large box'), we performed an informal analysis of across-word-class homonymy based on automatic part-of-speech tags generated by the NLTK toolkit (Bird et al., 2009). This analysis suggests that tokens belonging to the most frequent word class for a given spelling account for over 95% of the data set. The remaining 5% along with the presumably even smaller proportion of within word class homonyms is unlikely to be a source of major biases.

4.2. Definitions of key variables

This section is an overview of the key predictors related to our hypotheses. The statistical models presented later in this paper also contain some additional control variables, which will be discussed in Section 6 as part of the model descriptions. Predictors marked by the label *log* are transformed logarithmically.

Following standard practice in sociolinguistics (Bailey, 2008), the variable YEAR OF BIRTH (the year of birth of the speaker who uttered the word form) is used as a tool for tracking change over time. As discussed elsewhere (Hay et al., 2015), using time of recording would not be practical in this corpus, as it would simply cluster speakers into three main groups, in a way that is highly collinear with year of birth. Note that if speakers move in the direction of change during their lifespan, then the main danger of using year of birth as a proxy for change is that it may somewhat underestimate the speed of change.

4.2.1. Local predictors

While we control for many local predictors, the following local predictors are key to our hypotheses. This is because we separately hypothesize that there are cumulative word-level effects of these local biases (cf. the related lexical predictors in section 4.2.2).

PREVIOUS AND FOLLOWING PREDICTABILITY (log): Abbreviated as PREV/FOLL PREDICTABILITY in tables. Two separate variables representing the bigram probability of the word form based on the previous and the following word (cf. equation (1) in section 2.1). These values were calculated using all 2.1 million words in our corpus. Following Jurafsky et al. (2001) and Seyfarth (2014), we smoothed the probabilities using the modified Kneser-Ney algorithm in the SRILM toolkit (Stolcke et al., 2011).

UTTERANCE-FINAL: A binary variable indicating whether the word is utterance-final or not. As part of the transcription and alignment process for ONZE, transcribers manually demarcated the interviews into a series of intervals. The guidelines were open to interpretation, asking transcribers to 'start each major utterance with a breakpoint.' The guidelines provide an example showing intervals containing 2-11 words, with breaks at substantial pauses or major clause boundaries. Transcribers also likely used intonational cues to guide their decisions about utterance breaks. As these guidelines are not very specific, this variable is highly correlated with true utterance-finality, but also contains some noise. There is no reason to think that this noise might be unevenly distributed, however, or bias the results in any systematic way.

4.2.2. Lexical predictors

The following variables are designed to test the role of lexical factors, including lexical factors that result from accummulated local distributions.

PREVIOUS AND FOLLOWING INFORMATIVITY: Abbreviated as PREV/FOLL INFORMATIVITY in tables. Calculated from previous and following predictability using equation (2) in section 2.1.

PROPORTION UTTERANCE-FINAL (log): Abbreviated as PROPORTION UTTR-FINAL in tables. The proportion of tokens that were utterance-final for a given word form in our corpus.

FREQUENCY (log): The frequency of the word based on the British subset of the Google Books N-gram Cor-

[2] The complete data set and the code for the main analysis are available from `https://osf.io/q5wgh/`.

pus (Michel et al. 2011; restricted to texts published between 1851 and 1987). The British subset of Google N-grams contains frequency counts for words extracted from books published in Great Britain, with separate frequency counts for each year of publication. We obtained our static word frequency measure by averaging over the counts across the entire time period. Although using frequency estimates from a written corpus that is based on a different dialect is not without problems, there are two good reasons for relying on Google N-grams instead of our own corpus. First, Google N-grams contains between 300 million and one billion word tokens for every year, which makes both our static and dynamic estimates extremely robust. Second, as we show in section 5, our corpus displays a substantial overall decline in average word duration over time. As a result, a word that is more frequent in later recordings will be, on average, shorter than other words simply because it is more strongly represented in the later section of the corpus (which shows shorter word durations in general). Word duration slopes are also likely affected by this confound. Although Google N-grams frequencies are correlated with within-corpus frequencies, this correlation is only medium-strength (Pearson's $r = 0.45$), which means that the confounding effects described above can be attenuated by using the former in the place of the latter. In order to prevent issues due to major discrepancies between Google N-grams estimates and actual frequencies in spoken NZE, a small number of frequent New Zealand place names were removed from our corpus at the filtering stage.

4.2.3. Dynamic lexical predictors

The variables designed to test the main hypotheses in section 3 are derived from the lexical predictors as follows:

Change in previous/following informativity: We first divided the corpus into two halves according to speaker year of birth. The first half contained data from speakers born before or in 1930, and the second half contained data from speakers born after 1930. We decided to place the dividing line at 1930 in an effort to create two sections that each spanned a substantial time period and contained roughly equal numbers of tokens. Separate bigram language models were fit to the the pre-1930 and the post-1930 sections of the corpus. The smoothed estimates of predictability from these models were then used to calculate following and previous informativity for each half of the corpus. Change in these predictors was calculated simply by subtracting the pre-1930 value from the post-1930 value. Both of these predictors are approximately normally distributed.

Change in proportion utterance-final: Again, proportion utterance-final was calculated separately for the pre-1930 and post-1930 sections of the corpus. Change in proportion utterance-final is the logarithm of the ratio of the post-1930 value and the pre-1930 value, and is approximately normally distributed.

Change in frequency: The slope of a regression line fit through log frequency as a function of year of publication in the Google N-grams corpus. This is more or less equivalent to the logarithm of the expected growth / decrease in frequency over a single year expressed as a ratio. It is approximately normally distributed with a slight positive skew.

Admittedly, the way we have operationalized these measures results in a loss of information about potential non-linearities in the development of the relevant quantities. For instance, word frequencies (or, indeed, any of the other measures) may show U-shaped trajectories, starting high, falling and then rising again. Our dynamic predictors cannot capture such tendencies.

In the case of informativity and proportion utterance-final, it is not possible to reliably estimate non-linear changes. Both of these measures are derived from our own corpus, where a substantial number of words are only represented by 50–100 tokens. Such a small sample is not sufficient to generate robust non-linear estimates of change over time.

Our estimates of changes in frequency come from the Google N-grams corpus, and are based on many millions of tokens for each word, which makes them more suitable for a non-linear analysis. In order to keep the presentation of our results more streamlined, we will not attempt such an analysis in the main body of this paper. However, we have included a systematic comparison of estimation methods with varying degrees of non-linearity in Appendix B. This comparison shows that linear estimators actually perform better than non-linear ones, which suggests that short-term non-linearities in Google N-grams frequencies do not affect changes in word duration in our corpus.

5. An overview of changes to word duration and usage

In this section, we present a brief outline of general patterns of word-level change in our corpus. This overview will clarify the size and direction of observed changes to word duration and usage factors, and will therefore serve as a useful baseline for the discussion in the following sections. An investigation of the relationship between changing usage and changing duration is only

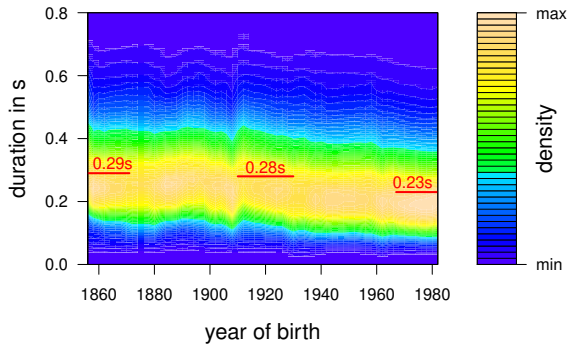

Figure 1: A heatmap showing changes in the distribution of word durations (vertical axis) as a function of speaker year of birth (horizontal axis). Yellow represents higher density areas, while blue represents lower density areas. Density distributions at given time points were calculated by generating kernel density estimates for all word tokens within a 10 year window centred on the time point. The three horizontal lines represent median word duration values calculated for speakers born between 1951–1971 (left), 1910–1930 (middle) and 1967–1987 (right). Both the median values and the density estimates show a clear pattern of shortening over time.

likely to bear fruit in a corpus in which these factors are indeed somewhat in flux.

Figure 1 plots the distribution of word durations against speaker year of birth in the form of a heatmap, and also displays median word durations at three different time points. The median word duration decreases by more than 20 per cent (around 60 ms) between the oldest and youngest speakers, showing an especially steep decline after 1920. This pattern is mirrored by changes in speech rate: the median syllable duration decreases by about 20 per cent from 202 ms (1851–1871) to 166 ms (1967–1987). These changes are probably due to the following factors. Speakers from the first two corpora (the Mobile Unit and the Intermediate Archive) were generally older at the time when they were recorded. Moreover, being recorded was an unusual experience for them, and the strangeness of the situation would likely elicit more formal speech. For some proportion of the interviews in the most recent collection of recordings (the Canterbury Corpus), the interviewee and interviewer were known to each other, which is likely to have further affected formality levels. Both older age (Yuan et al., 2006; Horton et al., 2010) and higher formality (Jacewicz et al., 2010) have been shown to correlate with slower speech. Therefore, it is unlikely that the patterns of increasing speech rate and decreasing word duration observed in our sample reflect general changes that have affected NZE. Nonetheless, it is important to take this trend into account when looking at raw word

CHANGE IN	MEDIAN	2.5%	97.5%
PREV INFORMATIVITY	1.04	−1.7	2.03
FOLL INFORMATIVITY	1.03	−1.7	1.93
PROPORTION UTTR-FINAL	1.01	−4.76	4.74
FREQUENCY	1.12	−1.68	3.92

Table 1: The distribution of changes in usage factors calculated over all 698 words in the data set. The second column shows the median, the third column the 2.5th percentile and the fourth column the 97.5th percentile. The figures show fold changes, that is, the ratio of the higher and the lower values, along with an indicator of the direction of the change (positive values indicate increases, while negative values indicate decreases). For instance, words at the 2.5th percentile of the variable change in proportion utterance-final show a 4.76-fold decrease in how often they occur utterance-finally (e.g. 10% → 2%).

duration trajectories, as it implies that certain words that show an apparent decrease in duration (e.g. 10% over the observed time period) are, in fact, getting longer relative to the rest of the words in the corpus. To adjust for this confound, the mixed models in section 6 include speaker year of birth as a main predictor and corpus as a random intercept. Additionally, graphs of raw word duration trajectories in this paper all include a line that represents the baseline decrease for words of a similar duration.

Table 1 presents overall trends in how words change with respect to the key usage factors involved in our hypotheses. We use fold changes to express these patterns, which, in this paper, are defined as the ratio of the higher value and the lower value. The direction of the change is indicated by the sign of the value: positive values indicate increases, while negative values indicate decreases (as a result of this definition, fold change values are not defined in the interval $(-1, 1]$). As noted above, estimates of the CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY and CHANGE IN PROPORTION UTTERANCE-FINAL were obtained by comparing predictor values for speakers born before 1930 and those born after 1930. Since the separate predictor values are essentially averages over two different halves of the corpus, their differences do not necessarily represent changes across the entire time period. If we were to assign a single time point to the average PREVIOUS/FOLLOWING INFORMATIVITY and PROPORTION UTTERANCE-FINAL values calculated within the two halves of the corpus, the best choice would be the midpoint of each period. These midpoints are separated by about 65-70 years. Therefore, the estimates for CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY and CHANGE IN PROPORTION UTTERANCE-FINAL in table 1 are best interpreted as fold changes over a period of 65-70 years. CHANGE IN FREQUENCY is operationalized in a slightly different

way, by fitting regression lines to log frequency values over speaker year of birth. In order to make these values comparable with the estimates for the other predictors, they are rescaled to represent fold changes over a period of 68 years.

As shown by the medians in table (1), all four usage factors show a very slight overall increase. However, the degree of this increase is relatively small compared to the degree of overall variation in the usage factors, which is indicated by the 2.5th and the 97.5th percentiles in table 1 (these provide a relatively good sense of the range of variation without including outliers). CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY values range roughly between two-fold decrease and two-fold increase. Decreasing CHANGE IN FREQUENCY values are similar to decreasing CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY values, but we see much more substantial, nearly four-fold increases. CHANGE IN PROPORTION UTTERANCE-FINAL shows even more extreme values, ranging between 5-fold decrease and 5-fold increase. Thus, we see evidence of relatively small adjustments over time to the informativity of individual words, and evidence of somewhat more substantial adjustments to word frequencies, and the proportion of tokens occurring utterance-finally.

While frequency and informativity are relatively straightforward notions that have been explored before, it will be useful to provide a few examples for how a word may change with respect to typical position within the utterance. The word *awful* shows a 9-fold increase in PROPORTION UTTERANCE-FINAL, moving from 1.7% in the pre-1930 half of the corpus to 15.4% in the post-1930 half. Perhaps the main reason for this change is that speakers born earlier appear to use *awful* mainly attributively or as an adverb (e.g. *awful frightened*, *awful lot*, *awful smell*), while a predicative use becomes more frequent for speakers born later (e.g. *bloody awful!*, *it was awful*). The word *beer* shows the opposite trend: it moves from 14% to 2.3% PROPORTION UTTERANCE-FINAL. This change is likely motivated by a combination of two factors: a shift in the length of the noun phrases in which the word *beer* appears, and a widely observed correlation between the length (or weight) of noun phrases and their position within the sentence. While *beer* is often quantified in the first half of the corpus (e.g. *a bottle of beer*, *a barrel of beer*), it tends to occur without quantification in the second half (e.g. *a beer*, *the beer*). By Behaghel's (1909) law of increasing length, we expect longer noun phrases to be found nearer the end of the utterance (cf. the generative notion of heavy NP shift; Ross 1967; Wasow 2002), which could bring about the observed change in proportion utterance-final.

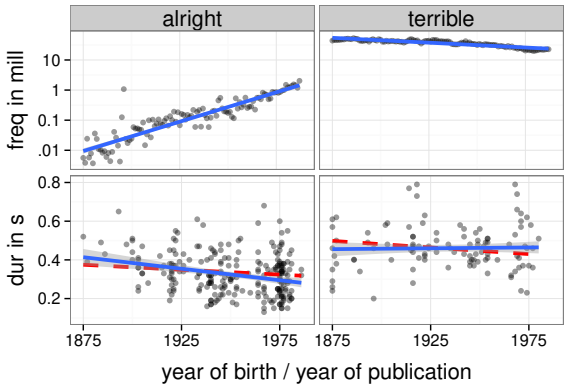

Figure 2: Two examples of parallels between changes in usage factors and word duration: *alright* (left) and *terrible* (right). The top panels plot word frequency (adjusted for a million-word corpus) against year of publication in the Google N-grams corpus. The bottom panels plot word duration against year of speaker birth. All panels include regression lines (blue solid lines) fitted to the data along with 95% confidence intervals. The bottom panels also show the expected decrease in duration for words of similar length (red dashed lines).

While these explanations are admittedly speculative, they at least provide examples of plausible scenarios under which we may observe changes in the typical position of a word within the utterance.

Figure 2 shows two examples that illustrate the types of parallels that we hypothesize to exist between changes in usage factors and changes in word durations. All panels plot time along the horizontal axis. The vertical axis represents frequency for the top panels, and word duration for the bottom panels. The panels on the left show how the word *alright* changed over time: its frequency increased, while its duration decreased (even in relation to the rest of the words; compare the solid blue regression line with the red dashed line, which is an estimate of the baseline decrease in duration). The panels on the right represent the word *terrible*, which displays the opposite pattern, slightly decreasing in frequency and increasing in duration (relative to the rest of the words). These words were handpicked as examples of the predictions in section 3. There are, of course, many other words that do not fit the predicted patterns as closely, or that actually go against them. The models presented in the next sections were designed to evaluate our hypotheses in a statistically more robust way. We suggest that the reader should take a mental note of figure 2, as similar illustrations will be used as insets in our model prediction graphs in section 6.2.2.

The example words in figure 2 exhibit almost entirely linear changes in both log frequency and word duration. However, as noted in the previous section,

not all changes in our corpus are linear. Appendix B provides a brief overview of non-linear changes in frequency, and demonstrates that short-term fluctuations in our frequency measure are generally not matched by corresponding fluctuations in word duration.

6. Statistical analysis and results

Our main approach to statistically evaluating our hypotheses will be to use a two-stage modelling strategy. The first stage is to fit a control model, which accounts for the many local and lexical factors that affect word duration in the corpus. From this model, we extract by-word random slopes over year of birth, representing the degree of durational change for each word, *once these factors are held constant*. These slopes form the input to a treatment model, which assesses each of our key predictors, in order to determine whether any of these predictors significantly contribute to word-level changes in duration. Word duration slopes are simply real numbers, with positive values indicating an increase in duration, and negative values a decrease. To give an example, we expect that word duration slopes will be positive and relatively high for word forms that are increasingly frequent in final position (cf. hypothesis 2b in section 3).

The control model is a large linear mixed-effects regression model based on all 271,764 observations in the data set. This model has two goals. First, we want to extract word duration slopes adjusted for nuisance variables and the local effects of predictability and position within the utterance. These adjustments are required because of our focus on changes to lexical representations: it is not sufficient to look at raw word duration trajectories, as these may reflect trivial non-lexical changes due to local effects (cf. section 2.1). Our second goal is to verify that static lexical factors such as frequency, informativity and typical position within the utterance contribute to overall word durations independently of local effects (cf. section 3). The outcome variable in the control model is word duration. The predictor variables include year of birth, nuisance variables such as speech rate, local factors such as predictability and lexical factors such as informativity. A fuller list of predictors is presented in the next section. We also include random intercepts for speakers, word forms and for the three separate corpora, random slopes for speaker year of birth by words and for each lexical measure by speakers. The full model specification is presented in the supplementary materials.

The treatment model is a linear fixed-effects regression model. This model is fit to the word duration slopes obtained by extracting the by-word random slopes for year of birth. These residual random slopes capture the direction and the magnitude of word duration changes for each word relative to the rest of the words after controlling for local effects and other variables. The predictor variables in this treatment model are the lexical and dynamic lexical predictors listed in section 4.2. The former are included as control variables, while the latter test our main diachronic hypotheses. Since changes in word duration are also likely to be affected by the baseline duration of the words (e.g. a word that is short to start with may be less likely to show further shortening), baseline duration is also included in the model, as well as its interactions with each lexical and dynamic lexical predictor. We did not perform variable selection (e.g. stepwise regression) in order to avoid the inflated rate of false positives associated with such methods (Harrell, 2001).

There are several reasons for following a two-stage approach in our analysis. While it is possible to test our main hypotheses using a single model, this requires complex two and three-way interactions that are somewhat difficult to plot or interpret. In contrast, using word duration slopes as a dependent variable in our treatment model makes the coefficients easily interpretable and allows us to create relatively straightforward plots that show both model predictions and individual data points (see figure 3). Moreover, since the treatment model is not a mixed model, it is possible to estimate how much of the variance in word duration slopes is accounted for by individual predictors (we will refer to this value as ΔR^2; cf. 6.2.2). This would be much less straightforward to calculate for a single mixed effects regression model.

To ensure that our findings are not an artefact of the two-stage design, we repeated the analysis with a single-stage model. We report the results from this model in section 6.2.2 alongside results from the treatment model. Although it has been argued that regressing on random coefficients extracted from another model may lead to anti-conservative results (Hadfield et al., 2010), in our case the results from the two-stage approach appear more conservative than those from the single-stage one, in the sense that the predicted effects are somewhat smaller.

6.1. Control model

6.1.1. Methods

As outlined above, the primary motivation for the control model is to obtain estimates of changes in word duration, which hold extraneous and local predictors constant. The outcome variable for the control model

is raw word duration measured in seconds. Although some previous research used the logarithm of word duration instead of an untransformed measure (Bell et al., 2009; Kuperman and Bresnan, 2012; Seyfarth, 2014), we found that the raw word duration trajectories in our data set do not show clear logarithmic properties. The results of the control model are the same regardless of whether it is fit to untransformed or log durations. We decided to use raw durations, since the treatment model is more directly interpretable if the slopes that serve as its outcome variable are extracted from a control model fit to untransformed word durations.

Our control model tested the key local and lexical predictors introduced in section 4.2. The local predictors are UTTERANCE-FINAL and PREVIOUS/FOLLOWING PREDICTABILITY, while the lexical predictors are PROPORTION UTTERANCE FINAL, PREVIOUS/FOLLOWING INFORMATIVITY and FREQUENCY. The control model also includes speaker YEAR OF BIRTH, used as a proxy for the time dimension of change. In addition, we also included the following control variables: part-of-speech based on automatic parsing using NLTK (Bird et al., 2009) (PoS); the number of segments in the word form based on a phonemic transcription of the citation form (SEGMENT NUMBER); the number of syllables (SYLLABLE NUMBER); the average length of a syllable in the utterance that the token came from (i.e. the inverse of speech rate, which correlates more linearly with word duration; AVG SYLL DURATION); whether the token was utterance-initial (UTTERANCE-INITIAL); whether the word form had been produced by the speaker in the last 20 seconds (i.e. whether it is a repeated word form; REPETITION); and the baseline duration of the token in its five-word context (in seconds; BASELINE DURATION). As for the last predictor, we follow Demberg et al. (2012) and Seyfarth (2014) in using the MARYTTS speech synthesis toolkit (Schröder and Trouvain, 2003) to generate baseline durations. BASELINE DURATION is simply the duration of the token in a synthesized speech segment where it occurs in the same context as in the real recording. Including BASELINE DURATION in the model allows us to 'control for the segmental length, content and context of each word form' (Seyfarth, 2014, p. 144). All continuous variables are scaled and centred. We did not test any interactions in this model.

6.1.2. Results

Table 2 presents a summary of the fixed effects in the control model.

The significance values in this table all come from log-likelihood ratio tests using the χ^2 statistic (cf. Seyfarth 2014). For each predictor, we compared the full model containing the predictor and a nested model without it. We report p-values for each predictor based on this comparison. The p-value thus relates to the predictor as a whole. As a result, we only report a single p-value for PoS, despite the fact that it is represented by more than a single coefficient in the model.

	β	SE	t	$p(\chi^2)$
INTERCEPT	0.281	0.0055	51.32	
YEAR OF BIRTH	−0.012	0.0016	−7.79	< 0.0001
BASELINE DURATION	0.014	0.0005	27.66	< 0.0001
SEGMENT NUMBER	0.033	0.0021	15.37	< 0.0001
SYLLABLE NUMBER	0.011	0.0020	5.61	< 0.0001
AVG SYLL DURATION	0.032	0.0002	172.74	< 0.0001
UTTERANCE-FINAL	−0.010	0.0011	−8.94	< 0.0001
	0.012	0.0008	14.65	< 0.0001
UTTERANCE-INITIAL	−0.010	0.0014	−7.48	< 0.0001
PROPORTION UTTR-FINAL	0.008	0.0023	3.34	0.0009
PREV PREDICTABILITY	−0.002	0.0002	−9.67	< 0.0001
FOLL PREDICTABILITY	−0.017	0.0002	−78.50	< 0.0001
PREV INFORMATIVITY	−0.004	0.0021	−1.89	0.0588
	0.006	0.0020	3.08	0.0020
FOLL INFORMATIVITY	0.015	0.0026	5.66	< 0.0001
FREQUENCY	−0.002	0.0017	−1.10	0.2749
REPETITION	−0.007	0.0004	−17.65	< 0.0001
PoS = ADVERB	−0.014	0.0059	−2.33	< 0.0001
PoS = NOUN	0.007	0.0042	1.59	–
PoS = VERB	−0.015	0.0050	−3.02	–

Table 2: Summary of fixed effects in control model. The p-values were generated using model comparisons based on χ^2 tests. For UTTERANCE-FINAL and PREVIOUS INFORMATIVITY we display both values from the original model (black) and from a model where collinear predictors (BASELINE DURATION for UTTERANCE-FINAL; FOLLOWING PREDICTABILITY and FOLLOWING INFORMATIVITY for PREVIOUS INFORMATIVITY) were removed (grey).

As is evident from table 2, all factors in the model reach significance by this criterion, with the exception of FREQUENCY. We conducted careful checks on collinearity for the model and found that the signs for two of the predictors were substantially affected by collinearity. Values for these predictors from models which exclude collinear predictors (as outlined in Appendix A) are shown below the original estimates in grey.

Since the control model is not the main focus of this paper, we only discuss effects that are related to the synchronic predictions outlined in section 3. The majority of our predictions are borne out by the model: FOLLOWING and PREVIOUS INFORMATIVITY both have a significant positive effect on word duration (although, in the case of PREVIOUS INFORMATIVITY, only when collinear predictors are removed), and PROPORTION UTTERANCE-FINAL also has a significant positive effect on word duration. These effects are significant regardless of whether by-speaker random slopes for the relevant predictors are included

in the model or not (the summary in table 2 shows the results with random slopes).

However, one of our predictions does not seem to be supported by the model: although the effect of FREQUENCY is in the right direction (i.e. it is negative), it does not reach significance. While this seems to go against previous findings in the literature, this contradiction is only apparent. Previous research that has reported significant frequency effects on word duration has relied on models that did not control for informativity (e.g. Gahl 2008; Bell et al. 2009). Seyfarth (2014), on the other hand, also included informativity as a predictor alongside frequency, and failed to find a significant frequency effect (Seyfarth, 2014). He argues that this is likely due to collinearity between frequency and informativity. In line with these observations, when PREVIOUS and FOLLOWING INFORMATIVITY are removed from our control model, FREQUENCY becomes a significant factor ($\beta = -0.004$, SE = 0.0016, $t = -2.23$, $p(\chi^2) = 0.0266$). This suggests that while informativity-based predictors (and FOLLOWING INFORMATIVITY in particular) are more robust than frequency, the separate contributions of these predictors are not easy to estimate due to issues of collinearity.

In sum, our data set shows the same effects reported by Seyfarth (2014), Bell et al. (2009) and Gahl (2008) and an additional lexical effect of the frequency with which a form occurs utterance-finally. Note also that previous informativity is a weaker predictor of word duration than following informativity, which is in line with Seyfarth's (2014) findings.

6.2. Treatment model

6.2.1. Methods

The outcome variable for the treatment model is word duration slope, obtained by extracting by-word random slopes for speaker year of birth from the control model.

To aid interpretability, the slopes were re-scaled so that they represent the expected degree of deviation from the overall trend in word duration trajectories over a period of 100 years. For instance, a value of 0.02 would indicate that a given word increased its duration by 20 ms over 100 years compared to the rest of the words.

The predictors include all of the lexical predictors outlined in section 4.2: PREVIOUS and FOLLOWING INFORMATIVITY, PROPORTION UTTERANCE-FINAL, and FREQUENCY. These were included mainly as control variables to allow for the possibility that word durations may change differently as a function of static usage factors (e.g. a frequent word may become shorter over time even if its frequency does not change). We also include the key dynamic lexical predictors: CHANGE IN FREQUENCY, CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY, and CHANGE IN PROPORTION UTTERANCE FINAL. These predictors constitute a test of hypotheses 3a-3c.

In addition, the model also includes AVERAGE BASELINE DURATION (calculated by averaging the context-specific baseline duration values for a given word) and its interactions with all other predictors. This allows the estimated effects of our predictors to vary as a function of the words' baseline duration. All predictors were scaled and centred.

We also fit an alternative single-stage model to the data. This model is a mixed effects regression model, which is identical to the *control* model in terms of its general structure, including the data set, the outcome variable and its random effects structure. The only difference between the control model and the single-stage model is that the latter also includes interactions between year of birth and all the lexical and dynamic lexical predictors listed in the previous paragraphs (including interaction terms between lexical / dynamic lexical variables and average baseline duration, yielding three-way interactions). These added interaction terms capture the degree to which the slope of the regression line corresponding to year of birth changes as a function of lexical and dynamic lexical predictors, providing an alternative test of our hypotheses. The resulting model is slightly unusual in that it contains interaction terms between year of birth and dynamic predictors, but does not contain the dynamic predictors themselves as main terms. This is because these main terms are theoretically meaningless in the context of the current model. Although we expect a correlation between changes in usage factors and *changes* in word duration, we do not expect a correlation between changes in usage factors and a word's *average* duration (which is what the main terms would capture). Note, however, that we also ran the model with the main terms included, and obtained exactly the same results (with no significant main terms for dynamic predictors).

6.2.2. Results

Table 3 shows a combined model summary for the treatment model and the single-stage model. Both sets of estimates are taken from full models, but the table only includes terms that were significant in at least one of the two models (and CHANGE IN FREQUENCY, which is not significant as a main term, but is part of a significant interaction with AVERAGE BASELINE DURATION). The left-hand side of table 3 shows each of the estimates from the treatment model along with the corresponding stan-

	Treatment model				Single-stage model			
	β	SE	t	$p(>\|t\|)$	β	SE	t	$p(>\|\chi^2\|)$
change in previous informativity	0.0016	0.0008	2.02	0.0434	0.0034	0.0013	2.61	0.0094
change in proportion utterance-final	0.0018	0.0008	2.36	0.0185	0.0026	0.0012	2.24	0.0252
change in frequency	−0.0012	0.0010	−1.26	0.2087	−0.0021	0.0015	−1.42	0.1553
average baseline duration	−0.0059	0.0013	−4.40	< 0.0001	−0.0060	0.0019	−3.21	0.0017
change in proportion utterance-final × average baseline duration	−0.0017	0.0006	−3.07	0.0022	−0.0029	0.0009	−3.03	0.0060
change in frequency × average baseline duration	−0.0012	0.0006	−2.24	0.0254	−0.0023	0.0009	−2.44	< 0.0001

$R^2 = 0.147$ (14.7%); $F(17, 680) = 6.877$, $p < 0.0001$

Table 3: Summary of effects in the treatment model (left) and the single-stage model (right). Effects with an × symbol are interaction terms. The numeric columns provide standard information about the estimated effects.

dard error, t-value and p-value. The right-hand side of the table shows estimates, standard errors, t-values and p-values from the single-stage model.[3] Although not shown in the table, all of the terms from the single-stage model are interactions with year of birth. In order to make the results more comparable across the two models, the estimates and standard errors from the single-stage model were rescaled to represent predicted deviations from the general decrease in word duration over a period of 100 years (see 6.2.1). All of the non-dynamic lexical factors were included in these models, but none of them reached significance.

Figure 3 provides a visual summary of the same findings in the form of model prediction plots. The solid lines and the confidence intervals represent model predictions from the treatment model, while the dashed lines represent model predictions from the single-stage model. Separate plots are presented for change in previous informativity, change in proportion utterance-final and change in frequency. The latter two also display interactions with baseline duration. Although the predictors were centred and scaled in both models, the prediction plots show the original untransformed scales to aid interpretability. In addition, the horizontal axes also show equivalent fold changes calculated in the same way as in section 5. The plots also include insets that illustrate predicted changes in word duration (blue solid lines) at different values of the predictor variable, along with the baseline decrease in word duration observed in the corpus (red dashed line; see previous section). These lines were generated from the raw word duration data by averaging over regression lines fit to multiple words with slope values in a specific range.

Collinearity is not an issue for the treatment model. There is only one pair of variables that are correlated at $|R| > 0.5$, following informativity and proportion utterance-final. Removing either of these variables did not affect the estimates for any of other the variables. Collinearity is a more complex matter for the single-stage model, as it includes a wide range of control variables alongside our main predictors. We did not perform separate collinearity checks for this model as it only plays a supporting role in our analysis and the results from this model are quantitatively very similar to those from the treatment model.

The results support all three hypotheses in section 3. All the observed patterns are in the expected direction. The descriptions in the summary below should all be interpreted in relation to the overall decrease in word duration. Thus, when a given set of words is described as 'increasing in duration', this increase is relative to the rest of the words (i.e. the red line in the insets). In absolute terms, the words may still be getting shorter, although less so than other words.

Words that are becoming less informative based on the previous context are becoming shorter, while words that are becoming more informative show the opposite pattern (figure 3a; cf. hypothesis 3a in 3). Short words that are increasingly frequent in utterance-final position are becoming longer, while long words do not seem to be affected by change in proportion utterance-final (figure 3b; cf. hypothesis 3b in 3). Note also that short words in general seem to show an increase in duration relative to the rest of the words, which is likely due to slight non-linearities in the way word durations vary (i.e. long words have more room to shorten, while short words have more room to lengthen). Finally, long words that are becoming more frequent show a decrease

[3] The t-values and p-values for the single-stage model are not in full correspondence as the p-values for this model are based on log-likelihood ratio tests, not t-tests (cf. section 6.1.2).

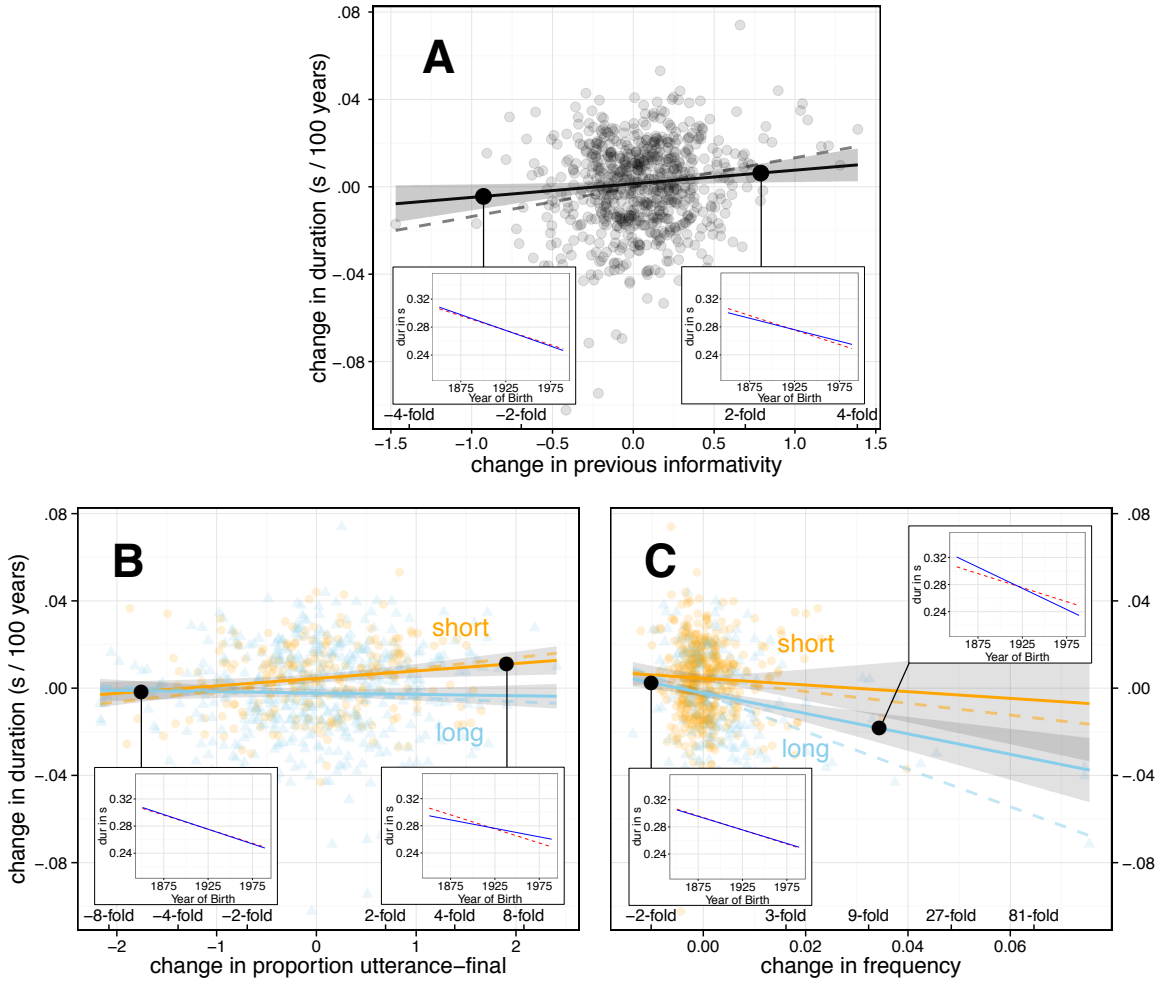

Figure 3: Raw data from the treatment model and model predictions from the treatment and single-stage models. The three main panels are set up as follows: the vertical axis indicates change in word duration (the outcome variable); the horizontal axis indicates (a) change in previous informativity, (b) change in proportion utterance-final, (c) change in frequency. To aid interpretability of dynamic predictors, equivalent fold changes are indicated above the values on the horizontal axes. Panel (a) shows the data points (grey dots) along with model predictions from the treatment model (solid black line) and the corresponding 95% confidence intervals (grey areas around regression line) as well as model predictions from the single-stage model (dashed grey line). Panels (b) and (c) also display the interaction between baseline duration and the relevant predictors by (i) including separate regression lines for short words (orange; at the lower quartile of baseline duration values) and long words (aqua; at the upper quartile of baseline duration values), and (ii) using the same colours to distinguish data points representing short and long words. As before, solid lines and the grey areas show predictions and confidence intervals from the treatment model, while the dashed lines show predictions from the single-stage model. For all panels, the insets show the baseline decrease in word duration (red dashed line) and the predicted change for words with a given value along horizontal axis (blue solid line). Note that the insets show predicted values, not specific words.

in duration, while short words are more or less unaffected by changes in frequency (figure 3c; cf. hypothesis 3c in 3).[4] Somewhat surprisingly, of the two dynamic informativity-based predictors only CHANGE IN PREVIOUS INFORMATIVITY is significant and not CHANGE IN FOLLOWING INFORMATIVITY. This is unexpected in light of the observation that the synchronic effects of following predictability and informativity are stronger than the effects of previous predictability and informativity.

The effect sizes vary both across predictors and models. The single-stage estimates are consistently more extreme than those from the treatment model, which is especially clear in figure 3. Following Bell et al. (2009), we use ΔR^2 to quantify how much of the variability in the word duration slopes is accounted for by specific predictors. This value is calculated by comparing the R^2 value of a version of the treatment model including the relevant predictors and another version excluding them. R^2 is measured in percentages in order to make the figures easier to interpret. We compare the full model with nested models where both the relevant main term and its interaction with AVERAGE BASELINE DURATION are dropped.

The ΔR^2 value for CHANGE IN PREVIOUS INFORMATIVITY is relatively low at 0.64. Looking at the predictions from the more moderate treatment model, it appears that even large changes in this predictor only lead to small changes in word duration slopes. For instance, the difference between the two solid black dots in figure 3a is only about 10 ms, meaning that words that are evolving in opposite directions with respect to previous informativity are predicted to diverge only by 10 ms in duration over a hundred years. To put this figure into perspective, 10 ms is only about 4% of the median word duration (260 ms) in the corpus. The effects of CHANGE IN PROPORTION UTTERANCE-FINAL and CHANGE IN FREQUENCY along with their interactions with baseline duration are substantially more robust, with ΔR^2 values of 1.29 and 3.59, respectively. The strength of these effects is also clearly visible in the prediction plots in 3. For instance, words whose frequencies are changing in opposite directions may diverge by more than 20 ms in duration over a hundred years (cf. the solid black dots in figure 3c). The estimates from the single-stage model are even more extreme, and are nearly twice the size of those from the treatment model.

[4] Since the distribution of CHANGE IN FREQUENCY values is positively skewed with a few outliers at the positive end, we have refit the model to a data set where 18 words with the highest CHANGE IN FREQUENCY values were removed (this constitutes 2.5 per cent of the data set). This actually slightly increased the strength of the effect for long words, and did not change the effect for short words at all.

7. Discussion and conclusions

Let us briefly summarize the main findings presented in the previous sections.

In section 6.1.2, the data set was shown to exhibit robust overall synchronic lexical effects of informativity and typical position in the utterance, replicating and extending previous findings in the literature (Seyfarth, 2014; Bell et al., 2009; Gahl, 2008). The effect of frequency was found not to be significant in the full model (in line with Seyfarth's 2014 findings), but did reach significance when the model was fitted without collinear informativity-based predictors (similar to Gahl 2008 and Bell et al. 2009). These results reinforce a number of existing results showing that local biases exert a cumulative effect on the lexicon, leading to variation in lexical representation.

Our primary hypotheses regarded the degree to which changing usage might predict change over time. Indeed, we found evidence for dynamic effects for all three of our predictors (hypotheses 3a-3c). Words that were increasing/decreasing in informativity (based on the previous contexts) showed a change in the same direction in duration (e.g. increasing informativity is associated with increasing duration). Long words that were becoming more frequent were also becoming shorter. Short words that were increasingly appearing utterance finally were also becoming longer. It is important to recall that the treatment model holds constant the local effects of the predictors. Thus, the results show – for example – that words that are increasing in utterance-finality are also increasing in duration, even when the local position of each token is accounted for.

7.1. Effect sizes

Although there is good evidence for the role of the dynamic predictors, the effect sizes are not large. Neither the individual ΔR^2 values, nor the overall R^2 value for the treatment model are particularly high, indicating that there is a substantial amount of variation in word duration slopes that is not explained by static and dynamic predictors based on usage factors. Some of the predicted differences in word duration slopes shown in figure 3 are also small, although CHANGE IN PROPORTION UTTERANCE-FINAL and CHANGE IN FREQUENCY have a more robust influence on durations, leading to differences in word duration slopes that peak around 20-30 ms / 100 years. The relatively small contribution of usage factors to word duration variation is not surprising given previous findings in the literature. Although frequency and predictability are more or less consistently found

to be significant predictors of word duration (see section 2.1), the multiple regression models presented by Bell et al. (2009) show that individual predictors based on frequency and predictability rarely account for more than 1-2% of the variance in word durations (Bell et al., 2009, 101). These figures are very similar to the ΔR^2 values reported in section 6.2.2. Since synchronic effects based on usage factors are relatively small, there is no reason to expect that changes in word duration should show more pronounced effects.

An important distinction is between informativity *versus* frequency and position within the utterance. The hypotheses related to CHANGE IN FREQUENCY (3b) and CHANGE IN PROPORTION UTTERANCE-FINAL (3c) receive strong support from the treatment model. Note that both of these effects are mediated by baseline duration. Our predictions based on informativity receive less support from the data. While CHANGE IN PREVIOUS INFORMATIVITY is significant in the final treatment model, CHANGE IN FOLLOWING INFORMATIVITY is not, even though one would expect this effect to be stronger based on the synchronic results. We elaborate on these observations below.

The relative robustness of CHANGE IN FREQUENCY and CHANGE IN PROPORTION UTTERANCE-FINAL as opposed to informativity-related dynamic predictors is likely linked to the observation that the former exhibit a much wider range of changes than the latter. As shown in table 1, changes in frequency and proportion utterance-final are often around 2-3 times greater than changes in previous and following informativity (when quantified using fold changes). Although it is difficult to say whether such a comparison across different types of predictors is meaningful, it appears that informativity is somewhat more stable in our corpus than frequency and typical position in the utterance. This relative stability may contribute to the small size of informativity-based effects on word duration slopes.

We suspect that there is also another, slightly more mundane reason for these differences in effect size. While CHANGE IN PROPORTION UTTERANCE-FINAL and especially CHANGE IN FREQUENCY were estimated in a robust way, the reliability of the language models required to calculate predictability and informativity decreases when separate models are constructed for the two halves of the corpus. This makes the CHANGE IN INFORMATIVITY measures rather noisy, which can, in turn, lead to smaller effect sizes and non-significant estimates. This may also explain the apparent contradiction that FREQUENCY is less robust than PREVIOUS/FOLLOWING INFORMATIVITY in the control model, but CHANGE IN FREQUENCY is more robust than CHANGE IN PREVIOUS/FOLLOWING INFORMATIVITY in the treatment model. The informativity measures in the control model are robust and therefore act as suppressors for the frequency effect. However, our estimates of change in informativity are noisier, allowing the effect of CHANGE IN FREQUENCY to surface. It is perhaps also a result of this noisiness that the effect of CHANGE IN FOLLOWING INFORMATIVITY is weaker than that of CHANGE IN PREVIOUS INFORMATIVITY, which is the opposite of what we would expect based on the synchronic results and previous observations in the literature.

7.2. Interactions with baseline duration

As noted above, the effects of CHANGE IN FREQUENCY and CHANGE IN PROPORTION UTTERANCE-FINAL vary as a function of AVERAGE BASELINE DURATION. Specifically, only long words are affected by the former and only short words are affected by the latter. This finding is relatively easy to interpret if we look more carefully at the effects of these two predictors. As shown in figure 3b, the main manifestation of the effect of CHANGE IN PROPORTION UTTERANCE-FINAL is a *lengthening* of words that are becoming more frequently utterance-final. On the other hand, the main manifestation of the effect of CHANGE IN FREQUENCY is a *shortening* of words that are becoming more frequent in general. The interaction with baseline duration follows straightforwardly from these observations: long words have simply more room to shorten, while short words have more room to lengthen.

7.3. Possible accounts

We interpret these results as evidence in favour of the accumulation of local effects at the lexical level. As outlined in section 2.2, there are a number of different views about the nature of the information that accumulates in lexical representations. The most straightforward interpretation is that details of fine phonetic variation are directly represented in the lexicon. However, it is worth reviewing possible alternate accounts of the results. We do this by stepping through the different approaches considered by Seyfarth (2014) (and briefly discussed in 2.2) to account for his finding that informativity is a robust predictor of word-duration.

One potential account invokes abstract lexical representations in which a word may be represented by a non-reduced and one or more reduced variants (Bürki et al., 2010). In such an account, increased usage of a reduced variant due to changing local factors would lead to the reduced variant being more easily accessible. This variant would then be more likely to be selected even in the absence of favouring local factors. In principle, such an account would predict that word duration distributions should be bimodal or multimodal (representing the different reduced and non-reduced variants), and that the

durational shifts we have observed should simply represent shifts in the frequencies of the modes. In practice, these predictions are impossible to test in our highly variable data set, where individual modes are likely obscured by the large amount of noise due to measurement errors, across-speaker variation, fluctuations in speech rate, and so on.

Another possibility raised by Seyfarth is that speakers construct lexical representations relying on a 'rational speech production' strategy. This strategy consists in choosing a default form for any given word that minimizes the overall need to deviate from that form in production. The representation does not contain a distribution of past encounters of a word, but is nonetheless still shaped by past experience. In such an account, deviations from the default representation involve planning costs, and are minimized (Seyfarth, 2014). This emphasis on motor planning is consistent with the model of speech production advocated in (MacDonald, 2013). An account based on 'efficient articulation' would posit a similar mechanism at the level of the articulatory gesture, which allows speakers to learn how tight the gestural timing needs to be for a particular word (cf. Lavoie 2002). Our results would require the 'default' form of these accounts to be highly dynamic, shifting gradually in order to keep track of changes in usage patterns. Therefore, these accounts still need to rely on a tight feedback loop at the lexical level, which is highly responsive to the distribution of experiences of a particular word.

Finally, Seyfarth suggests that it is also possible to account for the informativity results without phonetically detailed lexical representations. In such an account, individuals store not the overall phonetic detail of previous encounters, but rather some abstracted information about the average predictability of a word. Any individual production, then, would be influenced by this probability, such that words that are overall more probable are produced in a more reduced way. This account requires tracking of usage patterns at the word level in order to establish the overall probability. In addition, as pointed out by Seyfarth, it also requires speakers to balance multiple types of local and lexical probability in order to settle on a production target. Our own results add further to this complexity, by requiring that the probabilities are constantly updated, and also that they exist at multiple levels, including the likelihood of a word occurring utterance-finally.

Regardless of the exact nature of lexical representations, any account of our results must involve updating of *something* at the word level – be it phonetic detail, probability distributions, default productions or gestural patterns. The evidence presented here for a feedback loop between local usage patterns and lexical representations is unequivocal. We believe the storage of fine phonetic detail is the most parsimonious account. It is also consistent with a myriad recent results in the literature, showing that word-level representations are shaped in phonetically gradient ways by the distribution of linguistic and social environments in which we encounter them (Walker and Hay, 2011; Hay and Maclagan, 2012; Sóskuthy et al., 2015; Hay and Foulkes, 2016; Raymond et al., 2016).

7.4. Implications for language change

The observed patterns of co-adaptation between usage and duration also have important implications for studies of language change. Following Weinreich et al. (1968), many scholars argue that the selection of a specific pathway of change in a given language at a given point in time is a fundamentally social affair (e.g. Milroy and Milroy 1985; Labov 1994, 2002; Croft 2000). Otherwise, how could it be that different languages and varieties undergo different sets of changes seemingly at random, even in cases where the same changes could be applicable? While we acknowledge the crucial role of social factors in language change, the results in this paper suggest another potential contributor to such cross-linguistic differences. We have demonstrated that changes in one linguistic domain (the distribution of a word across different contexts) can be related to changes in a different domain (word duration). While two languages or varieties may appear very similar in a given linguistic domain (e.g. they have the same phoneme inventories), they may be quite different in other domains (e.g. the frequency distributions of different words). The types of mechanisms that we have identified could lead to situations where such differences in one domain result in different patterns of change in another domain (cf. Sóskuthy 2013, 2015).

7.5. Conclusions

Taken together, our data provide solid evidence that lexical representations respond to changes in usage factors. The parallel changes we have found are not simply the result of superficial shifts in exposure to local effects, but manifest themselves at a deeper level. To our knowledge, this is the first diachronic demonstration of the emergence of lexical effects on word durations based on quantitative evidence.

The production-perception feedback loop provides a straightforward account of the observed phenomena by

suggesting that lexical effects are simply the cumulative consequences of local biases in cognitive representations. The current study is unique in that it provides a view of the production-perception feedback loop in action, looking directly at the emergence of lexicalized effects. It thereby provides even stronger evidence for this mechanism than previous studies, which have only been able to investigate it by looking at its end results.

Acknowledgements

This work has been supported by a Rutherford Discovery Fellowship granted by the Royal Society of New Zealand to the second author. In addition, this project was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the Royal Society of New Zealand. The ONZE data was collected by the Mobile Disc recording Unit of the NZ Broadcasting Service, Rosemary Goodyear, Lesley Evans, members of the NZ English class of the Linguistics Department, University of Canterbury, and members of the ONZE team. The work done by members of the Origins of New Zealand English Project (ONZE) in preparing the data, making transcripts, and obtaining background information is also gratefully acknowledged. The Corpus was created and supported with funding from the following sources: University of Canterbury, Foundation for Research, Science and Technology (the New Zealand Public Good Science Fund), the Royal Society of New Zealand, The New Zealand Lotteries Board Fund, and the Canterbury History Foundation. We are particularly grateful to Robert Fromont for his work programming LaBB-CAT – the ONZE Corpus search engine and interactive interface, and for his help with extracting the data for this paper. This manuscript has benefited from feedback from Clay Beckner, Susanne Gahl, two anonymous reviewers, colleagues at the University of York and the New Zealand Institute of Language, Brain and Behaviour, and audiences at the University of Glasgow, University of Leeds, the PAC2015 conference in Toulouse, UKLVC10 in York and mFiL 2015 in Manchester.

References

Aylett, M., Turk, A., 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. The Journal of the Acoustical Society of America 119 (5), 3048–3058.

Bailey, G., 2008. Real and apparent time. In: Chambers, J., Trudgill, P., Schilling-Estes, N. (Eds.), The handbook of language variation and change. Blackwell Publishing, Oxford, pp. 312–332.

Behaghel, O., 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen, 110–142.

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.

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.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly Media, Inc., Sebastopol, CA.

Bürki, A., Ernestus, M., Frauenfelder, U. H., 2010. Is there only one fenêtre in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. Journal of Memory and Language 62 (4), 421–437.

Bybee, J., 2001. Phonology and language use. Cambridge University Press, Cambridge.

Bybee, J., 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. Language Variation and Change 14 (03), 261–290.

Cohen Priva, U., 2015. Informativity affects consonant duration and deletion rates. Laboratory Phonology 6 (2), 243–278.

Croft, W., 2000. Explaining language change: An evolutionary approach. Pearson Education, Harlow, UK.

Demberg, V., Sayeed, A. B., Gorinski, P. J., Engonopoulos, N., 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, pp. 356–367.

Friedman, L., Wall, M., 2005. Graphical views of suppression and multicollinearity in multiple linear regression. The American Statistician 59 (2), 127–136.

Fromont, R., Hay, J., 2008. ONZE Miner: the development of a browser-based research tool. Corpora 3 (2), 173–193.

Fromont, R., Watson, K., in press. Factors influencing automatic segmental alignment of sociophonetic corpora. Corpora.

Gahl, S., 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. Language 84 (3), 474–496.

Goldinger, S. D., 2000. The role of perceptual episodes in lexical processing. In: Cutler, A., McQueen, J. M., Zondervan, R. (Eds.), Proceedings SWAP (Spoken Word Access Processes). Max Planck Institute for Psycholinguistics, Nijmegen, pp. 155–159.

Gordon, E., Maclagan, M., Hay, J., 2007. The ONZE Corpus. In: Beal, J. C., Corrigan, K. P., Moisl, H. L. (Eds.), Models and methods in the handling of unconventional digital corpora: Volume 2, Diachronic Corpora. Vol. 2. Palgrave Macmillan, Basingstoke, Hampshire, pp. 82–104.

Hadfield, J. D., Wilson, A. J., Garant, D., Sheldon, B. C., Kruuk, L. E. B., 2010. The misuse of BLUP in ecology and evolution. The American Naturalist 175 (1), 116–125.

Harrell, F. E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York.

Hay, J., Foulkes, P., 2016. The evolution of medial (-t-) in real and remembered time. Language 92 (2), 298–330.

Hay, J., Maclagan, M., 2012. /r/-sandhi in early 20th century New Zealand English. Linguistics 50 (4), 745–763.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., LaShell, P., 2015. Tracking word frequency effects through 130 years of sound change. Cognition 139, 83–91.

Horton, W. S., Spieler, D. H., Shriberg, E., 2010. A corpus analysis of patterns of age-related change in conversational speech. Psychology and Aging 25 (3), 708–713.

Jacewicz, E., Fox, R. A., Wei, L., 2010. Between-speaker and within-speaker variation in speech tempo of American English. The Journal of the Acoustical Society of America 128 (2), 839–850.

Jaeger, T. F., Buz, E., to appear. Signal reduction and linguistic encoding. In: Fernández, E. M., Cairns, H. M. I. (Eds.), Handbook of Psycholinguistics. Wiley-Blackwell.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D., 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee, J. L., Hopper, P. (Eds.), Frequency and the emergence of linguistic structure. John Benjamins, Amsterdam, pp. 229–254.

Klatt, D. H., 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. The Journal of the Acoustical Society of America 59 (5), 1208–1221.

Kuperman, V., Bresnan, J., 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. Journal of Memory and Language 66 (4), 588–611.

Labov, W., 1994. Principles of Linguistic Change. Vol. 1: Internal Factors. Blackwell, Oxford.

Labov, W., 2002. Driving forces in linguistic change. In: Proceedings of the 2002 International Conference on Korean Linguistics, Seoul National University.

Lavoie, L., 2002. Some influences on the realization of *for* and *four* in American English. Journal of the International Phonetic Association 32 (02), 175–202.

Lieberman, P., 1963. Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech 6 (3), 172–187.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., Willerman, R., 1995. Is sound change adaptive? Rivista di Linguistica 7 (1), 5–37.

MacDonald, M. C., 2013. How language production shapes language form and comprehension. Frontiers in Psychology 4, 226.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182.

Milroy, J., Milroy, L., 1985. Linguistic change, social network and speaker innovation. Journal of Linguistics 21 (2), 339–384.

Oudeyer, P.-Y., 2006. Self-Organization in the Evolution of Speech. Oxford University Press, Oxford.

Paul, H., 1880. Principien der Sprachgeschichte (2nd Ed: 1886). Max Niemeyer, Halle.

Piantadosi, S. T., Tily, H., Gibson, E., 2011. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108 (9), 3526–3529.

Pierrehumbert, J. B., 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee, J. L., Hopper, P. (Eds.), Frequency effects and the emergence of lexical structure. John Benjamins, Amsterdam, pp. 137–157.

Pierrehumbert, J. B., 2002. Word-specific phonetics. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory phonology, Vol. VII. Mouton de Gruyter, Berlin, pp. 101–140.

Raymond, W. D., Brown, E. L., Healy, A. F., 2016. Cumulative con-

text effects and variant lexical representations: Word use and english final t/d deletion. Language Variation and Change 28 (2), 175–202.

Ross, J., 1967. Constraints on variables in syntax. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Schröder, M., Trouvain, J., 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6 (4), 365–377.

Seyfarth, S., 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. Cognition 133, 140–155.

Sóskuthy, M., 2013. Phonetic biases and systemic effects in the actuation of sound change. Ph.D. thesis, University of Edinburgh.

Sóskuthy, M., 2015. Understanding change through stability: a computational study of sound change actuation. Lingua 163, 40–60.

Sóskuthy, M., Foulkes, P., Haddican, B., Hay, J., Hughes, V., 2015. Word-level distributions and structural factors co-determine GOOSE fronting. In: 18th International Congress of Phonetic Sciences.

Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011. SRILM at sixteen: Update and outlook. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop.

Tabachnick, B. G., Fidell, L. S., 2007. Multivariate Statistics, 5th Edition. Pearson Education, Boston.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., Bresnan, J., 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. Language and Cognition 1 (2), 147–165.

Turk, A. E., Shattuck-Hufnagel, S., 2007. Multiple targets of phrase-final lengthening in American English words. Journal of Phonetics 35 (4), 445 – 472.

Walker, A., Hay, J., 2011. Congruence between 'word age' and 'voice age' facilitates lexical access. Laboratory Phonology 2 (1), 219–237.

Wasow, T., 2002. Postverbal Behaviour. Center for the Study of Language and Information, Stanford.

Wedel, A. B., 2006. Exemplar models, evolution and language change. The Linguistic Review 23, 247–274.

Wedel, A. B., 2007. Feedback and regularity in the lexicon. Phonology 24, 147–185.

Weinreich, U., Labov, W., Herzog, M. R., 1968. Empirical foundations for a theory of language change. In: Lehmann, W. P., Malkiel, Y. (Eds.), Directions for historical linguistics. University of Texas Press, Austin, TX, pp. 95–188.

Whalen, D. H., 1991. Infrequent words are longer in duration than frequent words. The Journal of the Acoustical Society of America 90 (4), 2311–2311.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P. J., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. The Journal of the Acoustical Society of America 91 (3), 1707–1717.

Wood, S., 2006. Generalized Additive Models: An Introduction with R. CRC Press, Boca Raton.

Wurm, L. H., Fisicaro, S. A., 2014. What residualizing predictors in regression analyses does (and what it does not do). Journal of Memory and Language 72, 37–48.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al., 1997. The HTK book. Entropic Cambridge Research Laboratory, Cambridge.

Yuan, J., Liberman, M., Cieri, C., 2006. Towards an integrated understanding of speaking rate in conversation. In: Proceedings of Interspeech 2006. pp. 541–544.

Appendix A. Checks on control model collinearity

The estimates in the control model summary need to be treated with a certain amount of caution, as collinearities in linear regression models are known to distort parameter estimates (Friedman and Wall, 2005). In certain extreme cases these distortions may manifest as a change of sign for variables that are less strongly correlated with the outcome variable than another collinear predictor (Friedman and Wall 2005; Wurm and Fisicaro 2014; see also Seyfarth 2014 for examples of such cases from a related study). Whether such distortions are problematic from the point of view of model interpretation depends on the role of the predictors in question. If the distorted predictors are included in the model purely as controls, there is no reason to worry about their estimates, as they do not affect the interpretation of the treatment variables or the model as a whole (Tabachnick and Fidell, 2007; Wurm and Fisicaro, 2014). One such case arises with UTTERANCE-FINAL, which has a surprising negative estimate in the control model, suggesting that words in utterance-final position are, in fact, shorter than elsewhere (see the values in black next to UTTERANCE-FINAL in table 2). This estimate goes against previous observations in the literature, and is also the opposite of what we observe in the raw data (without controlling for other variables): utterance-final words are on average 36 ms longer than other words, and this relationship is strong and significant by an unpaired two-tailed t-test ($t = 27.663$, df = 11914.12, $p < 0.0001$). The negative estimate in the original model turns out to be an artefact due to the inclusion of BASELINE DURATION. The baseline durations generated by the MARYTTS system already include contextual effects such as lengthening in final position, which makes them highly collinear with the variable UTTERANCE-FINAL. When the model is refit without BASELINE DURATION, the estimate for UTTERANCE-FINAL changes sign in the expected direction (see the values in grey next to UTTERANCE-FINAL in table 2).

Particular caution has to be exercised when at least one of the collinear predictors is a treatment variable. Therefore, we have performed two checks for each of the lexical variables that are essential to the hypotheses presented in section 3: PROPORTION UTTERANCE-FINAL, PREVIOUS INFORMATIVITY, FOLLOWING INFORMATIVITY and FREQUENCY. First, for each lexical variable we compared the model estimate to the zero-order correlation between the lexical variable and the outcome variable (word duration; cf. Wurm and Fisicaro 2014, fn. 4). Only one discrepancy was found: PREVIOUS INFORMATIVITY is positively correlated with word duration when other control variables are not included ($R = 0.35$, df = 271,763, $p < 0.0001$), but the model shows a negative estimate (see the values in black next to PREVIOUS INFORMATIVITY). This suggests that the model estimate for PREVIOUS INFORMATIVITY may have been biased by collinear variables. Second, we checked for strong correlations between each lexical variable and all other variables. If any variable x in the model was found to correlate with one of the lexical variables at around $|R| > 0.5$, we refit the model without x and checked whether there were any substantial changes to the estimate or the significance value for the lexical variable. Again, there was only one case where a change was found: when both FOLLOWING PREDICTABILITY and FOLLOWING INFORMATIVITY (both relatively strongly correlated with PREVIOUS INFORMATIVITY: $R = 0.64$, df = 271,763, $p < 0.0001$ for FOLL INFORMATIVITY and $R = -0.44$, df = 271,763, $p < 0.0001$ for FOLL PREDICABILITY) were removed from the model, the sign of the estimate for PREVIOUS INFORMATIVITY changed to positive (see the grey values next to previous informativity in table 2). All other lexical variables remained stable when collinear predictors were removed, though, as noted in the main text, the effect of FREQUENCY increases in strength when PREVIOUS and FOLLOWING INFORMATIVITY are removed. We interpret these findings as a strong indication that the negative estimate for PREVIOUS INFORMATIVITY in the model is simply due to collinearities, and assume that the estimates obtained after the removal of collinear predictors are a better reflection of the real effect of PREVIOUS INFORMATIVITY (which was also found to be positive in Seyfarth's 2014 study).

Appendix B. Non-linear changes in frequency

Our corpus does not allow us to study how non-linear changes in informativity or proportion utterance-final affect word duration (cf. section 4.2.3). However, it is possible to look at non-linearities in word frequency thanks to the level of detail in the Google N-grams data. In this appendix, we look at the degree to which words in our corpus show non-linear changes in frequency, and then attempt to see whether these non-linearities have any visible effects on the evolution of word durations in our corpus.

The analyses in this section use generalized additive models (GAMs; Wood 2006). GAMs are an extension of linear regression modelling, which allow the inclusion of so-called smooth terms in regression models alongside traditional linear terms. Smooth terms in GAMs are similar to more conventional ways of representing non-linearity in regression models, such as

polynomial regression. In the case of polynomial regression, several transformed versions of the same variable are created by raising them to different powers, and all of these are included as predictors in the same model. The degree of the polynomial (i.e. the number of transformed terms in the model; e.g. $y \sim x + x^2$ vs. $y \sim x + x^2 + x^3 + x^4$) has to be decided on somewhat arbitrarily before the analysis is performed, and this directly affects the amount of non-linearity or 'wiggliness' that the model can support. In contrast, smooth terms in GAMs use penalty terms to determine the degree of 'wiggliness', and estimate these penalty terms directly from the data using generalized cross-validation (Wood, 2006).

In this analysis, we use very simple GAMs with Google N-grams word frequency as the output variable, and a single smooth term corresponding to year of publication as the only predictor variable. The number of basis functions (i.e. the maximum amount of wiggliness allowed by the models) is set to 90 for all of the models described below.

In the first set of models, we fit separate GAMs to each of the 698 word frequency trajectories, and extracted the estimated penalty term from each model. Since lower penalty terms correspond to more wiggly word frequency curves, we can use the estimated values as a rough measure of non-linearity. Figure B.4 shows a density plot of the smoothing penalties and word frequency trajectories over time for three words exemplifying different levels of smoothness. The bulk of the penalty values lie in a region with at least some degree of non-linearity, as shown by the trajectory for *dry*, which comes from the centre of the distribution. Since a single penalty value can correspond to many different trajectory shapes, there are plenty of other types of curves near the median value, but they are all characterized by a certain amount of wiggliness. There are relatively few words that show no non-linearity at all (i.e. words with penalty values close to or higher than that of *rich*). In sum, the majority of the frequency trajectories corresponding to the words in our corpus show at least some level of non-linearity.

This leads us to ask whether non-linear changes in frequency are reflected in word duration trajectories. In order to answer this question, we need to return to our control model (cf. 6.1.2). The control model contains log frequency as a predictor. Since changes in frequency have been shown to affect word durations in 6.2.2, we expect that replacing this static measure with a time-varying measure (i.e. one that provides different values for two tokens of the same word produced by speakers born in different years) would improve the performance of the treatment model. Moreover, we can construct different versions of this dynamic frequency measure with different levels of smoothing over time. If non-linearities in the frequency trajectories affect changes in word duration, we expect that non-linear versions of the dynamic measure should outperform those that smooth over the trajectory in a linear fashion.

We implemented this comparison by fitting different sets of GAMs to the frequency trajectories corresponding to the words in our corpus. The smoothing penalty was fixed at the same value for all words within a single set, and was varied systematically between sets. For each set of GAMs, we refit the control model replacing the static measure of word frequency with model predictions from the by-word GAMs. These model predictions represent dynamic estimates of word frequency that are smoothed over time. The top panels in figure B.5 illustrate different degrees of smoothing over the same trajectory corresponding to different smoothing penalties. At low values, the smoother essentially links individual data points without any smoothing. At intermediate values, we see higher degrees of smoothing, moving towards a straight-line approximation (cf. the third panel from the left). At very high values, the smoother becomes a flat line, which corresponds to a static estimate of frequency (cf. the fourth panel). This flattening occurs as a result of using a thin plate regression spline with shrinkage (otherwise the smoother would converge to a linear regression line at high values).

The bottom panel in figure B.5 shows AIC values from different versions of the control model, which all include dynamic estimates of word frequency based on model predictions from by-word GAMs. The smoothing penalty for the GAMs increases gradually from left to right. Two important observations can be made about this graph. First, true dynamic estimates of frequency clearly result in better model fits than static estimates. This is shown by the fact that AIC values are markedly lower in the first two-thirds of the graph, before the GAM estimates flatten out. Second, the lowest AIC value comes from a model with a relatively high smoothing penalty, where the smoother is close to a linear regression line. In fact, an even lower AIC value can be obtained by using linear regression models instead of GAMs to calculate the dynamic estimates of frequency (-585110.4 for linear models vs. -585109.8 for GAMs). This means that linear estimates of frequency yield better model fits than non-linear ones.

In sum, the first set of models show that changes in word frequency include substantial non-linearities, while the second set of models show that word dura-

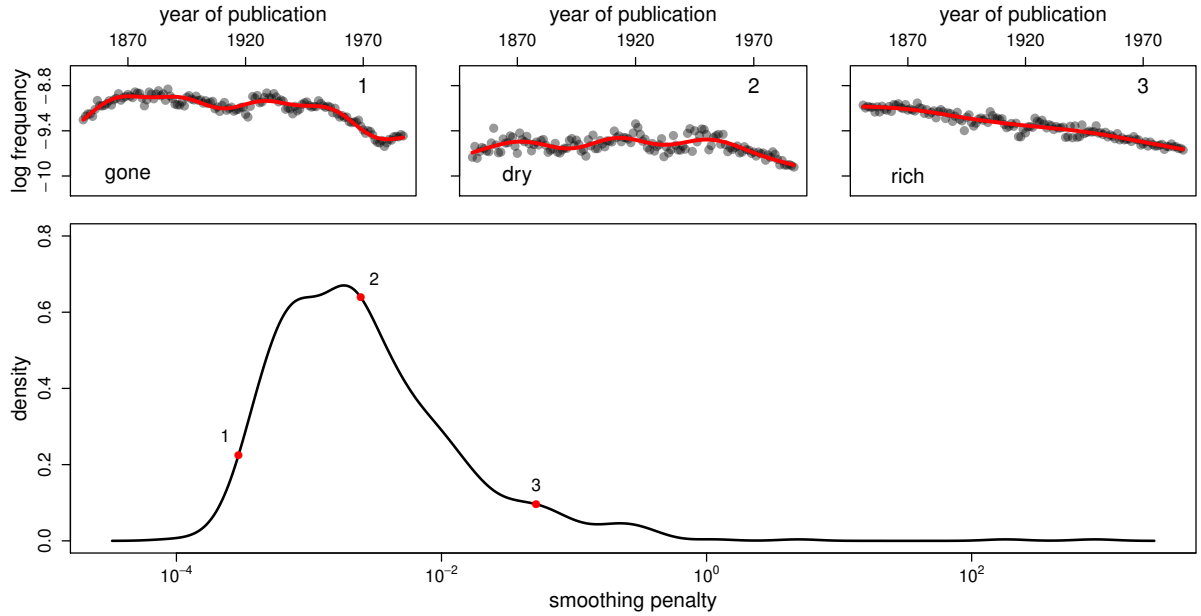

Figure B.4: Bottom panel: a density curve representing the distribution of smoothing penalty values from a set of GAMs fit separately to each word frequency trajectory. Top panels: three word frequency trajectories exemplifying different smoothing penalties. The example words are *gone* (left), *dry* (centre) and *rich* (right). The grey dots show raw log frequencies for each year from the Google N-grams corpus, while the red lines show model predictions from the corresponding GAMs.

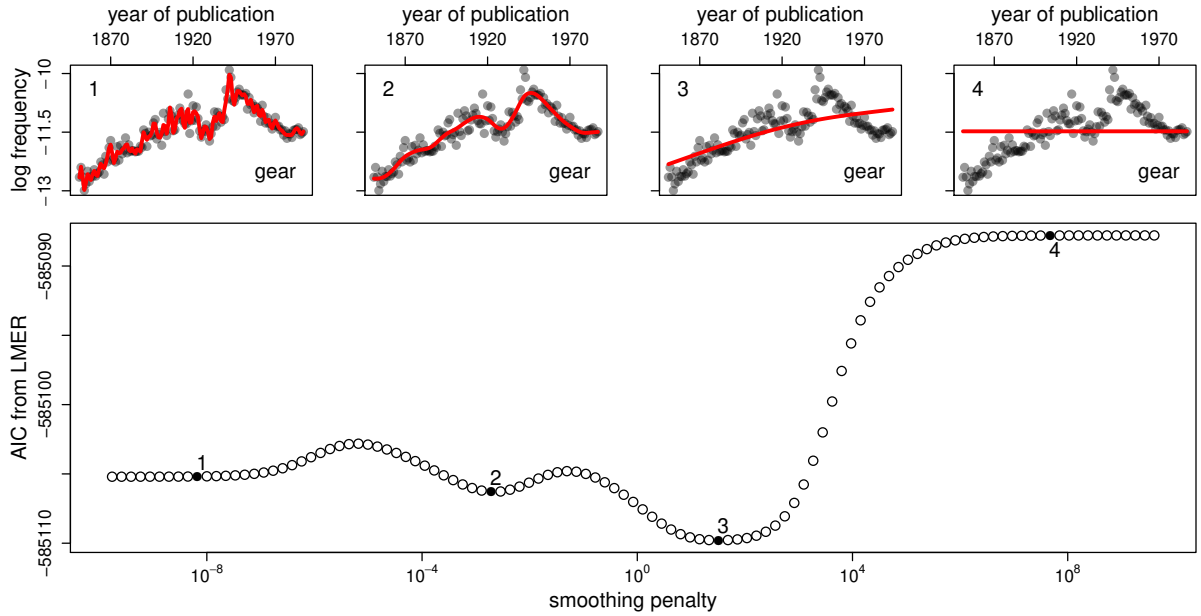

Figure B.5: Bottom panel: AICs from versions of the control model with different dynamic estimates of word frequency. The AIC values are shown as a function of the smoothing penalty for the by-word GAMs, which increases from left to right. Top panels: four GAM smoothers for the same word (*gear*) with different smoothing penalties. Each panel illustrates a different smoothing penalty from the bottom panel, as shown by the indices.

tions only seem to react to broad changes in frequency, and are not affected by non-linearities. This does not necessarily imply that short-term changes in frequency are irrelevant to changes in word duration. It may simply be the case that our corpus-based estimates of frequency and word duration are too course-grained to capture such finer parallels. Moreover, our frequency estimates come from a written corpus representing a different (though closely related) variety, which may further weaken short-term interactions between the two measures. Since our measures of informativity and typical position in utterance are even noisier, it is safe to conclude that our data set is not sufficiently detailed to look for very short-term parallels between word usage and duration.